SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1120-0001 $01.00/0

ON AN EXTENSION OF THE STURM COMPARISON THEOREM*

SHAIR AHMAD† AND ALAN C. LAZER‡

Abstract. The main purpose of this paper is to extend the Sturm comparison theorem for second-order differential equations to the equation $Ly + p(t)y = 0$, where L is a disconjugate linear differential operator of order n, $n \geqq 2$, and where $p(t)$ is a continuous function of constant sign.

In this paper we study the differential equation

$$Ly + p(t)y = 0, \tag{1}$$

where $p(t)$ is a continuous function of constant sign on a compact interval $[a, b]$, and L is a disconjugate linear differential operator of order $n \geqq 2$ which we may assume to be factored as a product of first-order operators: $L_0 y = r_0 y$, $L_h y = r_h (L_{h-1} y)'$ for $h = 1, \cdots, n$, $Ly = L_n y$. This differential equation and special cases have been investigated many times. (See, for example, [1], [3], [4], [5], [8], [9], [10], [11].) An extensive bibliography is contained in [1]. Here we assume that $r_h(t) > 0$ for $t \in [a, b]$, and $r_h \in C^{n-h}$ for $h = 1, \cdots, n$, although if we generalize the notion of a solution of (1) as in [5] or [9] our results will still hold if these functions are only assumed to be continuous. Throughout, $i_1, i_2, \cdots, i_k, j_1, \cdots, j_{n-k}$ will denote fixed integers such that $1 \leqq k \leqq n-1$ and $0 \leqq i_1 < i_2 < \cdots < i_k \leqq n-1$, and $0 \leqq j_1 < j_2 < \cdots < j_{n-k} \leqq n-1$. In our main theorems we shall always make the assumption

(A) *For $a < c \leqq b$ there exists no nontrivial solution of $Ly = 0$ such that*

$$\begin{aligned} (L_{i_s} y)(a) &= 0, \qquad s = 1, 2, \cdots, k, \\ (L_{j_t} y)(c) &= 0, \qquad t = 1, \cdots, n-k. \end{aligned} \tag{2}$$

The following result which is a special case of a more general theorem due to Elias [4, Corollary 3] shows that this condition depends only on the integers i_s, $s = 1, \cdots, k$, and j_t, $t = 1, \cdots, n-k$.

THEOREM 1. *A necessary and sufficient condition for* (A) *to hold is that for any integer l with $1 \leqq l \leqq n-1$, at least l terms of the sequence of integers $i_1, i_2, \cdots, i_k$, $j_1, \cdots, j_{n-k}$ are less than l.*

It is not difficult to show that if the above condition is not satisfied for some l, then some linear combination of l independent solutions of $L_l y = 0$ will satisfy the conditions (2) and the equation $Ly = 0$. Sufficiency can be derived as a consequence of Lemma 1 given below.

Our main result concerning (1) is the following extension of the well-known Sturm comparison theorem for second-order equations:

THEOREM 2. *Let $p_1(t)$ and $p_2(t)$ be continuous on $[a, b]$, and suppose that*

$$(-1)^{n-k} p_2(t) \leqq (-1)^{n-k} p_1(t) \leqq 0 \quad \text{for all } t \in [a, b]. \tag{3}$$

* Received by the editors August 27, 1979, and in final form April 7, 1980.

† Department of Mathematics, College of Arts and Sciences, University of Miami, Coral Gables, Florida 33124. This work was completed while the author was at Oklahoma State University, Stillwater, Oklahoma 74074, and was partially supported by the National Science Foundation under grant MCS 78-01480.

‡ University of Cincinnati, Cincinnati, Ohio 45220. The work of this author was partially supported by the National Science Foundation under grant MCS 77-25147.

If $p_1 \not\equiv p_2$, if condition (A) *holds, and if there exists a nontrivial solution u of*

$$
\begin{aligned}
&Lu + p_1(t)u = 0, \\
&(L_{i_s}u)(a) = 0, \qquad s = 1, \cdots, k, \\
&(L_{j_t}u)(b) = 0, \qquad t = 1, \cdots, n-k,
\end{aligned} \tag{4}
$$

then there exists a nontrivial solution y of

$$Ly + p_2(t)y = 0, \tag{5}$$

such that y satisfies the boundary conditions (2) *for some c with $a < c < b$.*

Remark. Elias [3, p. 256] has shown that if $p(t)$ is not identically zero on any subinterval of $[a, c]$ and does not change sign, then the existence of a nontrivial solution of (1) satisfying the boundary conditions (2) implies that $(-1)^{n-k}p(t) \leqq 0$ on $[a, c]$.

For the special cases $i_s = s-1$, $s = 1, \cdots, n-1$, $j_1 = 0$, and $i_1 = 0$, $j_t = t-1$, $t = 1, \cdots, n-1$, Theorem 2 was established by Schmitt [10] in a slightly weaker form. Later, in [1], the authors established Theorem 2 for the special case $i_s = s-1$, $s = 1, \cdots, k$, $j_t = t-1$, $t = 1, \cdots, n-k$, where k satisfies $1 \leqq k \leqq n-1$. Utilizing recently published results of Elias [4], we are now able to give a simpler proof of the more general theorem.

We present a simple example in which (A) does not hold and Theorem 2 does not apply. Let $r_0(t) = r_1(t) = r_2(t) \equiv 1$, and $Ly = L_2y = y''$. Let $i_1 = j_1 = 1$, so that $L_{i_1}y = L_{j_1}y = y'$. We see that (A) is not satisfied, since for any a and any $c > a$, $y'' = 0$, $y'(a) = y'(c) = 0$ has nontrivial solutions. If $0 < b-a < \pi$, $p_1(t) \equiv 0$, $p_2(t) \equiv 1$, then $u(t) \equiv 1$ is a nontrivial solution of $Lu + p_1(t)u = 0$, $(L_1u)(a) = (L_1u)(b) = 0$, but $Ly + p_2(t)y = 0$, $(L_1y)(a) = (L_1y)(c) = 0$ has no nontrivial solution for $a < c \leqq b$.

As a byproduct of our methods used to prove Theorem 2 we shall obtain a simple comparison theorem concerning an eigenvalue problem for the equation

$$Ly + \lambda p(t)y = 0. \tag{6}$$

We shall illustrate this comparison theorem with an example from fluid mechanics.

We make use of some known results concerning (6) which are summarized in

THEOREM 3. *If $p(t)$ is continuous on $[a, b]$, if $p(t)$ does not vanish identically on any subinterval of $[a, b]$, if* (A) *holds, and if $(-1)^{n-k}p(t) \leqq 0$ for all $t \in [a, b]$, then for any $c \in (a, b]$ there exists a unique number $\lambda_0(c) > 0$ such that when $\lambda = \lambda_0(c)$,* (6) *has a solution y which is strictly positive on (a, c) and which satisfies the boundary conditions* (2). *Moreover, $\lambda_0(c)$ is continuous in c.*

The first statement follows from results of Elias [4, pp. 34, 54]. It also follows as a special case of a result of Karlin [7, § 4] and the theory of oscillation kernels. The second statement follows from a result of [4, Corollary 5] which actually shows that $\lambda_0(c)$ is continuously differentiable on $(a, b]$.

The number $\lambda_0(c)$ is the smallest eigenvalue of (6) with the boundary conditions (2).

Let f be continuous on $[a, b]$. A maximal closed subinterval of $[a, b]$, which may consist of a single point, on which f is identically zero will be called a *zero component of f*. We say that *f changes sign h times on a subinterval* of $[a, b]$ if there exist $h+1$ points $x_1 < x_2 < \cdots < x_{h+1}$ belonging to this subinterval such that $f(x_j)f(x_{j+1}) < 0$ for $j = 1, \cdots, h$. We say *f changes sign* on a subinterval if f changes sign at least once on the subinterval.

LEMMA 1. *Let $f_1, f_2, \cdots, f_{m+1}$, $m \geqq 1$, be continuous on $[a, b]$, and suppose that for $1 \leqq h \leqq m$ the function f_h does not vanish identically on $[a, b]$. Suppose also that for all h, $1 \leqq h \leqq m$, f_{h+1} changes sign on the open interval between any two zero components of f_h.*

(I) *If $m+s$ of the numbers $f_1(a), f_2(a), \cdots, f_m(a), f_1(b), f_2(b), \cdots, f_m(b)$ are zero, where $s \geqq 1$, then f_{m+1} changes sign at least s times on (a, b).*

(II) *If $m \geqq 2$, if f_1 and f_2 have a common zero component contained in (a, b) and if $m-2+s$ of the $2m$ numbers $f_j(a)$, $j = 1, \cdots, m$, and $f_j(b)$, $j = 1, \cdots, m$ are zero with $s \geqq 1$, then f_{m+1} changes sign at least s times on (a, b).*

Rolle's theorem implies that if $r_0, r_1, \cdots, r_n$ are as above and if $f \in C^n[a, b]$, then the functions $f_h = L_{h-1}f$, $h = 1, \cdots, n+1$, will satisfy the first conditions of Lemma 1 provided that $f_{n+1} = Lf$ does not vanish identically. In this case Lemma 1 follows from Lemma 2 of [4]. In the general case it is easy to prove Lemma 1 by induction, starting with $m = 1$ for part (I), and starting with $m = 2$ for part (II).

LEMMA 2. *Suppose that $p(t)$ does not vanish identically on any subinterval of $[a, c]$, and that y is a nontrivial solution of (1) which satisfies the boundary conditions (2). If $0 \leqq h \leqq n-1$, then $L_h y$ must vanish at some point of $[a, c]$.*

Proof. Suppose on the contrary that for some l satisfying $0 \leqq l \leqq n-1$, $(L_l y)(t) \neq 0$ for all $t \in [a, c]$. If $l = n-1$, let $f_h(t) = (L_{h-1}y)(t)$ for $t \in [a, c]$ and $h = 1, \cdots, n$. If $0 \leqq l < n-1$, let $f_h(t) = (L_{l+h}y)(t)$ for $1 \leqq h \leqq n-1-l$, and $f_h(t) = (L_{h-n+l}y)(t)$ for $n-l \leqq h \leqq n$. In either case the hypotheses of the lemma imply that for $h = 1, \cdots, n-1$, f_{h+1} changes sign on the open interval between any two zero components of f_h, and that f_h does not vanish identically on $[a, b]$ for $h = 1, 2, \cdots, n$. Since $f_n(t) = L_l(t) \neq 0$ for all $t \in [a, c]$, the boundary conditions (2) imply that at least n of the numbers $f_1(a), \cdots, f_{n-1}(a)$, $f_1(b), \cdots, f_{n-1}(b)$ are equal to zero. Therefore, according to Lemma 1, $f_n(t) = L_l(t)$ must change sign on $[a, c]$, which is a contradiction. This contradiction proves the lemma.

LEMMA 3. *If the hypotheses of Theorem 3 hold, then* $\lim_{c \to a+} \lambda_0(c) + \infty$.

Proof. Assume the contrary; it follows that, since $\lambda_0(c) > 0$ for $c \in [a, b]$, there exists a sequence of numbers $\{c_m\}_1^\infty$ such that $c_m > a$ for all m, $\lim_{m \to \infty} c_m = a$, and the sequence $\{\lambda_0(c_m)\}_1^\infty$ is bounded. By considering a suitable subsequence we may assume, without loss of generality, that $\lim_{m \to \infty} \lambda_0(c_m) = \lambda^* \geqq 0$. For each $m \geqq 1$, let y_m be a nontrivial solution of the boundary value problem $Ly + \lambda_0(c_m)p(t)y = 0$, $(L_{i_s}y)(a) = 0$, $s = 1, 2, \cdots, k$, $(L_{j_t}y)(c_m) = 0$, $t = 1, 2, \cdots, n-k$. We may assume that y_m is defined on $[a, b]$ and by multiplying y_m by a suitable positive constant we may suppose that $\sum_{h=0}^{n-1} (L_h y_m(b))^2 = 1$. By compactness of the unit sphere in $\mathbb{R}^n$ we may assume, by considering a suitable subsequence, the existence of numbers $d_1, d_2, \cdots, d_n$ such that

$$\begin{aligned} &\lim_{m \to \infty} (L_h y_m)(b) = d_{h+1}, \qquad h = 0, \cdots, n-1, \\ &\sum_{h=1}^{n} d_h^2 = 1. \end{aligned} \tag{7}$$

If $m \geqq 1$ and we set $x_h = L_{h-1}y_m$, $h = 1, \cdots, n$, then

$$x_h' = \frac{1}{r_h(t)} x_{h+1}, \qquad h = 1, \cdots, n-1,$$

$$x_n' = -\frac{\lambda_0(c_m)p(t)}{r_n(t)r_0(t)} x_1(t).$$

Since $\lambda_0(c_m) \to \lambda^*$ as $m \to \infty$, it follows from (7) and the standard theory of the continuity

of solutions of linear differential systems with respect to parameters and initial conditions, that if $z_1, \cdots, z_n$ are the functions which satisfy the system of differential equations

$$
\begin{aligned}
z_h' &= \frac{1}{r_h(t)} z_{h+1}, \qquad h = 1, \cdots, n-1, \\
z_n' &= -\frac{\lambda^* p(t)}{r_n(t) r_0(t)} z_1,
\end{aligned} \tag{8}
$$

and the initial conditions

$$z_h(b) = d_h, \qquad h = 1, 2, \cdots, n, \tag{9}$$

then

$$\lim_{m \to \infty} (L_h y_m)(t) = z_{h+1}(t), \qquad h = 0, \cdots, n-1, \tag{10}$$

uniformly on $[a, b]$. Now the conditions satisfied by y_m at a and at c_m and Lemma 2 imply the existence of numbers $\xi_{0m}, \cdots, \xi_{(n-1)m}$ such that $\xi_{hm} \in [a, c_m]$ for $h = 0, \cdots, n-1$ and $L_h y_m(\xi_{hm}) = 0$. Since (10) holds uniformly on $[a, b]$, and since $\lim_{m \to \infty} \xi_{hm} = a$, it follows that $z_h(a) = 0$ for $h = 1, \cdots, n$. But since $x_h \equiv 0$ is a solution of $x_h' = (1/r_h) x_{h+1}$ for $h = 1, \cdots, n-1$, $x_n' = -(\lambda^* p / r_n r_0) x_1$, it follows from the uniqueness theorem for linear differential systems that $z_h(t) \equiv 0$ for $h = 1, \cdots, n$. As this contradicts (9) and (7), the lemma is proved.

LEMMA 4. *If* (A) *holds, if* p_1 *and* p_2 *are continuous on* $[a, c]$ *with* $(-1)^{n-k} p_2(t) \leqq (-1)^{n-k} p_1(t) \leqq 0$ *for all* $t \in [a, c]$, *if for* $m = 1, 2$,

$$Lu_m + p_m(t) u_m = 0, \qquad t \in [a, c], \tag{11}$$

$$(L_{i_s} u_m)(a) = 0, \qquad s = 1, \cdots, k, \tag{12}$$

$$(L_{j_t} u_m)(c) = 0, \qquad t = 1, \cdots, n-k, \tag{13}$$

if $u_2(t) > 0$ *for* $t \in (a, c)$, *and* $u_1(t) \not\equiv 0$ *on* $[a, c]$, *then* $p_1 \equiv p_2$ *and there exists a number* γ *such that* $u_1 \equiv \gamma u_2$.

Proof. By replacing u_1 by $-u_1$, if necessary, we may assume $u_1(\bar{t}) > 0$ for some $\bar{t} \in (a, c)$. Let q be the smallest integer such that $0 \leqq q \leqq n-1$ and $(L_q u_2)(a) \neq 0$, and let r be the smallest integer such that $0 \leqq r \leqq n-1$ and $(L_r u_2)(c) \neq 0$. Since $u_2(t) > 0$ on (a, c), it follows that $(L_q u_2)(a) > 0$ and $(-1)^r (L_r u_2)(c) > 0$. Since $L_h u_1(a) = L_h u_2(a) = 0$ if $0 \leqq h < q$, and $L_h u_2(c) = L_h u_1(c) = 0$ if $0 \leqq h < r$, there exist numbers $\alpha_1 > 0$ and $\delta > 0$ such that $L_q(u_2 - \alpha u_1)(a) > 0$, $(-1)^r L_r(u_2 - \alpha u_1)(c) > 0$ and $u_2(t) - \alpha u_1(t) > 0$ for $t \in (a, a+\delta)$ and $t \in (c-\delta, c)$, provided that $0 \leqq \alpha \leqq \alpha_1$. Since u_2 is positive on the compact interval $[a+\delta, c-\delta]$, there exists a number $\alpha_2 > 0$ such that $u_2(t) - \alpha u_1(t) > 0$ if $t \in [a+\delta, c-\delta]$ and $0 \leqq \alpha \leqq \alpha_2$. Therefore, if $\alpha_3 = \min\{\alpha_1, \alpha_2\} > 0$, we have for $0 \leqq \alpha \leqq \alpha_3$

$$L_q(u_2 - \alpha u_1)(a) > 0, \tag{14}$$

$$(-1)^r L_r(u_2 - \alpha u_1)(c) > 0, \tag{15}$$

and

$$u_2(t) - \alpha u_1(t) > 0 \quad \text{for all } t \in (a, c). \tag{16}$$

Since $u_1(\bar{t}) > 0$, (14), (15), and (16) cannot hold for all $\alpha > 0$. Consequently, there exists $\gamma > 0$ such that (14), (15), and (16) hold if $\alpha < \gamma$ but at least one of the inequalities fails

to hold for $\alpha > \gamma$. Clearly

$$u_2(t) - \gamma u_1(t) \geqq 0, \qquad t \in (a, c). \tag{17}$$

We claim that either

$$L_q(u_2 - \gamma u_1)(a) = 0, \tag{18a}$$

$$L_r(u_2 - \gamma u_1)(c) = 0, \tag{18b}$$

or

$$u_2(t_0) - \gamma u_1(t_0) = 0 \quad \text{for some } t_0 \in (a, c). \tag{18c}$$

Indeed, if none of (18a), (18b) or (18c) held, then, by applying the same reasoning that was applied to u_2 and u_1 to $w \equiv u_2 - \gamma u_1$ and u_1, it would follow that (14), (15), and (16) would hold for values of α larger than γ. From the hypotheses of the lemma and from (17) it follows that

$$(-1)^{n-k}(Lw)(t) = (-1)^{n-k}[-p_1(t)w(t) + (p_1(t) - p_2(t))u_2(t)] \geqq 0, \tag{19}$$

for all $t \in (a, c)$.

We claim that $w(t) = u_2(t) - \gamma u_1(t) = 0$ for all $t \in [a, c]$. Assume the contrary; it follows that Lw does not vanish identically on $[a, c]$ because of assumption (A). Consequently, $L_h w$ does not vanish identically on $[a, c]$ for $h = 0, 1, \cdots, n-1$, so the functions $f_h = L_{h-1}w$, $h = 0, \cdots, n$, fulfill the conditions in Lemma 1. If either (18a) or (18b) held, it would follow by (12) and (13) that, since $q \neq i_s$, $1 \leqq s \leqq k$, and $r \neq j_t$, $1 \leqq t \leqq n-k$, at least $n+1$ of the numbers

$$(L_0 w)(a), \cdots, (L_{n-1}w)(a), \qquad (L_0 w)(b), \cdots, (L_{n-1}w)(b)$$

would be zero. Thus, by Lemma 1, $L_n w = Lw$ would change sign on (a, c), contradicting (19). Therefore, both (18a) and (18b) are impossible, so (18c) must hold for some $t_0 \in (a, c)$. If either $w(t) = 0$ for all $t \in [t_0, c]$, or $w(t) = 0$ for all $t \in [a, t_0]$, then either $(L_h w)(a) = 0$ for $0 \leqq h \leqq n-1$ or $(L_h w)(b) = 0$ for $0 \leqq h \leqq n-1$. Since $1 \leqq k \leqq n-1$, it would follow from (12) and (13) that least $n+1$ of the numbers

$$(L_0 w)(a), \cdots, (L_{n-1}w)(a), \qquad (L_0 w)(b), \cdots, (L_{n-1}w)(b)$$

would be zero. By the reasoning used above, this gives a contradiction. Hence, (18c) implies that w has a zero component contained in (a, c) and since, according to (17), $w(t) \geqq 0$ on (a, c), we see that $L_0 w = r_0 w$ and $L_1 w = r_1(L_0 w)'$ have a common zero component contained in (a, c). Since, by (14) and (15), at least $n = n - 2 + 2$ of the numbers

$$(L_0 w)(a), \cdots, (L_{n-1}w)(a), \qquad (L_0 w)(b), \cdots, (L_{n-1}w)(b)$$

are zero, part (II) of Lemma 1 implies that $L_n w = Lw$ changes sign twice on (a, c), which contradicts (19).

This contradiction shows that $w(t) = u_2(t) - \gamma u_1(t) \equiv 0$ on $[a, c]$. Hence, $Lw \equiv 0$ on $[a, c]$ and, since $u_2(t) > 0$ on (a, c), it follows from (19) that $p_1(t) \equiv p_2(t)$ on $[a, c]$. This proves the lemma.

As an application of Lemma 4 we prove a simple comparison theorem for the eigenvalue problem (6).

THEOREM 4. *Let $p(t)$ and $q(t)$ be continuous on $[a, b]$ and let*

$$(-1)^{n-k}q(t) \leqq (-1)^{n-k}p(t) \leqq 0, \qquad t \in [a, b]. \tag{20}$$

Suppose that $p(t)$ does not vanish identically on any subinterval of $[a, b]$. If $p(t) \not\equiv q(t)$ on $[a, b]$, and $\mu_0(c)$ has the same meaning relative to the differential equation $Ly + \mu q(t)y = 0$ as $\lambda_0(c)$ has relative to (6) *in Theorem* 3, *then $\mu_0(b) < \lambda_0(b)$.*

Proof. According to the definition of $\lambda_0(b)$ and $\mu_0(b)$ there exists a solution v of $Lv + \lambda_0(b)p(t)v = 0$ which satisfies the boundary conditions (2) for $b = c$ with $v(t) > 0$ on (a, b). Similarly, there exists a solution $w(t)$ of $Lw + \mu_0(b)q(t)w = 0$ satisfying the same boundary conditions with $w(t) > 0$ on (a, b). If, contrary to the assertion of the lemma, $0 < \lambda_0(b) \leqq \mu_0(b)$ then, by (20), $(-1)^{n-k}\mu_0(b)q(t) \leqq (-1)^{n-k}\lambda_0(b)p(t) \leqq 0$ for all $t \in [a, b]$. Applying Lemma 4 with $p_2(t) = \mu_0(b)q(t)$ and $p_1(t) = \lambda_0(b)p(t)$, it would follow that $\lambda_0(b)p(t) \equiv \mu_0(b)q(t)$ for all $t \in [a, b]$, which contradicts (20), the assumed inequality $0 < \lambda_0(b) \leqq \mu_0(b)$, and the hypothesis $p \not\equiv q$. Thus, $0 < \mu_0(b) < \lambda_0(b)$, and the lemma is proved.

Remark. If L and the boundary conditions (2) are self-adjoint, Theorem 4 is a consequence of a well-known variational characterization of λ_0 in terms of Rayleigh quotients.

Example. The eigenvalue problem given by the differential equation (6) and the boundary conditions (2) includes as a special case an eigenvalue problem which occurs frequently in the study of stability and bifurcation of fluid motions. This is the eigenvalue problem

$$\begin{aligned} &Lu + \lambda p(t)v = 0, \\ &L^2v - \lambda q(t)u = 0, \\ &u(a) = u(b) = v(a) = v(b) = v'(a) = v'(b) = 0, \end{aligned} \tag{21}$$

an excellent discussion of which is given by Joseph in [6, vol. 1, pp. 251–255]. Here L is a second-order iterated operator of the type considered above, and p and q are assumed to be positive and sufficiently differentiable. We note that (21) is equivalent to either of the two problems

$$\begin{aligned} &Mu + \lambda^2 q(t)u = 0, \\ &(M_0u)(a) = (M_2u)(a) = (M_3u)(a) = 0, \\ &(M_0u)(b) = (M_2u)(b) = (M_3u)(b) = 0, \end{aligned} \tag{22}$$

where $Mu = L^2p^{-1}Lu$; or

$$\begin{aligned} &Nv + \lambda^2 p(t)v = 0, \\ &(N_0v)(a) = (N_1v)(a) = (N_4v)(a) = 0, \\ &(N_0v)(b) = (N_1v)(b) = (N_4v)(b) = 0, \end{aligned} \tag{23}$$

where $Nv = Lq^{-1}L^2v$.

Since the boundary conditions for both problems satisfy the hypothesis of Theorem 1, and since λ_0 is the smallest eigenvalue of (21) if and only if λ_0^2 is the smallest eigenvalue of either (22) or (23), it follows from Theorem 4 that the smallest eigenvalue of (21) decreases (increases) if either $p(t)$ or $q(t)$ are increased (decreased). This fact has been suggested by numerical studies such as the one given by Chandrasekhar [2, pp. 298–305] that concerns the problem

$$\begin{aligned} &(D^2 - a^2)^2u = (1 + \alpha t)v, \\ &(D^2 - a^2)v = -Ta^2u, \end{aligned} \tag{24}$$

$$u(0) = (Du)(0) = v(0) = 0,$$
$$u(1) = (Du)(1) = v(1) = 0. \tag{25}$$

Here $D = d/dt$, α and a are constants, and T is to be determined. Since

$$Ly = (D^2 - a^2)y = \frac{1}{\cosh at}\frac{d}{dt}\left(\cosh^2 at \frac{d}{dt}\left(\frac{y}{\cosh at}\right)\right)$$

is an iterated operator of the type considered, and since the substitutions $T = \lambda^2$, $w = \lambda u$ transform the system (24) into $Lv + \lambda a^2 w = 0$, $L^2 w - \lambda(1+\alpha t)v = 0$, the problem (24), (25) is equivalent to the form (21) if $\alpha > -1$. The numerical study in [2] shows, among other things, that the smallest eigenvalue T_c of (24), (25) increases as $\mu \equiv 1 + \alpha$ decreases.

Proof of Theorem 2. For each integer $m \geqq 1$ let

$$Q_m(t) = p_2(t) - \frac{(1)^{n-k}}{m}, \tag{26}$$

and let $\Lambda_{m0}(c) > 0$ denote the number for which there exists a solution of the differential equation

$$Ly + \Lambda_{m0}(c) Q_m(t) y = 0, \tag{27}$$

such that y satisfies the boundary conditions (2) and y is positive on (a, c). Since $(-1)^{n-k} Q_r(t) < (-1)^{n-k} Q_s(t)$ if $r < s$ and $t \in [a, b]$, it follows from Theorem 4 that

$$\Lambda_{r0}(c) < \Lambda_{s0}(c), \tag{28}$$

if $r < s$ and $c \in (a, b]$. If for some $m \geqq 1$, $\Lambda_{m0}(b) \geqq 1$, then (3) and (26) would imply that

$$(-1)^{n-k} \Lambda_{m0}(b) Q_m(t) < (-1)^{n-k} p_1(t). \tag{29}$$

Since (4) has a nontrivial solution, and since the boundary value problem

$$Ly + \Lambda_{m0}(b) Q_m(t) y = 0,$$
$$(L_{i_s} y)(a) = 0, \qquad s = 1, \cdots, k,$$
$$(L_{j_t} y)(b) = 0, \qquad t = 1, \cdots, n-k,$$

has a solution which is positive on (a, b), (29) and Lemma 4 would imply that $\Lambda_{m0}(b) Q_m(t) \equiv p_1(t)$ which is absurd. Therefore,

$$\Lambda_{m0}(b) < 1, \qquad m \geqq 1. \tag{30}$$

Since $\lim_{c \to a+} \Lambda_{10}(c) = +\infty$, there exists c_1 with $a < c_1 < b$ such that $\Lambda_{10}(c_1) = 1$. Suppose that for some $r \geqq 1$ we have shown the existence of numbers $c_1 < \cdots < c_r$ with $a < c_h < b$ such that $\Lambda_{h0}(c_h) = 1$ for $h = 1, \cdots, r$. Since $\Lambda_{(r+1)0}(b) < 1$ and, according to (28), $1 = \Lambda_{r0}(c_r) < \Lambda_{(r+1)0}(c_r)$, it follows, by continuity of $\Lambda_{(r+1)0}$, that there exists c_{r+1}, with $c_r < c_{r+1} < b$, such that $\Lambda_{(r+1)0}(c_{r+1}) = 1$. Thus, there exists a sequence $\{c_m\}_1^\infty$ such that for all $m \geqq 1$

$$a < c_m < b, \qquad c_m < c_{m+1}, \qquad \Lambda_{m0}(c_m) = 1. \tag{31}$$

Let $c = \lim_{m \to \infty} c_m \leqq b$. By (31) and the definition of Λ_{m0}, there exists for each $m \geqq 1$ a solution v_m of

$$Lv_m + Q_m(t) v_m = 0, \tag{32}$$

such that

$$v_m(t) > 0, \qquad t \in (a, c_m), \tag{33}$$

and

$$\begin{aligned} &(L_{i_s} v_m)(a) = 0, \qquad s = 1, \cdots, k, \\ &(L_{t_j} v_m)(c_m) = 0, \qquad t = 1, \cdots, n-k. \end{aligned} \tag{34}$$

By multiplying each v_m by a suitable positive constant we may assume that $\sum_{h=0}^{n-1} (L_h v_m)(a)^2 = 1$ for all m. Thus by compactness, there exists a subsequence $\{v_{m_l}\}_{l=1}^{\infty}$ of $\{v_m\}_1^{\infty}$ and numbers $d_0, \cdots, d_{n-1}$ such that $\sum_{h=0}^{n-1} d_h^2 = 1$ and $\lim_{l\to\infty} (L_h v_{m_l})(a) = d_h$ for $h = 0, \cdots, n-1$. Since, by (26), $\lim_{l\to\infty} Q_{m_l}(t) = p_2(t)$ uniformly on $[a, b]$, it follows by the same reasoning used to prove Lemma 2, that if v denotes the solution of (5) such that $(L_h v)(a) = d_h$ for $h = 0, \cdots, n-1$, then

$$\lim_{l\to\infty} (L_h v_{m_l})(t) = (L_h v)(t) \quad \text{uniformly on } [a, b]. \tag{35}$$

Since (34) implies that $(L_{i_s} v)(a) = 0$, $s = 1, \cdots, k$ and $(L_{j_t} v)(c) = \lim_{l\to\infty} (L_{j_t} v_{m_l})(c_{m_l}) = 0$, $t = 1, \cdots, n-k$, and since v is a nontrivial solution of (5), the proof of Theorem 2 is complete if $c < b$.

We show that $c = b$ is impossible. Assume the contrary; if $a < t < b$ then $t < c_{m_l}$ for l sufficiently large, so by (35) for $h = 0$ and (33), $v(t) = \lim_{l\to\infty} v_{m_l}(t) \geqq 0$. Hence,

$$(-1)^{n-k}(Lv)(t) = -(-1)^{n-k} p_2(t) v(t) \geqq 0 \quad \text{for all } t \in [a, b]. \tag{36}$$

By the uniqueness theorem, the zeros of v on (a, b) are isolated. If $v(t_0) = 0$ for some $t_0 \in (a, b)$, then, since $v(t) \geqq 0$, $v'(t_0) = 0$. Hence, $\{t_0\}$ would be a common zero component of $L_0 v$ and $L_1 v$, contained in (a, b). Since v satisfies the boundary conditions (2) for $b = c$, at least $n = n - 2 + 2$ of the numbers

$$(L_0 v)(a), \cdots, (L_{n-1} v)(a), \qquad (L_0 v)(b), \cdots, (L_{n-1} v)(b)$$

are zero, so part (II) of Lemma 1 would imply the existence of at least two sign changes of $L_n v = Lv$ on (a, b), contradicting (36). Thus, if $b = c$, $v(t) > 0$ on (a, b).

Since the boundary value problem (4) has a nontrivial solution, and since v is a positive solution of (5) satisfying the conditions (2) with $c = b$, the inequality (3) and Lemma 4 imply that $p_1(t) \equiv p_2(t)$ on $[a, b]$. Since this contradicts an assumption of Theorem 2, $b = c$ is impossible. Hence, $c < b$ so, by an earlier remark, the proof is complete.

Remark. The reasoning used in the proof of Theorem 2 actually shows the existence of a solution of (5), satisfying conditions (2), which is positive on (a, c). Lemma 4 with $p_1 = p_2$ shows that this solution is unique up to constant multiples.

REFERENCES

[1] S. AHMAD AND A. C. LAZER, *On n-th order Sturmian theory*, J. Differential Equations, 35 (1980), pp. 87–112.

[2] S. CHANDRASEKHAR, *Hydrodynamic and Hydromagnetic Stability*, Oxford University Press, Oxford, 1961.

[3] U. ELIAS, *The extremal solutions of the equation $Ly + p(x)y = 0$, II*, J. Math. Anal. Appl., 55 (1976), pp. 253–265.

[4] ———, *Eigenvalue problems for the equation $Ly + \lambda p(x)y = 0$*, J. Differential Equations, 29 (1978), pp. 28–57.

[5] G. JOHNSON, *The k-th conjugate point function for an even order linear differential equation*, Proc. Amer. Math. Soc., 42 (1974), pp. 563–568.
[6] D. D. JOSEPH, *Stability of Fluid Motions*, vol. 1, Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[7] S. KARLIN, *Total positivity, interpolation by splines, and Green's functions of differential operators*, J. Approx. Theory, 4 (1971), pp. 91–112.
[8] A. JU. LEVIN, *Distribution of the zeros of solutions of a linear differential equation*, Dokl. Akad. Nauk SSSR, 156 (1964), pp. 1281–1284; transl. in Soviet Math. Dokl., 5 (1964), pp. 818–821.
[9] Z. NEHARI, *Disconjugate linear differential operators*, Trans. Amer. Math. Soc., 129 (1969), pp. 500–516.
[10] K. SCHMITT, *Boundary value problems and comparison theorems for ordinary differential equations*, SIAM J. Appl. Math., 26 (1974), pp. 670–678.
[11] U. ELIAS, *Oscillatory solutions and extremal points for a linear differential equation*, Arch. Rat. Mech. Anal., 71 (1979), pp. 177–198.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1120-0002 $01.00/0

AN EXTENSION OF THE ENESTRÖM-KAKEYA THEOREM AND ITS SHARPNESS*

N. ANDERSON,† E. B. SAFF‡ AND R. S. VARGA¶

Abstract. The classical Eneström-Kakeya Theorem, for obtaining bounds for the moduli of the zeros of any polynomial with positive coefficients, is extended to the case of any complex polynomial having no zeros on the ray $[0, +\infty)$. It is shown that this extension is sharp in the sense that, given such a complex polynomial $p_n(z)$ of degree $n \geq 1$, a sequence of polynomials $\{Q_{m_i}(z)\}_{i=1}^{\infty}$ can be found for which the classical Eneström-Kakeya Theorem, applied to the products $Q_{m_i}(z)p_n(z)$, yields, in the limit as $i \to \infty$, the maximum of the moduli of the zeros of $p_n(z)$.

A computational algorithm, based on linear programming, is also described whereby nearly "optimal" multiplying polynomials $Q_{m_i}(z)$ can be computed.

1. Introduction. With π_n denoting the set of all complex polynomials of degree exactly n, and with

$$\pi_n^+ := \{p_n(z) = \sum_{j=0}^{n} a_j z^j : a_j > 0 \text{ for all } j = 0, 1, \cdots, n\}, \tag{1.1}$$

a useful form of the classical Eneström-Kakeya Theorem [4], [13], due in fact to Eneström [4], is the following:

THEOREM A. *For any $p_n(z) = \sum_{j=0}^{n} a_j z^j$ in π_n^+ with $n \geq 1$, define*

$$\alpha = \alpha[p_n] := \min_{0 \leq i < n} \left\{ \frac{a_i}{a_{i+1}} \right\}, \qquad \beta = \beta[p_n] := \max_{0 \leq i < n} \left\{ \frac{a_i}{a_{i+1}} \right\}. \tag{1.2}$$

Then, all the zeros of $p_n(z)$ lie in the annulus

$$\alpha \leq |z| \leq \beta. \tag{1.3}$$

Evidently, if

$$\rho(p_n) := \max \{|z_j| : p_n(z_j) = 0\} \tag{1.4}$$

denotes the *spectral radius* of any complex polynomial $p_n(z)$ of degree at least unity, then it follows from (1.3) of Theorem A that

$$\beta[p_n] \geq \rho(p_n) \quad \forall p_n(z) \in \pi_n^+, \quad \forall n \geq 1. \tag{1.5}$$

Naturally, it is of interest to know when the inequality of (1.5) is sharp. This was first studied by Hurwitz [11], and the following result of [1] is a corrected form of Hurwitz's original contribution. (A similar result can be analogously obtained for the sharpness of $\alpha[p_n]$ in estimating the minimum of the moduli of the zeros of $p_n(z)$; see [1].)

THEOREM B. *For any $p_n(z) = \sum_{j=0}^{n} a_j z^j$ in π_n^+ with $n \geq 1$, define*

$$\bar{S} = \bar{S}[p_n] := \{j = 1, 2, \cdots, n+1 : \beta a_{n+1-j} - a_{n-j} > 0\}, \quad \textit{where } a_{-1} := 0, \tag{1.6}$$

* Received by the editors August 24, 1979, and in final revised form March 24, 1980.

† Department of Mathematics, Kent State University, Kent, Ohio 44242. The research of this author was supported in part by the National Science Foundation.

‡ Department of Mathematics, University of South Florida, Tampa, Florida 33620. The research of this author was supported in part by the U.S. Air Force Office of Scientific Research.

¶ Department of Mathematics, Kent State University, Kent, Ohio 44242. The research of this author was supported in part by the U.S. Air Force Office of Scientific Research and by the U.S. Department of Energy.

and

$$\bar{k} = \bar{k}[p_n] := \text{g.c.d.}\,\{j \in \bar{S}\}. \tag{1.7}$$

Then, equality in (1.5) *is valid iff* $\bar{k} > 1$. *If* $\bar{k} > 1$, *the zeros of* $p_n(z)$ *on* $|z| = \beta$ *are all simple, and are precisely given by*

$$\beta \exp\{2\pi ij/\bar{k} : j = 1, 2, \cdots, \bar{k} - 1\}. \tag{1.8}$$

Moreover, $p_n(z)$ *has the form*

$$p_n(\beta z) = \{1 + z + z^2 + \cdots + z^{\bar{k}-1}\} q_m(z^{\bar{k}}), \tag{1.9}$$

where $q_m \in \pi_m^+$. *If* $m \geqq 1$, *all the zeros of* $q_m(w)$ *lie in* $|w| < 1$, *and* $\beta[q_m] \leqq 1$.

Now, the Eneström-Kakeya upper bound $\beta[p_n]$ for $\rho(p_n)$ from (1.5) is certainly an easy quantity to compute. But, it suffers from two serious deficiencies. First, this upper bound can be applied only to the rather limited set of polynomials $\bigcup_{n=1}^{\infty} \pi_n^+$. For example, it cannot be applied as such to the particular polynomial $f_1(z) = 1 + z^2$. Second, the upper bound $\beta[p_n]$ may be a poor estimate of $\rho(p_n)$, and it is not apparent how this situation can be improved. For example, if $f_2(z) = 1 + \varepsilon z + z^2$ where $0 < \varepsilon \leqq 1$, we find that $\beta[f_2] = \varepsilon^{-1}$, which is a crude upper bound for $\rho(f_2) = 1$, when ε is small.

To explain our approach of generalizing the Eneström-Kakeya Theorem, note in the first example above that if $Q_1(z) = 1 + z$, then the product $Q_1(z) \cdot f_1(z) = 1 + z + z^2 + z^3$ is an element of π_3^+. On applying Theorem A, we obtain that $\beta[Q_1 \cdot f_1] = 1 \geqq \rho(Q_1 f_1)$. Moreover, since $\rho(Q_1 f_1) \geqq \rho(f_1)$ from (1.4), then

$$1 = \beta[Q_1 f_1] \geqq \rho(f_1),$$

and this last inequality is sharp since $\rho(f_1) = 1$. Similarly, for the second example above we find that

$$\beta[Q_1 f_2] = 1 + \varepsilon > \rho(f_2) = 1,$$

this upper bound being a sharper estimate of $\rho(f_2) = 1$ than the classical Eneström-Kakeya bound ε^{-1}, when ε is small.

More generally, for some complex polynomial $p_n(z)$ in π_n with $n \geqq 1$, suppose that there is a nonnegative integer m and a multiplier polynomial $Q_m(z)$ in π_m such that $Q_m(z) \cdot p_n(z) \in \pi_{n+m}^+$. Then, on applying (1.5), we have $\beta[Q_m p_n] \geqq \rho(Q_m p_n) \geqq \rho(p_n)$, i.e.,

$$\beta[Q_m p_n] \geqq \rho(p_n), \tag{1.10}$$

and we call $\beta[Q_m p_n]$ a *generalized Eneström-Kakeya functional* for $p_n(z)$.

Several questions now arise, the first being to find the precise class of polynomials $p_n(z)$ for which the generalized Eneström-Kakeya functional is defined. This is answered in

PROPOSITION 1. *Given* $p_n(z) \in \pi_n$ *with* $n \geqq 1$, *there exists a nonnegative integer* m *and a* $Q_m(z) \in \pi_m$ *for which* $Q_m(z) \cdot p_n(z) \in \pi_{m+n}^+$ *iff* $p_n(z)$ *has no zeros on the ray* $[0, +\infty)$.

The proof of this result will be given in § 3. Because of Proposition 1, it is convenient then to set

$$\hat{\pi}_n := \{p_n(z) \in \pi_n : p_n(z) \text{ has no zeros on the ray } [0, +\infty)\} \quad \text{for } n \geqq 1. \tag{1.11}$$

The next results, aimed at the sharpness of the inequality of (1.10), are our main results. Their proofs are given in §§ 4 and 5.

THEOREM 1. *For each $p_n(z) \in \hat{\pi}_n$ with $n \geqq 1$, there exists a sequence of polynomials $\{Q_{m_i}(z)\}_{i=1}^{\infty}$, with $Q_{m_i}(z) \in \pi_{m_i}$ and with $Q_{m_i}(z) \cdot p_n(z) \in \pi^+_{m_i+n}$ for all $i \geqq 1$, such that*

$$\lim_{i \to \infty} \beta[Q_{m_i} p_n] = \rho(p_n). \tag{1.12}$$

In essence, Theorem 1 gives us that the generalized Eneström-Kakeya functional is asymptotically *sharp* in the sense of (1.12).

Another question that can be asked is to characterize those elements $p_n \in \hat{\pi}_n$ with $n \geqq 1$ for which equality holds in (1.10) for *some* polynomial $Q_m(z)$, as opposed to equality holding in the limit as in (1.12) of Theorem 1. This is answered in

THEOREM 2. *Given $p_n(z) \in \hat{\pi}_n$ with $n \geqq 1$, there exists a nonnegative integer m and a polynomial $Q_m(z)$ in π_m with $Q_m(z) \cdot p_n(z) \in \pi^+_{m+n}$, such that*

$$\beta[Q_m p_n] = \rho(p_n) \tag{1.13}$$

iff all of the following hold:

(1.14)
- (i) *All zeros of $p_n(z)$ of modulus $\rho(p_n)$ are simple.*
- (ii) *If $\{\zeta_j\}_{j=1}^{r}$ denotes the set of all zeros of $p_n(z)$ on the circle $|z| = \rho(p_n)$, then $\arg \zeta_j$ is a (nonzero) rational multiple of 2π, i.e., $\arg \zeta_j = 2\pi n_j/d_j$ (in lowest terms), where n_j and d_j are positive integers with $0 < n_j < d_j$ for all $j = 1, 2, \cdots, r$.*
- (iii) *If $D := \text{l.c.m.}\ \{d_j\}_{j=1}^{r}$, there is a positive integer σ such that, for every zero ζ of $p_n(z)$ with $|\zeta| < \rho(p_n)$; we have $\zeta^{\sigma D} \notin [0, +\infty)$.*

It is interesting to note that the motivation for Theorems 1 and 2 comes directly from Theorem B, in the sense that the polynomial $p_n(\beta z)$ of (1.9) of Theorem B is such that its zeros have a *ring-like character*; i.e., $p_n(\beta z)$ has $\bar{k} - 1$ zeros nearly uniformly distributed on $|z| = 1$, while its remaining zeros are distributed as the $\bar{k}$th roots of zeros of $q_m(w)$ (cf. (1.9)). This pattern persists, as we shall see, both in our examples as well as in the spirit of the proofs of Theorems 1 and 2.

In the next section, we show how linear programming techniques can be used to determine nearly "optimal" polynomial multipliers $Q_m(z)$ of a specific degree such that $Q_m(z) \cdot p_n(z) \in \pi^+_{m+n}$. In addition, the results of some numerical experiments will be given and discussed.

Because of the continuing interest in the classical Eneström-Kakeya Theorem and its many generalizations, we have gathered in the References a number of books and papers which deal in part with this topic, in the hope that such a list may be of value to the readers.

2. Optimization of the generalized Eneström-Kakeya functional. For any $p_n \in \hat{\pi}_n$, set

$$\omega_m(p_n) := \{Q_m(z) \in \pi_m : Q_m(z) \cdot p_n(z) \in \pi^+_{m+n}\} \quad \text{for any } m \geqq 0. \tag{2.1}$$

Note that $\omega_m(p_n)$ may be empty for a particular nonnegative integer m, but from Proposition 1, it follows there is a nonnegative integer m_0 such that $\omega_{m_0}(p_n) \neq \varnothing$. As is easily seen, $\omega_m(p_n) \neq \varnothing$ implies $\omega_{m+k}(p_n) \neq \varnothing$ for every $k \geqq 1$. Thus, as a consequence of Proposition 1, there is a *least* nonnegative integer $\sigma(p_n)$ for each $p_n \in \hat{\pi}_n$ such that

$$\omega_m(p_n) \neq \varnothing, \quad \text{for all } m \geqq \sigma(p_n). \tag{2.2}$$

Note also that $\omega_m(p_n) \neq \varnothing$ implies that $\omega_m(p_n)$ is a *convex* subset of π_m; i.e., if $q_1(z)$ and $q_2(z)$ are in $\omega_m(p_n)$, then so is $\alpha q_1(z) + (1-\alpha) q_2(z)$ for all $0 \leqq \alpha \leqq 1$.

Given a $p_n \in \hat{\pi}_n$, and given that $\omega_m(p_n) \neq \varnothing$, it is of interest to determine computationally a nearly "optimum" element $\hat{Q}_m(z)$ in $\omega_m(p_n)$, i.e., one whose generalized Enestrom-Kakeya functional satisfies:

$$\beta[\hat{Q}_m p_n] \doteq \inf\{\beta[Q_m p_n]: Q_m \in \omega_m(p_n)\}.$$

This can be done by solving a sequence of linear programming subproblems, each of which consists of finding a so-called feasible solution [17, § 3.5] to a set of linear inequalities. (Such computational subproblems are usually solved using "Phase I" of the simplex method; see [17].) Specifically, for any fixed $p_n(z) = \sum_{j=0}^{n} a_j z^j$ in $\hat{\pi}_n$ (which we may take, without loss of generality, to be real), assume $\omega_m(p_n) \neq \varnothing$, and consider any real $Q_m(z) = \sum_{j=0}^{m} b_j z^j$. If we set

$$Q_m(z) \cdot p_n(z) = \sum_{j=0}^{m+n} \gamma_j z^j,$$

then $Q_m \in \omega_m(p_n)$ iff

$$\gamma_j > 0 \quad \text{for all } j = 0, 1, \cdots, m+n, \tag{2.3}$$

which is a system of linear inequalities in the b_j's, since $\gamma_j = \sum_{i=\max(0;\, j-m)}^{\min(j;\, n)} a_i b_{j-i}$, for $j = 0, 1, \cdots, m+n$. We then say that $(\tau, Q_m(z))$ is a *feasible point* for $\omega_m(p_n)$ (cf. Luenberger [17, p. 18]) if, in addition to (2.3),

$$\gamma_j \leqq \tau \gamma_{j+1} \quad \text{for all } j = 0, 1, \cdots, m+n-1. \tag{2.4}$$

By definition, if $(\tau, Q_m(z))$ is a feasible point for $\omega_m(p_n)$, then $Q_m \in \omega_m(p_n)$ and $\beta[Q_m p_n] \leqq \tau$. Note that Theorem 1 implies that given any $\tau > \rho(p_n)$, a feasible point $(\tau, Q_m(z))$ is guaranteed to exist for m sufficiently large.

On the other hand, fixing m and given a feasible point $(\tau, Q_m(z))$ for $\omega_m(p_n)$ we can proceed (see below) to determine computationally a *least* feasible point $(\tau_m, \hat{Q}_m(z))$ in $\omega_m(p_n)$, where

$$\tau_m := \inf\{\tau: (\tau, Q_m(z)) \text{ is a feasible point in } \omega_m(p_n) \text{ for some } Q_m(z) \in \pi_m\}. \tag{2.5}$$

Note that since $(\beta[Q_m p_n], Q_m(z))$ is, by definition, a feasible point for $\omega_m(p_n)$ for each $Q_m(z) \in \omega_m(p_m)$, it follows from (2.4) and (2.5) that

$$\tau_m = \inf\{\beta[Q_m p_n]: Q_m(z) \in \omega_m(p_n)\}. \tag{2.6}$$

Thus, our computational technique finds in essence an "optimal" multiplier polynomial in $\omega_m(p_n)$, if $\omega_m(p_n) \neq \varnothing$.

The computational experiments were carried out as follows. Given a $p_n(z) \in \hat{\pi}_n$,

1. Compute a $\tau^{(0)} > \rho(p_n)$ from the coefficients of p_n using some standard upper bound for $\rho(p_n)$ (see, e.g., [18]).

2. For $m := 1, 2, 3, \cdots$, use linear programming to attempt to find a feasible point $(\tau^{(0)}, Q_m(z))$. Call the first m for which success occurs m_0.

3. For $m := m_0, m_0+1, \cdots$, find "optimal" multipliers $\hat{Q}_m(z)$, for each fixed m, by using a bisection technique on the variable τ, with (2.3) and (2.4) holding. For example, given $(\tau^{(0)}, Q_{m_0}(z))$, try to find a feasible point for $\tau := \tau^{(0)}/2$; if this is not possible, try with $\tau := \frac{3}{4} \cdot \tau^{(0)}$; otherwise try $\tau := \tau^{(0)}/4$, etc.

In our computations, the actual testing for feasibility ("Phase I" of the simplex method) was done using the program in Wilkinson and Reinsch [25, p. 152].

Now, let m_i be the sequence of integers and $Q_{m_i}(z)$ the sequence of polynomials in Theorem 1. By (2.5), we can compute a sequence of polynomials $\hat{Q}_{m_i}$ satisfying

$$\beta(\hat{Q}_{m_i} p_n) = \tau_{m_i} + \varepsilon_i, \tag{2.7}$$

where the ε_i are positive quantities which can be chosen to satisfy $\lim_{i\to\infty} \varepsilon_i = 0$. But,

$$\rho(p_n) \leqq \beta(\hat{Q}_{m_i} p_n) \leqq \beta(Q_{m_i} p_n) + \varepsilon_i,$$

since from (2.6), $\tau_{m_i} \leqq \beta(Q_{m_i} p_n)$. Thus, taking limits and using (1.12),

$$\lim_{i\to\infty} \beta(\hat{Q}_{m_i} p_n) = \rho(p_n), \tag{2.8}$$

and hence (ignoring roundoff), the sequence of estimates provided by the computational algorithm is guaranteed to converge to $\rho(p_n)$.

Example 1. $p_6(z) = (z^3+1)^2 \in \hat{\pi}_6$.

For this polynomial, an optimum multiplier polynomial $\hat{Q}_{32}(z)$ was computed. Its zeros are shown in Fig. 1. The value of τ_{32} (cf. (2.6)) is 1.03626 to 5D. The coefficients γ_j in

$$\hat{Q}_{32}(z)p_6(z) := \sum_{j=0}^{38} \gamma_j z^j$$

satisfy $\gamma_j/\gamma_{j+1} = 1.03626$ for *all* $0 \leqq j \leqq 37$ except for $\gamma_{26}/\gamma_{27} = 0.944348$ and $\gamma_{32}/\gamma_{33} = 0.051895$. Two of the zeros of $\hat{Q}_{32}$ are roughly equal to the zeros of $z^2 + Rz + R^2$, where $R = 1.03626 = \tau_{32}$ (compare (4.7) in the proof of Theorem 1).

Note the circular pattern of the zeros of $\hat{Q}_{32}(z) \cdot p_6(z)$. This idea is used in the proof of Theorem 1, although the multiplier polynomials used there are not "optimal" at each stage. For example, using the technique of this proof on $p_6(z)$ above yields $\beta[Q_{48}p_6] = 2^{1/5} \approx 1.14870$, which is not as good as the result $\beta[\hat{Q}_{32}p_6] = 1.03626$ obtained from linear programming.

Example 2. $p_4(z) = (z^2 - \sqrt{3}z + 1)(z^2 + (\sqrt{2}/2)z + \frac{1}{4}) \in \hat{\pi}_4$.

For this polynomial (which is not in π_4^+), an optimum multiplier polynomial $\hat{Q}_{17}(z)$ was computed. Its zeros are shown in Fig. 2. Again, note the tendency of the optimal multiplier $\hat{Q}_{17}(z)$ to "fill out" the rings ($|z| = 1, |z| = \frac{1}{2}$) on which the zeros of the original polynomial lie.

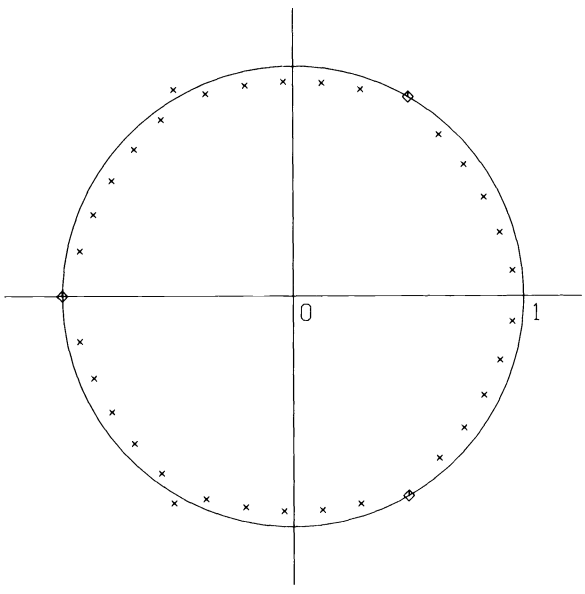


FIG. 1. *Zeros of* $p_6(z) = (z^3+1)^2$: *diamonds*; *zeros of optimal multiplier of degree* 32: *crosses* ($\tau_{32} \doteq 1.03626$).

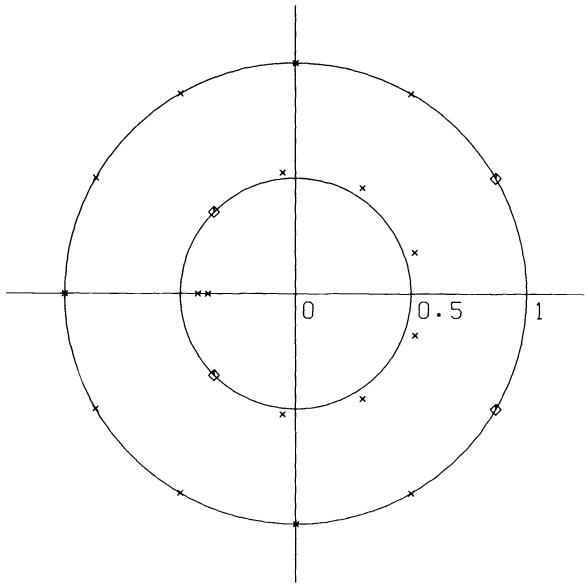


FIG. 2. *Zeros of* $p_4(z) = (z^2 - \sqrt{3}\, z + 1)(z^2 + (\sqrt{2}/2)z + \frac{1}{4})$: *diamonds*; *zeros of optimal multiplier of degree* 17: *crosses* ($\tau_{17} \doteq 1.00034$).

3. Proofs of Proposition 1 and lemmas. We begin with the

Proof of Proposition 1. First, assume that $p_n(z)$ is any polynomial which has no zeros on the ray $[0, +\infty)$. Without loss of generality, we may assume that $p_n(z)$ is monic and moreover real, for if $(z - \zeta)$ is a factor of $p_n(z)$ with ζ not real, then both ζ and $\bar{\zeta}$ are not contained in $[0, +\infty)$ and we may consider $p_n(z) \cdot (z - \bar{\zeta})$ in place of $p_n(z)$ if $(z - \bar{\zeta})$ is not a factor of $p_n(z)$. Hence, we can express $p_n(z)$, by hypothesis, as

$$p_n(z) = \prod_{i=1}^{\alpha_1} (z + \delta_i) \prod_{j=1}^{\alpha_2} (z - r_j e^{i\theta_j})(z - r_j e^{-i\theta_j}),$$

or equivalently

$$p_n(z) = \prod_{i=1}^{\alpha_1} (z + \delta_i) \prod_{j=1}^{\alpha_2} (z^2 - 2r_j \cos \theta_j z + r_j^2), \tag{3.1}$$

where $\delta_i > 0$ (if the first product is not vacuous), and where $r_j > 0$ and $0 < \theta_j < \pi$ (if the second product is not vacuous). If the second product is vacuous, then already $p_n \in \pi_n^+$. If the second product is not vacuous, consider the quadratic factor

$$z^2 - 2r_j \cos \theta_j z + r_j^2, \qquad r_j > 0,\ 0 < \theta_j < \pi.$$

If $\pi/2 < \theta_j < \pi$, this quadratic factor is an element of π_2^+. If not, this quadratic factor divides

$$(z^2 - 2r_j \cos \theta_j z + r_j^2)(z^2 + 2r_j \cos \theta_j z + r_j^2) = z^4 - 2r_j^2 \cos (2\theta_j) z^2 + r_j^4.$$

If $\pi/4 < \theta_j \leqq \pi/2$, this product, when multiplied by $(1 + z)$, is then a polynomial in π_5^+. If $0 < \theta_j \leqq \pi/4$, this process of doubling the argument θ_j can be continued, and eventually, since $\theta_j > 0$, one obtains in this manner an element in some π_ν^+. As this is true for each quadratic factor of (3.1), a polynomial multiplier can thus be found such that $Q_m(z) \cdot p_n(z) \in \pi_{m+n}^+$.

Conversely, supposing that $p_n(z)$ has a zero on the ray $[0, +\infty)$, the same is true for any product $Q_m(z) \cdot p_n(z)$, whence $Q_m(z) \cdot p_n(z) \notin \pi^+_{m+n}$ for any $Q_m(z)$. □

Before proceeding to the proof of Theorem 1 in § 4, we establish some results needed in the proof of that theorem.

LEMMA 1. *For any positive integer m, let $\{P_k(z)\}_{k=1}^m$ be any collection of m polynomials, each having positive coefficients and each being of degree at least unity. Then,*

$$\beta\Big[\prod_{k=1}^{m} P_k\Big] \leqq \sum_{k=1}^{m} \beta[P_k]. \tag{3.2}$$

Proof. The proof will be by induction on m. Obviously, (3.2) is valid for $m = 1$. Assume, then, that (3.2) is true for m, and consider any $(m+1)$ polynomials $\{P_k(z)\}_{k=1}^{m+1}$, each having positive coefficients and degree at least unity. Calling

$$Q(z) := \prod_{k=1}^{m} P_k(z) = \sum_{i=0}^{\gamma} a_i z^i \quad \text{and} \quad P_{m+1}(z) := \sum_{j=0}^{\lambda} b_j z^j, \tag{3.3}$$

and, noting that rearrangements of the P_k's have no effect in (3.2), we may assume that $\gamma \geqq \lambda$. On setting

$$Q(z) \cdot P_{m+1}(z) := \sum_{k=0}^{\gamma+\lambda} c_k z^k,$$

then from (3.3) and the hypotheses of this lemma, we obtain

$$c_k = \sum_{j=0}^{k} a_j b_{k-j} > 0, \quad \text{for } k = 0, 1, \cdots, \gamma+\lambda,$$

where $a_j := 0$ for all $j > \gamma$ and where $b_j := 0$ for all $j > \lambda$. With the inductive hypothesis that $\beta[Q] \leqq \sum_{k=1}^{m} \beta[P_k]$, we must show that

$$c_k/c_{k+1} \leqq \beta[Q] + \beta[P_{m+1}] \quad \text{for all } k = 0, 1, \cdots, \gamma+\lambda-1. \tag{3.4}$$

This is done by considering the three cases: $0 \leqq k < \lambda$, $\lambda \leqq k < \gamma$, and $\gamma \leqq k \leqq \gamma+\lambda-1$. Since the proofs of the cases are similar, we consider for brevity only the case $\gamma \leqq k \leqq \gamma+\lambda-1$. In this case, we have

$$\begin{aligned}
c_k/c_{k+1} &= \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big) \Big/ \Big(\sum_{j=0}^{k+1} a_j b_{k+1-j}\Big) = \Big(\sum_{j=k-\lambda}^{\gamma} a_j b_{k-j}\Big) \Big/ \Big(\sum_{j=k+1-\lambda}^{\gamma} a_j b_{k+1-j}\Big) \\
&= \Big(a_{k-\lambda} b_\lambda + \sum_{j=k+1-\lambda}^{\gamma} a_j b_{k-j}\Big) \Big/ \Big(\sum_{j=k+1-\lambda}^{\gamma} a_j b_{k+1-j}\Big) \\
&\leqq \Big(a_{k-\lambda} b_\lambda + \sum_{j=k+1-\lambda}^{\gamma} a_j \beta[P_{m+1}] b_{k+1-j}\Big) \Big/ \Big(\sum_{j=k+1-\lambda}^{\gamma} a_j b_{k+1-j}\Big) \\
&\leqq \frac{a_{k-\lambda} b_\lambda}{a_{k+1-\lambda} b_\lambda} + \beta[P_{m+1}] \leqq \beta[Q] + \beta[P_{m+1}] \leqq \sum_{k=1}^{m+1} \beta[P_k],
\end{aligned}$$

the last inequality making use of the inductive hypothesis $\beta[Q] \leqq \sum_{k=1}^{m} \beta[P_k]$, which completes the proof. □

Next, we establish a needed polynomial perturbation result.

LEMMA 2. *Given any complex number $re^{i\theta}$ with*

$$0 < \theta < \pi \quad \text{and} \quad r > 0, \tag{3.5}$$

and given any sequence $\{s_\nu\}_{\nu=1}^{\infty}$ *of positive integers such that*

$$s_\nu \geqq 2 \quad \forall \nu \geqq 1, \quad \textit{and} \quad \lim_{\nu\to\infty} s_\nu = +\infty, \tag{3.6}$$

there exists a sequence of monic polynomials $\{\tilde{P}_{s_\nu}(z)\}_{\nu=1}^{\infty}$ *with* $\tilde{P}_{s_\nu} \in \pi_{s_\nu}^{+}$ *for all* ν *sufficiently large such that*

$$\begin{cases} \text{(i)} & (z-re^{i\theta})(z-re^{-i\theta}) \textit{ divides } \tilde{P}_{s_\nu}(z) \textit{ for all } \nu \geqq 1, \textit{ and} \\ \text{(ii)} & \lim\limits_{\nu\to\infty} \beta[\tilde{P}_{s_\nu}] = r. \end{cases} \tag{3.7}$$

Proof. First, if we can show that $\{\tilde{P}_{s_\nu}(z)\}_{\nu=1}^{\infty}$ satisfies (3.7) for the special case $r=1$, then $\tilde{P}_{s_\nu}(z/r)$ will satisfy (3.7) for the general case. Thus, assuming $r=1$, we define the real monic polynomials $P_{s_\nu}(z)$ by

$$P_{s_\nu}(z) := 1 + z + \cdots + z^{s_\nu} = (z^{s_\nu+1}-1)/(z-1), \tag{3.8}$$

whose simple zeros are $\zeta_k(s_\nu) := \exp[2\pi ki/(s_\nu+1)]$, $k=1, 2, \cdots, s_\nu$. For each ν sufficiently large, choose the distinct zero, say $\zeta_{k_1}(s_\nu)$, of $P_{s_\nu}(z)$ which best approximates $e^{i\theta}$ in the upper-half complex plane. It is geometrically clear that

$$|\zeta_{k_1}(s_\nu) - e^{i\theta}| \leqq \frac{\pi}{s_\nu+1}, \tag{3.9}$$

for all ν sufficiently large, say $\nu \geqq \nu_0$. With this choice, define

$$\tilde{P}_{s_\nu}(z) := P_{s_\nu}(z) \frac{(z-e^{i\theta})(z-e^{-i\theta})}{(z-\zeta_{k_1}(s_\nu))(z-\bar{\zeta}_{k_1}(s_\nu))} \quad \forall \nu \geqq \nu_0, \tag{3.10}$$

so that $\tilde{P}_{s_\nu}(z)$ is a monic polynomial in π_{s_ν}, and $(z-e^{i\theta})(z-e^{-i\theta})$ divides $\tilde{P}_{s_\nu}(z)$. In essence, the pair of simple zeros $\zeta_{k_1}(s_\nu)$ and $\bar{\zeta}_{k_1}(s_\nu)$ is "perturbed" to the two simple zeros $e^{\pm i\theta}$ of $\tilde{P}_{s_\nu}(z)$. Expressing $\tilde{P}_{s_\nu}(z)$ in the form

$$\tilde{P}_{s_\nu}(z) = \sum_{i=0}^{s_\nu} a_i(s_\nu;\theta) z^i \quad \forall \nu \geqq \nu_0, \tag{3.11}$$

and writing $a_i = a_i(s_\nu;\theta)$ and $\arg \zeta_{k_1}(s_\nu) =: \psi$, we obtain on cross-multiplying in (3.10) that

$$(z^2 - 2\cos\psi z + 1)\tilde{P}_{s_\nu}(z) = (z^2 - 2\cos\theta z + 1)P_{s_\nu}(z). \tag{3.12}$$

On equating coefficients of z^j on both sides, we have (cf. (3.11))

$$a_{j-2} - 2\cos\psi a_{j-1} + a_j = 2(1-\cos\theta), \qquad j = 1, 2, \cdots, s_\nu, \tag{3.13}$$

where $a_0 = 1 =: a_{-1}$. We write:

$$b_j := a_j - 1, \qquad j = -1, 0, \cdots, s_\nu; \tag{3.14}$$

then (3.13) becomes

$$b_{j-2} - 2\cos\psi b_{j-1} + b_j = 2(\cos\psi - \cos\theta), \qquad j = -1, 0, \cdots, s_\nu, \tag{3.15}$$

where $b_0 = b_{-1} = 0$. The solution of this linear difference equation can be verified to be

$$b_j = \frac{[\cos\psi - \cos\theta]\{\sin\psi + \sin j\psi - \sin(j+1)\psi\}}{(1-\cos\psi)\sin\psi}, \qquad -1 \leqq j \leqq s_\nu, \tag{3.16}$$

so that

$$|b_j| \leqq \frac{3|\cos\psi - \cos\theta|}{(1-\cos\psi)\sin\psi}, \qquad -1 \leqq j \leqq s_\nu. \tag{3.17}$$

From our definition in (3.9) of $\zeta_{k_1}(s_\nu)$ and from (3.5), it then follows that there is a constant M, dependent only on θ, such that

$$|b_j| \leqq M/s_\nu \quad \text{for all } -1 \leqq j \leqq s_\nu, \text{ and for all } s_\nu \geqq 2. \tag{3.18}$$

Recalling (3.14), we see that (3.18) implies that $|a_j - 1| \leqq M/s_\nu$, from which it follows that $\tilde{P}_{s_\nu}(z) \in \pi^+_{s_\nu}$ for all ν sufficiently large, as well as (cf. (3.7 (ii))) $\lim_{\nu\to\infty} \beta[\tilde{P}_{s_\nu}] = 1$. □

4. Proof of Theorem 1. Consider (cf. (1.11)) any $p_n(z) \in \hat{\pi}_n$ with $n \geqq 1$, and assume, without loss of generality, that $p_n(z)$ is monic, real, and is normalized so that (cf. (1.4))

$$\rho(p_n) = 1.$$

Writing

$$p_n(z) = \prod_{i=1}^{n} (z - \zeta_i), \tag{4.1}$$

so that $|\zeta_i| \leqq 1$ for all i, we define

$$P_t(w) := \prod_{i=1}^{n} (w - \zeta_i^t), \tag{4.2}$$

where t is any positive integer, and set

$$S_i := \{t \in \mathbb{Z}_+ : \zeta_i^t \notin [0, +\infty)\}, \qquad 1 \leqq i \leqq n, \tag{4.3}$$

where $\mathbb{Z}_+$ denotes the set of all positive integers. Because $p_n(z) \in \hat{\pi}_n$, then $1 \in S_i$, and S_i is thus nonempty for all $1 \leqq i \leqq n$. Note that if some $\arg \zeta_i =: \theta_i$ is a rational multiple of π, i.e., (in lowest terms) $\theta_i = 2\pi\gamma/\delta$ where γ and δ are positive integers with $\gamma/\delta < 1$, then no multiple of δ is in S_i, while all $t \not\equiv 0 \pmod{\delta}$ are in S_i. In this case, it is evident that

$$S_i = \mathbb{Z}_+ \backslash \{m\delta\}_{m=1}^{\infty}.$$

On the other hand, if some $\arg \zeta_i$ is not a rational multiple of π, then $S_i = \mathbb{Z}_+$. Consequently, since $p_n(z)$ is a fixed polynomial in $\hat{\pi}_n$, then

$$\begin{cases} \text{(i)} \quad \bigcap_{i=1}^{n} S_i =: T = \{t_j\}_{j=1}^{\infty} \subset \mathbb{Z}_+, \quad \text{and} \\ \text{(ii)} \quad 1 = t_1 < t_2 < t_3 \cdots, \quad \text{with } \lim_{j\to\infty} t_j = +\infty. \end{cases} \tag{4.4}$$

We claim now that for each $t_j \in T$, there exists a polynomial $G_j(w)$ such that

$$\begin{cases} \text{(i)} \quad G_j(w) \text{ is monic and has positive coefficients for all } j \geqq 1, \\ \text{(ii)} \quad P_{t_j}(w) \text{ of (4.2) divides } G_j(w) \text{ for all } j \geqq 1, \text{ and} \\ \text{(iii)} \quad \beta[G_j] \leqq n, \end{cases} \tag{4.5}$$

where n is the degree of $p_n(z)$ in (4.1). To see this, consider from (4.2) any factor $(w - \zeta_i^{t_j})$ of $P_{t_j}(w)$, where $t_j \in T$. If $\zeta_i^{t_j}$ is real, i.e., $\arg \zeta_i^{t_j} = \pi$, then this factor is just $(w + |\zeta_i^{t_j}|)$, since T can contain only odd integers in this case, and moreover,

$\beta[(w+|\zeta_i^{t_j}|)]=|\zeta_i^{t_j}|\leqq 1$. If $\zeta_i^{t_j}$ is not real, the reality of the polynomial $P_{t_j}(w)$ gives us that the product

$$(w-\zeta_i^{t_j})(w-(\bar{\zeta}_i)^{t_j}) \tag{4.6}$$

divides $P_{t_j}(w)$, where we may assume that $0<\arg\zeta_i^{t_j}<\pi$. Applying Lemma 2 to the product of (4.6) gives a polynomial $\tilde{P}_{i,j}(w)$ having (4.6) as a factor, such that $\tilde{P}_{i,j}(w)$ has positive coefficients, and such that $\beta[\tilde{P}_{i,j}]\leqq 2$. Thus, multiplying all these $\tilde{P}_{i,j}(w)$'s together, thereby forming $G_j(w)$, gives that $G_j(w)$ is monic with positive coefficients. Applying Lemma 1 to the product defining $G_j(w)$ gives $\beta[G_j(w)]\leqq n$, and by construction, $P_{t_j}(w)$ of (4.2) divides $G_j(w)$, thereby establishing (4.5).

Next, for each $R>0$, form the product

$$H_j(z;R):=\{R^{t_j-1}+R^{t_j-2}z+\cdots+z^{t_j-1}\}G_j(z^{t_j}), \tag{4.7}$$

for each $t_j\in T$, where $G_j(w)$ satisfies (4.5). Because $G_j(w)$ has all positive coefficients from (4.5 (i)), the polynomial $H_j(z;R)$ defined in (4.7) similarly has all positive coefficients, and the Eneström-Kakeya functional β of (1.2) can be directly applied to it. Note that the given $p_n(z)$ in $\hat{\pi}_n$ divides $H_j(z;R)$ (cf. (4.2) and (4.5 (ii)) for each choice of $R>0$. Now, it can be easily verified that

$$\beta[H_j(z;R)]=\max\{R;\Gamma_j/R^{t_j-1}\}, \tag{4.8}$$

where $\Gamma_j:=\beta[G_j]$. On equating R and Γ_j/R^{t_j-1}, i.e., on setting $R_j:=\Gamma_j^{1/t_j}$, we obtain from (4.8) that

$$\beta[H_j(z;R_j)]=\Gamma_j^{1/t_j}=(\beta[G_j])^{1/t_j},\quad \forall t_j\in T. \tag{4.9}$$

To complete the proof of Theorem 1, it thus remains from (4.9) to show that

$$\lim_{j\to\infty}(\beta[G_j])^{1/t_j}=1. \tag{4.10}$$

Since $p_n(z)$ divides $H_j(z;R_j)$ and since, by normalization, $\rho(p_n)=1$, then from (4.9) and (4.5 (iii)),

$$1=\rho(p_n)\leqq\beta[H_j(z;R_j)]=(\beta[G_j])^{1/t_j}\leqq(n)^{1/t_j},$$

which yields, by way of (4.4 (ii)), the desired result of (4.10). □

5. Proof of Theorem 2. First, assume that for $p_n(z)\in\hat{\pi}_n$ with $n\geqq 1$, there exists a polynomial $Q_m(z)$ in π_m with $Q_m(z)\cdot p_n(z)\in\pi^+_{m+n}$, such that $\beta[Q_mp_n]=\rho(p_n)$. Without loss of generality, we may, as in § 4, normalize to the case $\rho(p_n)=1$, i.e.,

$$\beta[Q_mp_n]=\rho(p_n)=1.$$

From (1.5), $\beta[Q_mp_n]\geqq\rho(Q_mp_n)\geqq\rho(p_n)=1$, so that

$$\beta[Q_mp_n]=\rho(Q_mp_n)=\rho(p_n)=1.$$

Hence, from Theorem B, all zeros of $Q_m(z)\cdot p_n(z)$ on $|z|=1$ are necessarily simple, which establishes the necessity of (1.14 (i)) of Theorem 2. Next, again from Theorem B, there is a positive integer $\bar{k}>1$ such that the zeros of $Q_m(z)\cdot p_n(z)$ on $|z|=1$ are precisely of the form

$$\exp\{2\pi ij/\bar{k}: j=1,2,\cdots,\bar{k}-1\}. \tag{5.1}$$

Evidently, each zero of $p_n(z)$ (as well as each zero of $Q_m(z)$) on $|z|=1$ is a (nonzero) rational multiple of 2π; i.e., if $\{\zeta_j\}_{j=1}^r$ denotes the set of all zeros of $p_n(z)$ on $|z|=1$, then

$$\arg \zeta_j = 2\pi n_j/d_j \quad \text{(in lowest terms)}, \tag{5.2}$$

where n_j and d_j are positive integers with $0 < n_j < d_j$ for all $j = 1, 2, \cdots, r$, thereby establishing the necessity of (1.14 (ii)) of Theorem 2.

Next, again from Theorem B, we have that, for some nonnegative integer l,

$$Q_m(z) \cdot p_n(z) = (1 + z + z^2 + \cdots + z^{\bar{k}-1}) g_l(z^{\bar{k}}), \tag{5.3}$$

where $g_l(w) \in \pi_l^+$, and if $l \geqq 1$, all zeros of $g_l(w)$ lie in $|w| < 1$, and $\beta[g_l] \leqq 1$. Clearly, the zeros of $p_n(z)$ on $|z| = 1$ must be of the form (5.1), so that, for suitable integers ν_j,

$$\frac{n_j}{d_j} = \frac{\nu_j}{\bar{k}}, \qquad j = 1, 2, \cdots, r.$$

Thus, if $D := \text{l.c.m.}\ \{d_j\}_{j=1}^r$, then D divides $\bar{k}$, whence $\bar{k} = \sigma D$ for some positive integer σ. Now, consider any zero ζ of $p_n(z)$ with $|\zeta| < 1$. Evidently, $\zeta^{\bar{k}}$ is a zero of $g_l(w)$ from (5.3). But, since $g_l(w) \in \pi_l^+$, then $\zeta^{\bar{k}} = \zeta^{\sigma D}$ could not be contained in $[0, +\infty)$, which establishes the necessity of (1.14 (iii)).

Conversely, assume that $p_n \in \hat{\pi}_n$ with $n \geqq 1$, that $\rho(p_n) = 1$, and that (1.14) is valid. Defining $\tilde{p}_{n-r}(z) := p_n(z)/\prod_{j=1}^r (z - \zeta_j) = \prod_{j=1}^{n-r} (z - \mu_j)$, where again $\{\zeta_j\}_{j=1}^r$ is the set of all zeros of $p_n(z)$ on $|z| = 1$, then either $\tilde{p}_{n-r}(z)$ is a nonzero constant, or $\tilde{p}_{n-r}(z)$ is a polynomial of degree $n - r \geqq 1$, all of whose zeros lie in $|z| < 1$. In the former case, hypothesis (1.14 (iii)) holds vacuously, while in the latter case, hypothesis (1.14 (iii)) implies that $\hat{p}_{n-r}(W) := \prod_{j=1}^{n-r} (W - \mu_j^{\bar{k}})$ is an element of $\hat{\pi}_{n-r}$. Now, applying Theorem 1 to $\hat{p}_{n-r}(W) \in \hat{\pi}_{n-r}$, shows that there exists a sequence of polynomials $\{Q_i(W)\}_{i=1}^{\infty}$ such that

$$\begin{cases} \text{(i)} & Q_i(W) \cdot \hat{p}_{n-r}(W) \text{ has positive coefficients for all } i \geqq 1, \text{ and} \\ \text{(ii)} & \lim\limits_{i \to \infty} \beta[Q_i \cdot \hat{p}_{n-r}] = \rho(\hat{p}_{n-r}(W)) < 1. \end{cases} \tag{5.4}$$

To fix matters, choose from this sequence $\{Q_i(W)\}_{i=1}^{\infty}$ the polynomial $\tilde{Q}(W)$ of *least* degree such that

$$\begin{cases} \text{(i)} & \tilde{Q}(W) \cdot \hat{p}_{n-r}(W) \text{ has positive coefficients, and} \\ \text{(ii)} & \beta[\tilde{Q} \cdot \hat{p}_{n-r}] \leqq 1, \end{cases} \tag{5.5}$$

and set $g(W) := \tilde{Q}(W)\hat{p}_{n-r}(W)$.

Since $\prod_{j=1}^r (z - \zeta_j)$ divides, from (1.14 (ii)), the polynomial

$$1 + z + z^2 + \cdots + z^{\bar{k}-1}, \quad \text{where } \bar{k} = \sigma D,$$

it follows that there is a polynomial multiplier, $Q_m(z)$, such that

$$Q_m(z) \cdot p_n(z) = \{1 + z + \cdots + z^{\bar{k}-1}\} g(z^{\bar{k}}), \tag{5.6}$$

where $g(W)$, from (5.5), has positive coefficients, $\beta[g] \leqq 1$, and all zeros of $g(W)$ lie in $|z| \leqq 1$. By this construction, one directly verifies that

$$\beta[Q_m p_n] = 1 = \rho(p_n), \tag{5.7}$$

which establishes (1.13). □

We remark that the construction in the proof of the sufficiency of conditions (1.14) of Theorem 2 gives the multiplier polynomial $Q_m(z)$ of *least* degree for satisfying (5.7).

To illustrate the result of Theorem 2, consider

Example 3. $p_3(z) := (1 + z^2)(\frac{1}{2} + z) \in \hat{\pi}_3$.

The zeros of $p_3(z)$ are $\pm i$, $-\frac{1}{2}$, so that $\rho(p_3)=1$. For the zeros on $|z|=1$, their arguments are $2\pi/4$ and $6\pi/4$, whence (cf. (1.14 (iii))) $D=4$. But, since $(-\frac{1}{2})^{4\sigma}\in[0,+\infty)$ for *every* positive integer σ, hypothesis (1.14 (iii)) fails, and it is not possible to find a polynomial multiplier $Q_m(z)$ for which (1.13) is valid.

Example 4. $p_4(z)=(1+z^2)(\frac{1}{4}+\frac{1}{2}z+z^2)\in\hat{\pi}_4$.

The zeros of $p_4(z)$ are $\pm i$, and $\frac{1}{2}e^{\pm 2\pi i/3}$. As in the previous case, the zeros on $|z|=1$ have arguments $2\pi/4$ and $6\pi/4$, whence $D=4$, and as $(\frac{1}{2}e^{\pm 2\pi i/3})^{4\sigma}\notin[0,+\infty)$ for $\sigma=1$, 2, 4, $\cdots$, we choose $\sigma=1$. In this example, the conditions of (1.14) of Theorem 2 are valid, and with $Q_5(z):=(1+z)(\frac{1}{4}-\frac{1}{2}z+z^2)(\frac{1}{16}-\frac{1}{4}z+z^2)$, then

$$Q_5(z)p_4(z)=(1+z+z^2+z^3)g_2(z^4),\quad \text{with } g_2(w)=\frac{1}{256}+\frac{w}{16}+w^2,$$

so that

$$\beta[Q_5p_4]=1=\rho(p_4).$$

Acknowledgment. We wish to thank Mr. Howard Fraser of Kent State University for having programmed the calculations in Examples 1 and 2. We also wish to thank the referee for bringing the excellent paper by G. Pólya [20] to our attention, and Professor Karl Zeller for his kind comments on the manuscript.

REFERENCES

[1] N. ANDERSON, E. B. SAFF AND R. S. VARGA, *On the Eneström-Kakeya Theorem and its sharpness*, Linear Algebra and Appl., 28 (1979), pp. 5–16.

[2] G. T. CARGO AND O. SHISHA, *Zeros of polynomials and fractional order differences of their coefficients*, J. Math. Anal. Appl., 7 (1963), pp. 176–182.

[3] J. EGERVÁRY, *On a generalization of a theorem of Kakeya*, Acta Sci. Math. (Szeged), 5 (1931), pp. 78–82.

[4] G. ENESTRÖM, *Härledning af en allmän formel för antalet pensionärer . . .*, Ofv. af Kungl. Vetenskaps-Akademiens Förhandlingar N:0 6, 1893, Stockholm.

[5] LEOPOLD FEJÉR, *Über ein trigonometrisches Analogon eines Kakayasches Satzen*, Jahresber. Deutsch. Math.-Verein, 38 (1929), pp. 231–238.

[6] N. K. GOVIL AND V. K. JAIN, *On the Eneström-Kakeya Theorem II*, J. Approx. Theory, 22 (1978), pp. 1–10.

[7] N. K. GOVIL AND Q. I. RAHMAN, *On the Eneström-Kakeya Theorem*, Tôhoku Math. J., 20 (1968), pp. 126–136.

[8] EMIL GROSSWALD, *Bessel Polynomials*, Lecture Notes in Mathematics 698, Springer-Verlag, New York, 1978 (p. 77).

[9] F. HEIGL, *Über die Abschätung der Wurzeln algebraischen Gleichungen*, Monatsh. Math., 62 (1958), pp. 16–55.

[10] PETER HENRICI, *Applied and Computational Complex Analysis*, Vol. I, John Wiley, New York, 1974 (p. 284).

[11] A. HURWITZ, *Über einen Satz des Herrn Kakeya*, Tôhoku Math. J., 4 (1913), pp. 89–93.

[12] A. JOYAL, G. LABELLE, AND Q. I. RAHMAN, *On the location of zeros of polynomials*, Canad. Math. Bull., 10 (1967), pp. 53–63.

[13] S. KAKEYA, *On the limits of the roots of an algebraic equation with positive coefficients*, Tôhoku Math. J., 2 (1912), pp. 140–142.

[14] AUBREY KEMPNER, *Extract of a letter to the Editor*, Tôhoku Math. J., 4 (1914), pp. 94–95.

[15] P. V. KRISHNAIAH, *On Kakeya's Theorem*, J. London Math. Soc., 30 (1955), pp. 314–319.

[16] STEPHAN LIPKA, *Zur Theorie der algebraischen Gleichungen mit positiven Koeffizienten*, Acta Sci. Math. (Szeged), 5 (1931), pp. 69–77.

[17] DAVID G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1965.

[18] MORRIS MARDEN, *Geometry of Polynomials*, Mathematical Surveys Number 3, American Mathematical Society, Providence, RI, 1966.

[19] A. M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973 (p. 99).
[20] GEORG PÓLYA, *Über die Nullstellen gewisser ganzer Funktionen*, Math. Z., 2 (1918), pp. 352–383.
[21] GEORG PÓLYA AND GÁBOR SZEGÖ, *Problems and Theorems in Analysis*, Vol. 1, Springer-Verlag, New York, 1972 (p. 107).
[22] ZALMAN RUBENSTEIN, *Some results in the location of zeros of polynomials*, Pacific J. Math., 15 (1965), pp. 1391–1395.
[23] STEPHAN RUSCHEWEYH, *On the Kakeya-Eneström Theorem and Gegenbauer polynomial sums*, this Journal, 9 (1978), pp. 682–686.
[24] MIODRAG TOMIČ, *Généralisation et démonstration géométrique de certains théorèmes de Fejér et Kakeya*, Acad. Serbe Sci. Publ. Inst. Math., 2 (1948), pp. 146–156.
[25] J. H. WILKINSON AND C. REINSCH, *Handbook for Automatic Computation, Vol. II, Linear Algebra*, Springer-Verlag, New York, 1971.
[26] K. ZELLER AND W. BEEKMANN, *Theorie der Limitierungsverfahren*, Springer-Verlag, Berlin–New York, 1970 (p. 126).

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1120-0003 $01.00/0

ALGEBRAIC METHOD FOR SOLVING SYSTEMS OF LINEAR DIFFERENTIAL EQUATIONS WITH VARIABLE COEFFICIENTS*

SERGE VASILACH†

Abstract. In our previous papers [SIAM J. Math Anal., 6(1975), pp. 295–311; 10(1979), pp. 586–602; 10(1979), pp. 1077–1088.] we have given an algebraic method for solving linear differential equations and partial differential equations with variable coefficients.

The present article is devoted to the application of the same method for solving systems of linear differential equations whose coefficients are functions of the independent variable t of the real line R_t.

The important result of this method consists in the fact that the required solution is obtained in an arbitrary neighborhood of a fixed point t_0 of R_t.

1.1. Introduction. In the present paper, an algebraic method is given for solving systems of linear differential equations with variable coefficients. For a similar method concerning linear differential equations and partial differential equations, see our previous articles [1], [2], [3]. Our method is based on the construction of a new class of distributions (cf. [4]), which permits us to define certain composition algebras as tensor products of distributions of this class, and to transform linear differential equations (respectively, partial differential equations) into algebraic composition equations (cf., for example, [1, Thm. 1, p. 305] and [2, § 4, Thm. 4.1]).

1.2. The composition algebras, $\mathscr{D}_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_\alpha)}$ and $\mathscr{D}'_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}_{(-\Gamma\alpha)}$. Let R_t (resp. R_α) be the real line with variable t (resp. α).

Let $(-\Gamma_t)$ (resp. $(+\Gamma_\alpha)$) be the cone of R_t (resp. R_α), defined by

$$(-\Gamma_t) = (-\infty, t], \quad (\text{resp. } (+\Gamma_\alpha) = [\alpha, +\infty)).$$

Let (cf. [5, Chapt. II, § 2]) $\mathscr{D}_{(+\Gamma_\alpha)}$ (resp. $\mathscr{D}_{(-\Gamma_t)}$) be the locally convex space of indefinitely differentiable functions with support limited to the left (resp. to the right) with respect to the variable $\alpha \in R_\alpha$ (resp. $t \in R_t$).

Let $\mathscr{D}'_{(+\Gamma_t)}$ (resp. $\mathscr{D}'_{(-\Gamma_\alpha)}$) be the strong dual of

$$\mathscr{D}_{(-\Gamma_1)} \quad (\text{resp. } \mathscr{D}_{(+\Gamma_\alpha)}).$$

Let $\mathscr{D}_{(-\Gamma_t)(+\Gamma_\alpha)}$ be the locally convex space of indefinitely differentiable functions with respect to the variables $(t, \alpha) \in R_t \times R_\alpha$, and with support limited to the left for $\alpha \in R_\alpha$ and to the right for $t \in R_t$ (cf. [4, § 2, pp. 5–7]).

Let $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ be the strong dual of $\mathscr{D}_{(-\Gamma_t)(+\Gamma_\alpha)}$ (cf. [4, § 4, pp. 7–9]). One has (cf. [4, § 3, No. 2, Thm. 4]) the kernel theorem,

$$\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)} = \mathscr{D}'_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_\alpha)}. \tag{1.1}$$

If we consider $\mathscr{D}_{(+\Gamma_t)}$ as a subspace of $\mathscr{D}'_{(+\Gamma_t)}$, equipped with the topology induced by $\mathscr{D}'_{(+\Gamma_t)}$, we have

$$\mathscr{D}_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_\alpha)} \subset \mathscr{D}'_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_\alpha)},$$

and $\mathscr{D}_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_\alpha)}$ is a composition algebra for the composition operation given by

$$S \circ T = \int_\alpha^t S(t, \xi) T(\xi, \alpha)\, d\xi. \tag{1.2}$$

* Received by the editors September 20, 1977, and in revised form May 28, 1980.
† Departement de Mathématiques, Université Laval, Quebec G1K 7P4, Canada.

More precisely, $\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)}$ is a composition algebra, non-commutative, with zero divisors and which has Dirac kernel $\delta(t-\alpha)$ as unit element for the composition (cf. [1, § 1.2]).

Similar definitions and properties hold for $\mathscr{D}_{(-\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(+\Gamma_\alpha)}$. Moreover, in [2, § 1.3], we have seen that $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ is a right (resp. left) module with respect to the composition algebra $\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)}$ (resp. $\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)}$).

Let $m(t)$ be a multiplication operator in $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ (cf. [1, § 3.10] and [2, § 3.5]). Then, we have

$$A\circ(m(t)T)=(Am(\alpha))\circ T. \tag{1.3}$$

Indeed, using (1.2) we obtain

$$A\circ(m(t)T)=\langle A(t,\xi),m(\xi)T(\xi,\alpha)\rangle=(Am(\alpha))\circ T.$$

1.3. Composition product and derivatives into $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$. Let $\delta(t-\alpha)$ be the Dirac kernel with respect to the variables t and α. It is known that $\delta(t-\alpha)=\delta(\alpha-t)$ and that $\delta(t-\alpha)$ belongs to the composition algebras: $\mathscr{D}'_t\hat{\otimes}\mathscr{D}_\alpha$, $\mathscr{D}_t\hat{\otimes}\mathscr{D}'_\alpha$, $\mathscr{E}_t\hat{\otimes}\mathscr{E}'_\alpha$, $\mathscr{E}'\hat{\otimes}\mathscr{E}_\alpha$, $\mathscr{S}'_t\hat{\otimes}\mathscr{S}_\alpha$, $\mathscr{S}_t\hat{\otimes}\mathscr{S}'_\alpha$, $\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)}$, $\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)}$ (cf. [1, § 1, Remark p. 296]). Further, one has for the derivative of order j:

$$\delta_t^{(j)}(t-\alpha)=(-1)^j\delta_\alpha^{(j)}(t-\alpha), \tag{1.4}$$

and

PROPOSITION 1.1. *For all $S(t,\alpha)\in\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ we have*

(cf. [2, § 3, Formulas (3.36) and (3.37)]):

$$\delta_t^{(j)}(t-\alpha)\circ S(t,\alpha)=\frac{\partial^j S}{\partial t^j}, \tag{1.5}$$

$$S(t,\alpha)\circ\delta_t^{(j)}(t-\alpha)=(-1)^j\frac{\partial^j S}{\partial\alpha^j} \tag{1.6}$$

$$\delta_t^{(j)}(t-\alpha)\circ S(t,\alpha)\circ\delta_t^{(k)}(t-\alpha)=(-1)^k\frac{\partial^{j+k}S}{\partial t^j\,\partial\alpha^k}. \tag{1.7}$$

Let

$$Y(t-\alpha)=\begin{cases}1 & \text{for } t\geqq\alpha,\\ 0 & \text{elsewhere,}\end{cases}$$

be the Heaviside kernel with respect to the variables t and α (cf. [1, § 3.2]).

Let $Y(t-\alpha)^j$ be the jth composition power of $Y(t-\alpha)$ given by (cf. [1, § 3.6, formula (3.7)]:

$$Y(t-\alpha)^j=\begin{cases}\dfrac{(t-\alpha)^{j-1}}{(j-1)!} & \text{for } r\geqq\alpha,\\ 0 & \text{elsewhere.}\end{cases}$$

For $S(t,\alpha)=Y(t-\alpha)$, (1.5) and (1.6) give us

$$\delta_t^{(j)}(t-\alpha)\circ Y(t-\alpha)=\delta_t^{(j-1)}(t-\alpha)=Y_t^{(j)}(t-\alpha)$$

and

$$Y(t-\alpha)\circ\delta_t^{(k)}(t-\alpha)=(-1)^k\delta_t^{(k-1)}(t-\alpha)=(-1)^kY_\alpha^{(k)}(t-\alpha)=Y_t^{(k)}(t-\alpha).$$

In particular, we obtain

$$Y(t-\alpha)\,\delta_t'(t-\alpha)=\delta_t'(t-\alpha)\circ Y(t,\alpha)=\delta(t-\alpha),$$

which shows us that $Y(t-\alpha)$ (resp. $\delta_t'(t-\alpha)$) is the *inverse* of $\delta_t'(t-\alpha)$ (resp. $Y(t-\alpha)$) for the composition product. Therefore, it is natural to write

$$\text{(1.8)}\qquad \begin{cases} Y^{(j)}(t-\alpha)=\delta^{(j-1)}(t-\alpha) \\ \text{and} \\ Y_t^{(-k)}(t-\alpha)=\delta_t^{(k-1)}(t-\alpha)=\left\{\dfrac{(t-\alpha)^{k-1}}{(k-1)}\right\}, \end{cases}$$

for any positive integers j and k. Under these conditions the following commutativity formula holds:

$$\text{(1.9)}\qquad \delta_t^{(j)}(t-\alpha)\circ\delta_t^{(k)}(t-\alpha)=\delta_t^{(k)}(t-\alpha)\circ\delta_t^{(j)}(t-\alpha),$$

for any positive or negative integers j and k.

1.4. Derivatives of the composition products into $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$. For

$$A\in\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\,\mathscr{D}_{(-\Gamma_\alpha)},$$

$$B\in\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)},$$

$$C\in\mathscr{D}_{(+\Gamma_t)}\,\hat{\otimes}\,\mathscr{D}'_{(-\Gamma_\alpha)},$$

the composition products $A\circ B$ and $B\circ C$ belong to $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$. Moreover, the equations (1.5), (1.6), (1.7) imply

$$\text{(1.10)}\qquad \begin{aligned} &\delta_t^{(j)}(t-\alpha)\circ A\circ\delta_t^{(k)}(t-\alpha)\circ B\circ\delta_t^{(1)}(t-\alpha) \\ &\quad=(-1)^k\frac{\partial^{j+k}A}{\partial t^j\partial\alpha^k}\circ(-1)^l\frac{\partial^1 B}{\partial\alpha^l}=\frac{\partial^j A}{\partial t^j}\circ(-1)^l\frac{\partial^{k+l}B}{\partial t^k\partial\alpha^l} \end{aligned}$$

for any positive or negative integers j, k, l. Similar formulas can be obtained for the composition product $B\circ C$.

2. Composition algebras of matrices.

2.1. Preliminaries. In [1, § 2] and [2, § 2.3] we have defined the composition algebras

$$(L^p_{\mathrm{loc}})_t(L^q_{\mathrm{loc}})_\alpha \quad \text{for } 1/p+1/q=1,\quad 1\leqq p\leqq\infty,\quad 1\leqq q\leqq\infty,$$

of (classes of) functions $f(t,\alpha)$ with pth locally integrable powers with respect to Lebesgue measure on R_t for each fixed $\alpha\in R_\alpha$ and with qth locally integrable powers with respect to Lebesgue measure on R_α for each fixed $t\in R$.

Also in [1, § 2.3] we have defined the composition algebra $\mathscr{C}_{t\alpha}^{(l,k)}$ of functions $f(t,\alpha)$ continuously differentiable of order $\leqq l$ with respect to the variable $t\in R_t$ and of order $\leqq m$ with respect to the variable $\alpha\in R_\alpha$. On the other hand, for an element f of $(L^p_{\mathrm{loc}})_t(L^q_{\mathrm{loc}})_\alpha$ (resp. $\mathscr{C}_{t\alpha}^{(l,m)}$) we have defined (cf. [1, § 3.3]), the *kernel-function* $\{f\}$, as an element of $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$, by setting

$$\{f\}=Y(t-\alpha)f(t,\alpha)=\begin{cases} f(t,\alpha) & \text{for } t\geqq\alpha, \\ 0 & \text{elsewhere.}\end{cases}$$

Further, for elements f, g of $(L^p_{\mathrm{loc}})_t(L^q_{\mathrm{loc}})_\alpha$ (resp. of $\mathscr{C}^{(1,m)}_{t\alpha}$) we have seen (cf. [1, § 2.4]) that the composition product $\{f\}\circ\{g\}$ is given by

$$\{f\}\circ\{g\}=\begin{cases}\displaystyle\int_\alpha^t f(t,\xi)g(\xi,\alpha)d\xi & \text{for } t\geqq\alpha,\\ 0 & \text{elsewhere.}\end{cases}$$

It is clear that

$$\{f\}\circ\{g\}=\{f\circ g\},\qquad \{f+g\}=\{f\}+\{g\},$$

and $\{\lambda f\}=\lambda\{f\}$, for λ a real or complex parameter.

Therefore, the set of kernel-functions corresponding to the elements of $(L^p_{\mathrm{loc}})_t(L^q_{\mathrm{loc}})_\alpha$ (resp. $\mathscr{C}^{(l,m)}_{t\alpha}$) is a composition algebra. We denote by $\{(L^p_{\mathrm{loc}})_t(L^q_{\mathrm{loc}})_\alpha\}$ (resp. $\{\mathscr{C}^{(l,m)}_{t\alpha}\}$) this composition algebra.

2.2. The composition algebras $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ and $\mathrm{Mat}_{n\times n}(\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)})$. In this section we will extend the composition product to matrices whose elements belong to the composition algebra $(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ (resp. to $(\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)})$. We denote by $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ the class of square matrices of order n whose elements belong to the composition algebra $\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$.

For

$$M=\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ \cdots & \cdots & \cdots & \cdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{pmatrix},$$

and

$$N=\begin{pmatrix}b_{11} & b_{12} & \cdots & b_{1n}\\ \cdots & \cdots & \cdots & \cdots\\ b_{n1} & b_{n2} & \cdots & b_{nn}\end{pmatrix},$$

elements of $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$, the composition product $M\circ N$ is given by

$$M\circ N=\begin{pmatrix}a_{11} & a_{12} & & a_{1n}\\ \cdots & \cdots & \cdots & \cdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{pmatrix}\circ\begin{pmatrix}b_{11} & & & b_{1n}\\ \cdots & \cdots & \cdots & \cdots\\ b_{n1} & b_{n2} & \cdots & b_{nn}\end{pmatrix}$$
$$=\begin{pmatrix}d_{11} & d_{12} & & d_{1n}\\ \cdots & \cdots & \cdots & \cdots\\ d_{n1} & d_{n2} & \cdots & d_{nn}\end{pmatrix},$$

where

$$(d_{ij}=\sum_{k=1}^{n}a_{ik}\circ b_{kj})_{1\leqq i\leqq n,1\leqq j\leqq n},$$

belong to $\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)}$. Therefore, $M\circ N$ belongs to $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$.

It is clear that $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ is a composition algebra.

In a similar way, one can show that $\mathrm{Mat}_{n\times n}(\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)})$ also is a composition algebra.

2.3. The composition bimodule $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)})$. We denote by $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ the class of square matrices of order n, whose elements belong to the composition bimodule $\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)}$.

It is easy to see that $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$ is a right (resp. left) module with respect to the composition algebra $\mathrm{Mat}_{n\times n}(\mathscr{D}_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}'_{(-\Gamma_\alpha)})$ (resp. $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)}\hat{\otimes}\mathscr{D}_{(-\Gamma_\alpha)})$).

PROPOSITION 2.1. *Let $Q(t)$ be a $n\times n$ matrix whose elements $(m_{ij}(t))$, $1\leqq i\leqq n$, $1\leqq j\leqq n$, are multiplication operators in $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$.*

Then,

$$M\circ(Q(t)N)=(MQ(\alpha)\circ N. \tag{2.1}$$

Proof. We have

$$Q(t)N=\begin{pmatrix} m_{11}(t) & \cdots & m_{1n}(t)\\ \cdots & \cdots & \cdots\\ m_{n1}(t) & \cdots & m_{nn}(t)\end{pmatrix}\begin{pmatrix} b_{11}(t,\alpha) & \cdots & b_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ b_{n1}(t,\alpha) & \cdots & b_{nn}(t,\alpha)\end{pmatrix}$$
$$=\begin{pmatrix} g_{11}(t,\alpha) & \cdots & g_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ g_{n1}(t,\alpha) & \cdots & b_{nn}(t,\alpha)\end{pmatrix},$$

where

$$g_{ij}(t,\alpha)=\sum_{k=1}^{n} m_{ik}(t)b_{kj}(t,\alpha), \tag{2.2}$$

whence

$$M\circ Q(t)N=\begin{pmatrix} a_{11}(t,\alpha) & \cdots & a_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ a_{n1}(t,\alpha) & \cdots & a_{nn}(t,\alpha)\end{pmatrix}\circ\begin{pmatrix} g_{11}(t,\alpha) & \cdots & g_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ g_{n1}(t,\alpha) & \cdots & g_{nn}(t,a)\end{pmatrix}$$
$$=\begin{pmatrix} h_{11}(t,\alpha) & \cdots & h_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ h_{n1}(t,\alpha) & \cdots & h_{nn}(t,\alpha)\end{pmatrix},$$

in which, by virtue of (2.2),

$$h_{ij}(t,\alpha)=\sum_{k=1}^{n} a_{ik}(t,\alpha)\circ g_{kj}(t,\alpha)$$
$$=\sum_{k=1}^{n} a_{ik}(t,\alpha)\circ\sum_{k_1=1}^{n} m_{kk_1}(t)b_{k_1j}(t,\alpha)$$
$$=\sum_{k_1=1}^{n}\Big(\sum_{k=1}^{n} a_{ik}(t,\alpha)\circ(m_{kk_1}(t)b_{k_1j}(t,\alpha))\Big).$$

But, (cf. [1, § 3.10, p. 304]) we have

$$a_{ik}(t,\alpha)\circ(m_{kk_1}(t)b_{k_1j}(t,\alpha))=(a_{ik}(t,\alpha)m_{kk_1}(\alpha))\circ b_{k_1j}(t,\alpha),$$

whence (2.1).

2.4. The Dirac matrix $I(t-\alpha)$. We say that the square matrix of order n

$$I(t-\alpha)=\begin{pmatrix}\delta(t-\alpha) & 0 & \cdots & 0\\ 0 & \delta(t-\alpha) & 0 & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0\cdots 0 & & \delta(t-\alpha)\end{pmatrix} \tag{2.3}$$

is the Dirac $n\times n$ matrix.

It is easy to see that $I(t-\alpha)$ is a unit element for the composition with the objects of $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)})$.

2.5. The Heaviside matrix $\mathscr{Y}(t-\alpha)$. We say that the square matrix of order n

$$\mathscr{Y}(t-\alpha)=\begin{pmatrix}Y(t-\alpha) & & 0 & \cdots & & 0\\ 0 & & Y(t-\alpha) & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & & o & \cdots & & Y(t-\alpha)\end{pmatrix} \tag{2.4}$$

is the Heaviside $n\times n$ matrix.

$$\mathscr{Y}(t-\alpha)^p=\begin{pmatrix}Y(t-\alpha)^p & 0 & \cdots & \cdots & \cdots & 0\\ 0 & Y(t-\alpha)^p & 0 & \cdots & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 0 & & Y(t-\alpha)^p\end{pmatrix} \tag{2.5}$$

is the pth composition power of $\mathscr{Y}(t-\alpha)$.

2.6. Derivatives of $I(t-\alpha)$. The derivative of order j, for $j\geqq 0$, is given by

$$I^{(j)}(t-\alpha)=\begin{pmatrix}\delta_t^{(j)}(t-\alpha) & 0 & \cdots & & 0\\ 0 & \delta_t^{(j)}(t-\alpha) & & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \\ 0 & \cdots & \cdots & 0 & \delta_t^{(j)}(t-\alpha)\end{pmatrix}. \tag{2.6}$$

A similar definition holds for the derivatives with respect to the variable α.

2.7. Derivatives of $\mathscr{Y}(t-\alpha)$. The derivative of order j, $j\geqq 0$, of $\mathscr{Y}(t-\alpha)$, with respect to the variable t, is given by

$$\mathscr{Y}_t^{(j)}(t-\alpha)=\begin{pmatrix}\delta_t^{(j-1)}(t-\alpha) & 0 & \cdots & & 0\\ 0 & \delta_t^{(j-1)}(t-\alpha) & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & \cdots & \cdots & 0 & \delta_t^{(j-1)}(t-\alpha)\end{pmatrix}. \tag{2.7}$$

A similar definition holds for the derivatives with respect to the variable α.

2.8. Composition algebras of kernel-matrices. Let $\{K\}$ be the composition algebra of kernel-functions which corresponds to the composition algebra K of functions of the variables t and α. Let $\mathrm{Mat}_{n\times n}(K)$ be the class of matrices whose elements belong to K. Then for

$$M(t,\alpha)=\begin{pmatrix}a_{11}(t,\alpha) & \cdots & a_{1n}(t,\alpha)\\ a_{21}(t,\alpha) & \cdots & a_{2n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ a_{n1}(t,\alpha) & \cdots & a_{nn}(t,\alpha)\end{pmatrix},$$

belonging to $\mathrm{Mat}_{n\times n}(K)$ we denote by $\{M\}$ the matrix which corresponds to M, given by

$$\{M\}=\begin{pmatrix}\{a_{11}(t,\alpha)\} & \cdots & \{a_{1n}(t,\alpha)\}\\ \{a_{21}(t,\alpha)\} & \cdots & \{a_{2n}(t,\alpha)\}\\ \cdots & \cdots & \cdots\\ \{a_{n1}(t,\alpha)\} & \cdots & \{a_{nn}(t,\alpha)\}\end{pmatrix},$$

$$=\left\{\begin{pmatrix}a_{11}(t,\alpha) & \cdots & a_{1n}(t,\alpha)\\ a_{21}(t,\alpha) & \cdots & a_{2n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ a_{n1}(t,\alpha) & \cdots & a_{nn}(t,\alpha)\end{pmatrix}\right\}. \tag{2.8}$$

Let

$$N=\begin{pmatrix}b_{11}(t,\alpha) & \cdots & b_{1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ b_{n1}(t,\alpha) & \cdots & b_{nn}(t,\alpha)\end{pmatrix}$$

be a second object of $\mathrm{Mat}_{n\times n}(K)$; one has

$$M\circ N=\begin{pmatrix}c_{11} & \cdots & \cdots & c_{1n}\\ & \cdots & \cdots & \\ c_{n1} & \cdots & \cdots & c_{nn}\end{pmatrix},$$

where

$$\begin{aligned}c_{ij}&=\sum_{k=1}^{n} a_{ik}\circ b_{kj}\\ &=\sum_{k=1}^{n}\int_{\alpha}^{t} a_{ik}(t,\xi)b_{kj}(\xi,\alpha)\,d\xi.\end{aligned} \tag{2.9}$$

Then we can write

$$M\circ N=\int_{\alpha}^{t} M(t,\xi)N(\xi,\alpha)\,d\xi\in\mathrm{Mat}_{n\times n}(K), \tag{2.10}$$

and

$$\{M\}\circ\{N\}=\left\{\int_{\alpha}^{t} M(t,\xi)N(\xi,\alpha)\,d\xi\right\}\in\mathrm{Mat}_{n\times n}(\{K\}). \tag{2.11}$$

Therefore $\mathrm{Mat}_{n\times n}(\{K\})$ is a composition algebra.

2.9. Derivatives of the composition products of matrices. Let M, N be arbitrary objects of $\mathrm{Mat}_{n\times n}(\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)})$. One has

$$I_t^{(j)}(t-\alpha)\circ M=\frac{\partial^j M}{\partial t^j}, \tag{2.12}$$

$$M\circ I_\alpha^{(k)}(t-\alpha)=(-1)^k\frac{\partial^k M}{\partial \alpha^k}, \tag{2.13}$$

and

$$
\begin{aligned}
& I_t^{(j)}(t-\alpha)\circ M\circ I_t^{(1)}(t-\alpha)\circ N\circ I_t^{(k)}(t-\alpha) \\
& \qquad = (-1)^l \frac{\partial^{j+1} M}{\partial t^j\,\partial\alpha^l}\circ(-1)^k\frac{\partial^k N}{\partial\alpha^k} \\
& \qquad = \frac{\partial^j M}{\partial t^j}(-1)^k\frac{\partial^{l+k} N}{\partial t^l\,\partial\alpha^k}.
\end{aligned} \tag{2.14}
$$

These formulas are a consequence of the formulas (1.5), (1.6), (1.7) and of the definition of the composition product of matrices.

Of course, the derivatives of kernel-matrices are given by the same formulas.

For the primitives of kernel-matrices one has

$$
\mathcal{Y}(t-\alpha)^p\circ\{M\} = \left\{\int_\alpha^t \frac{(t-\xi)^{p-1}}{(p-1)!}M(\xi,\alpha)\,d\xi\right\} \tag{2.15}
$$

and

$$
\{M\}\circ\mathcal{Y}(t-\alpha)^q = \left\{\int_\alpha^t M(t,\xi)\frac{(\xi-\alpha)^{q-1}}{(q-1)!}\,d\xi\right\}. \tag{2.16}
$$

3. Algebraic method for solving systems of linear differential equations with variable coefficients.

3.1. Fundamental matrices. Let us consider in $\mathcal{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ the system of linear equations

$$
\frac{dT_i}{dt} - \sum_{k=1}^n a_{ik}(t)T_k = S_i, \qquad 1\leqq i\leqq n, \tag{3.1}
$$

in which $(a_{ij}(t))_{1\leqq i\leqq n,1\leqq j\leqq n}$ are operators of multiplication and $(S_i)_{1\leqq i\leqq n}$ is a given family of elements of $\mathcal{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$.

It is required to find the solution $(T_i)_{1\leqq i\leqq n}$ of this system by our algebraic method.

In this respect, using the fundamental identity (1.5), we transform the system (3.1) into the equivalent algebraic system of composition equations

$$
[\delta'_t(t-\alpha)-a_{il}(t)\delta(t-\alpha)]\circ T_i - \sum_{k=2}^n a_{ik}(t)\delta(t-\alpha)\circ T_k = S_i, \qquad 1\leqq i\leqq n. \tag{3.2}
$$

On the other hand, (3.2) is equivalent to the equation

$$
[I'_t(t-\alpha) - M(t)I(t-\alpha)]\circ T = S, \tag{3.3}
$$

in which we have set

$$
M(t) = \begin{pmatrix} a_{11}(t) & \cdots & a_{1n}(t) \\ \cdots & \cdots & \\ a_{n1}(t) & \cdots & a_{nn}(t) \end{pmatrix},
$$

$$
T = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{pmatrix}, \qquad = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{pmatrix}. \tag{3.4}
$$

Further, by composition to the left of both sides of (3.2) with $\mathcal{Y}(t-\alpha)$ we obtain

$$[\mathcal{Y}(t-\alpha)\circ I_t'(t-\alpha)-\mathcal{Y}(t-\alpha)\circ(M(t)I(t-\alpha))]\circ T=\mathcal{Y}(t-\alpha)\circ S. \tag{3.5}$$

But we have $\mathcal{Y}(t-\alpha)\circ I_t'(t-\alpha)=I(t-\alpha)$, and (cf. § 2, Prop. 2.1)

$$\mathcal{Y}(t-\alpha)\circ(Mt)I(t-\alpha))=\{M(\alpha)\}=\left\{\begin{pmatrix} a_{11}(\alpha) & \cdots & a_{1n}(\alpha)\\ \cdots & \cdots & \cdots\\ a_{n1}(\alpha) & \cdots & a_{nn}(\alpha)\end{pmatrix}\right\}. \tag{3.6}$$

Then (3.5) takes the form

$$[I(t-\alpha)-\{M(\alpha)\}]\circ T=\mathcal{Y}(t-\alpha)\circ S. \tag{3.7}$$

For solving (3.7), we will first determine the fundamental matrix, $\{E(t,\alpha)\}$, the solution of the equation

$$\frac{d\{E\}}{dt}-\{M(\Gamma)E\}=I(t-\alpha), \tag{3.8}$$

which is equivalent to the equation

$$[I(t-\alpha)-\{M(\alpha)\}]\circ\{E\}=\mathcal{Y}(t-\alpha). \tag{3.9}$$

For determining $\{E\}$, we will proceed as follows.

Let $\{M(\alpha)\}^\nu$ be the νth composition power of $\{M(\alpha)\}$, given by

$$\{M(\alpha)\}^\nu=\left\{\begin{pmatrix} m_{\nu 11}(t,\alpha) & \cdots & m_{\nu 1n}(t,\alpha)\\ \cdots & \cdots & \cdots\\ m_{\nu n1}(t,\alpha) & \cdots & m_{\nu nn}(t,\alpha)\end{pmatrix}\right\}. \tag{3.10}$$

Then, by definition, for $\nu=0$, $\nu=1$, $\nu=2$ we obtain

$$m_{0ij}(t,\alpha)=\begin{cases}\delta(t-\alpha) & \text{for } i=j,\\ 0 & \text{for } i\neq j,\end{cases}$$

$$m_{1ij}(t,\alpha)=a_{ij}(\alpha)\quad \text{for } 1\leqq i\leqq n,\quad 1\leqq j\leqq n,$$

whence

$$\{M(\alpha)\}^0=I(t-\alpha),\qquad \{M(\alpha)\}^1=\{M(\alpha)\}, \tag{3.11}$$

and

$$\{M(\alpha)\}^2=\left\{\int_\alpha^t M(\xi)M(\alpha)\,d\xi\right\}. \tag{3.12}$$

On the other hand, one has

$$\begin{aligned}\sum_{k=1}^n m_{1ik}\circ m_{2kj}&=\sum_{k=1}^n a_{ik}(\alpha)\circ m_{2kj}(t,\alpha)\\ &=\sum_{k=1}^n\int_\alpha^t a_{ik}(\xi)\Big(\sum_{k_1=1}^n\int_\alpha^\xi a_{kk_1}(\xi_1)\,d\xi_1\Big)a_{k_1j}(\alpha)\,d\xi\\ &=\sum_{k=1}^n\sum_{k_1=1}^n\int_\alpha^t a_{ik}(\xi)\,d\xi\int_\alpha^\xi a_{kk_1}(\xi_1)a_{k_1j}(\alpha)\,d\xi_1.\end{aligned}$$

Likewise, we have

$$\sum_{k=1}^{n} m_{2ik}(t,\alpha)\circ a_{kj}(\alpha) = \sum_{k=1}^{n}\int_{\alpha}^{t} m_{2ik}(t,\xi)a_{kj}(\alpha)\,d\xi$$
$$= \sum_{k=1}^{n}\sum_{k_1=1}^{n}\int_{\alpha}^{t} a_{k_1k}(\xi)\,d\xi\int_{\xi}^{t} a_{ik_1}(\xi_1)\,d\xi_1 a_{kj}(\alpha)$$
$$= a_{3ij}(t,\alpha).$$

It is easy to show that we have the following formula,

$$\int_{\alpha}^{t} a_{kk_1}(\xi)\,d\xi\int_{\xi}^{t} a_{ik_1}(\xi_1)\,d\xi_1 = \int_{\alpha}^{t} a_{1k_1}(\xi)\,d\xi\int_{\alpha}^{\xi} a_{kk_1}(\xi_1)\,d\xi_1,$$

which leads to the relations

$$m_{3ij} = \sum_{k=1}^{n} m_{2ik}(t,\alpha)\circ a_{kj}(\alpha)$$
$$= \sum_{k=1}^{n} a_{ik}(\alpha)\circ m_{2kj}(t,\alpha),$$

and

$$\{M\}\circ\{M\}^2 = \{M\}^2\circ\{M\} = \{M\}^3$$
$$= \left\{\int_{\alpha}^{t} M(\xi)\,d\xi\int_{\alpha}^{\xi} M(\xi_1)\,d\xi_1 M(\alpha)\right\}.$$

whence, by recurrence,

$$\{M\}^{\nu} = \left\{\int_{\alpha}^{t} M(\xi)\,d\xi\int_{\alpha}^{\xi} M(\xi_1)\,d\xi_1 \cdots\int_{\alpha}^{\xi_{\nu-3}} M(\xi_{\nu-2})\,d\xi_{\nu-2}M(\alpha)\right\}.$$

and

$$\{M\}\circ\{M\}^{\nu} = \{M\}^{\nu}\circ\{M\} = \{M\}^{\nu+1}. \tag{3.13}$$

Then for $\nu\in\mathbb{N}-\{0,1,2,3\}$ we obtain

$$m_{\nu_{ij}}(t,\alpha) = \sum_{k_1=1}^{n}\sum_{k_{\nu-1}=1}^{n}\int_{\alpha}^{t} a_{ik_1}(\xi)\,d\xi\int_{\alpha}^{\xi} a_{k_1k_2}(\xi_1)\,d\xi_1$$
$$\cdots\int_{\alpha}^{\xi\nu-3} a_{k_{\nu-2}k_{\nu-1}}(\xi_{\nu-2})a_{k_{\nu-1}j}(\alpha)\,d\xi_{\nu-2} \tag{3.14}$$
$$= \sum_{k_1=1}^{n}\cdots\sum_{k_{\nu-1}=1}^{n} a_{ik_1}\circ a_{k_1k_2}\circ\cdots\circ a_{k_{\nu-1}j},$$

which gives us

$$
\begin{aligned}
\sum_{\nu=0}^{\infty}\{M\}^{\nu} &= I(t-\alpha)+\{M\}+\{M\}^2+\cdots\{M\}^{\nu}+\cdots \\
&= I(t-\alpha)+\left\{\begin{pmatrix} a_{11}(\alpha) & \cdots & a_{1n}(\alpha) \\ \cdots & \cdots & \cdots \\ a_{n1}(\alpha) & \cdots & a_{nn}(\alpha)\end{pmatrix}\right\} \\
&\quad+\left\{\begin{pmatrix} m_{211}(t,\alpha) & \cdots & m_{21n}(t,\alpha) \\ \cdots & \cdots & \cdots \\ m_{2n1}(t,\alpha) & \cdots & m_{2nn}(t,\alpha)\end{pmatrix}\right\}+\cdots \\
&= I(t-\alpha)+\left\{\begin{pmatrix} G_{11}(t,\alpha) & \cdots & G_{1n}(t,\alpha) \\ \cdots & \cdots & \cdots \\ G_{n1}(t,\alpha) & \cdots & G_{nn}(t,\alpha)\end{pmatrix}\right\},
\end{aligned}
\tag{3.15}
$$

where we have set

$$
G_{ij}(t,\alpha)=\sum_{\nu=1}^{\infty} m_{\nu_{ij}}(t,\alpha). \tag{3.16}
$$

From (3.16) we obtain

$$
|G_{ij}(t,\alpha)|\leqq\sum_{\nu=1}^{\infty}|m_{\nu_{ij}}(t,\alpha)|<\infty,
$$

and, if we put

$$
A_p=\operatorname*{Sup}_{p\in N}|a_{ij}(t)| \quad \text{for } 1\leqq i\leqq n,\ 1\leqq j\leqq n,\ a_p\leqq\alpha\leqq t\leqq b_p,
$$

we find that

$$
|m_{\nu_{ij}}|\leqq A_p^{\nu}n^{\nu-1}\frac{(b_p-a_p)^{\nu-1}}{(\nu-1)!}.
$$

Therefore, $(G_{ij}(t,\alpha))_{1\leqq i\leqq n,1\leqq j\leqq n}$ are absolute and uniformly convergent series on each compact subset of $R^2_{t\alpha}$, $\alpha\leqq t$.

This assertion holds also for the derivatives

$$
(\partial G_{ij}/\partial t)_{1\leqq i\leqq n,1\leqq j\leqq n}.
$$

If the functions $(a_{ij}(t))_{1\leqq i\leqq n,1\leqq j\leqq n}$, (resp. $a_{ij}(\alpha))_{1\leqq i\leqq n,1\leqq j\leqq n}$, are of class $\mathscr{C}_t^{(m)}$ (resp. $\mathscr{C}_\alpha^{(m)}$) then $(G_{ij}(t,\alpha))_{1\leqq i\leqq n,1\leqq j\leqq n}$ are of class $\mathscr{C}_{t\alpha}^{(m,m)}$.

Then, by setting

$$
G(t,\alpha)=\begin{pmatrix} G_{11}(t,\alpha) & \cdots & G_{1n}(t,\alpha) \\ \cdots & \cdots & \cdots \\ G_{n1}(t,\alpha) & \cdots & G_{nn}(t,\alpha)\end{pmatrix} \tag{3.17}
$$

we see that $G(t,\alpha)$ belongs to $\mathrm{Mat}_{n\times n}(\mathscr{C}_{t\alpha}^{(m,m)})$ and $\{G(t,\alpha)\}$ to $\mathrm{Mat}_{n\times n}(\{\mathscr{C}_{t\alpha}^{(m,m)}\})$.

On the other hand,

$$\sum_{\nu=0}^{\infty} \{M\}^{\nu} \circ [I(t-\alpha)-\{M\}]$$

$$= [I(t-\alpha)-\{M\}] \circ \sum_{\nu=0}^{\infty} \{M\}^{\nu} = I(t-\alpha), \tag{3.18}$$

shows that $\Sigma_{\nu=0}^{\infty} \{M\}^{\nu}$ is the inverse of $[I(t-\alpha)-\{M\}]$ for the composition operation.

Then, from (3.9), we obtain for the fundamental matrix $\{E\}$ the expression

$$\{E(t, \alpha)\} = \sum_{\nu=0}^{\infty} \{M\}^{\nu} \circ \mathcal{Y}(t-\alpha). \tag{3.19}$$

But, by virtue of (3.15), (3.16) and (3.17), we obtain

$$\{E\} = \mathcal{Y}(t-\alpha) + \left\{ \begin{pmatrix} G_{11}(t, \alpha) & \cdots & G_{1n}(t, \alpha) \\ \cdots & \cdots & \cdots \\ G_{n1}(t, \alpha) & \cdots & G_{nn}(t, \alpha) \end{pmatrix} \right\} \circ \mathcal{Y}(t-\alpha), \tag{3.20}$$

that is,

$$\{E\} = \mathcal{Y}(t-\alpha) + \left\{ \begin{pmatrix} E_{11}(t, \alpha) & \cdots & E_{1n}(t, \alpha) \\ \cdots & \cdots & \cdots \\ E_{n1}(t, \alpha) & \cdots & E_{nn}(t, \alpha) \end{pmatrix} \right\}, \tag{3.21}$$

in which we have set

$$E_{ij}(t, \alpha) = \begin{cases} Y(t-\alpha) + G_{ij}(t-\alpha) \circ Y(t-\alpha) & \text{for } i = j, \\ G_{ij}(t, \alpha) \circ Y(t-\alpha) & \text{for } i \neq j. \end{cases} \tag{3.22}$$

Then the solution T of (3.3) is given by

$$T = \begin{pmatrix} T_1 \\ \vdots \\ T_n \end{pmatrix} = \left\{ \begin{pmatrix} E_{11} & \cdots & E_{1n} \\ \cdots & \cdots & \cdots \\ E_{n1} & \cdots & E_{nn} \end{pmatrix} \right\} \circ \begin{pmatrix} S_1 \\ \vdots \\ S_n \end{pmatrix} = \{E\} \circ S, \tag{3.23}$$

equivalent to

$$T_i = \sum_{k=1}^{n} \{E_{ik}\} \circ S_k, \qquad 1 \leqq i \leqq n. \tag{3.24}$$

Of course, $(E_{ij}(t, \alpha))_{1 \leqq i \leqq n,\, 1 \leqq j \leqq n}$ belong to $\mathscr{C}_{t\alpha}^{(m,m)}$. Consequently, $\{E(t, \alpha)\}$ is a kernel-matrix which belong to $\mathrm{Mat}_{n \times n}(\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)})$, and the solution (3.23) has meaning, if $(S_k)_{1 \leqq k \leqq n}$ are elements of the composition algebra $\mathscr{D}_{(+\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(-\Gamma_t)}$.

4. Boundary-value problem for a system of linear differential equations with variable coefficients.

4.1. Statement of the problem and its solution. In this section we will determine, by means of our algebraic method, the solution of a boundary-value problem for a canonical system of linear differential equations with variable coefficients. Let

$$\frac{dx_i}{dt} = b_i(t) + \sum_{j=1}^{n} a_{ij}(t) x_j(t), \qquad 1 \leqq i \leqq n, \tag{4.1.}$$

be a system of n linear equations of order 1, in which $(a_{ij}(t))_{1 \leqq i \leqq n, 1 \leqq j \leqq n}$ and $(b_i(t))_{1 \leqq i \leqq n}$ are given functions satisfying certain conditions of differentiability in a closed interval $[a, b]$ of the real line R_t.

It is required to find the solution $x_i(t)$, $1 \leqq i \leqq n$, of (4.1) which satisfies the following conditions:

$$x_i(t_0) = c_i, \qquad 1 \leqq i \leqq n, \tag{4.2}$$

at a fixed point $t_0 \in [a, b]$, where $(c_i)_{1 \leqq i \leqq n}$ are real or complex constants. To do this, we will proceed as follows. Let

$$\begin{aligned} \left\{\frac{dx_i}{dt}\right\} &= Y(t-\alpha)\frac{dx_i}{dt}, \\ \{x_i(t)\} &= Y(t-\alpha)x_i(t), \\ \{b_i(t)\} &= Y(t-\alpha)b_i(t), \end{aligned} \tag{4.3}$$

be the kernel-functions of the bimodule $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$, which correspond to the functions (dx_i/dt), $x_i(t)$ and $b_i(t)$, $i \leqq i \leqq n$.

Further, let us transform (4.1) in $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ by multiplying both sides of (4.1) by $Y(t-\alpha)$. Then, according to (4.3), the system (4.1) takes the form

$$\left\{\frac{dx_i}{dt}\right\} = \{b_i(t)\} + \sum_{j=1}^{n} \{a_{ij}(t)x_j(t)\}, \qquad 1 \leqq i \leqq n. \tag{4.4}$$

On the other hand, from the fundamental formula (cf. [1, § 3.4, formula (3.2)]),

$$\left\{\frac{dx_i}{dt}\right\} = \frac{d\{x_i\}}{dt} - \delta(t-\alpha)x_i(\alpha), \qquad 1 \leqq i \leqq n, \tag{4.5}$$

and replacing in (4.4), $\{dx_i/dt\}$ by the right-hand side of (4.5), we find that the kernel-functions $(\{x_i(t)\})_{1 \leqq i \leqq n}$ are solutions in $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ of the system

$$\frac{d\{x_i\}}{dt} - \sum_{j=1}^{n} \{a_{ij}(t)x_j(t)\} = \{b_i(t)\} + \delta(t-\alpha)x_i(\alpha), \qquad 1 \leqq i \leqq n, \tag{4.6}$$

which is a particular case of the system (3.1), where one has

$$T_i = \{x_i\} \quad \text{and} \quad S_i = \{b_i\} + \delta(t-\alpha)x_i(\alpha), \qquad 1 \leqq i \leqq n.$$

Then, if we set

$$\{x\} = \left\{\begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}\right\},$$

and

$$\{B\} = \begin{pmatrix} \{b_1(t)\} + \delta(t-\alpha)x_1(\alpha) \\ \cdots \\ \{b_n(t)\} + \delta(t-\alpha)x_n(\alpha) \end{pmatrix},$$

we obtain

$$\{x\} = \{E\} \circ \{B\}, \tag{4.7}$$

equivalent to (cf. formula (3.24))

$$\{x_i\} = \sum_{j=1}^{n} \{E_{ij}\} \circ \{b_j\} + \sum_{j=1}^{n} \{E_{ij}\}x_j(\alpha), \tag{4.8}$$

which for $t \geqq \alpha$ takes the form

$$x_i(t) = \sum_{j=1}^{n} \int_{\alpha}^{t} E_{ij}(t, \xi) b_j(\xi)\, d\xi + \sum_{j=1}^{n} E_{ij}(t, \alpha) x_j(\alpha), \qquad 1 \leqq i \leqq n, \tag{4.9}$$

where, according to the formula (3.22), we have

$$E_{ij}(t, \alpha) = \begin{cases} 1 + \int_{\alpha}^{t} G_{ij}(t, \xi)\, d\xi & \text{for } i = j, \\ \int_{\alpha}^{t} G_{ij}(t, \xi)\, d\xi & \text{for } i \neq j. \end{cases} \tag{4.10}$$

Hence

$$\begin{aligned} &E_{ii}(\alpha, \alpha) = 1 \quad \text{for } 1 \leqq i \leqq n, \\ &E_{ij}(\alpha, \alpha) = 0 \quad \text{for } i \neq j, \quad 1 \leqq i \leqq n, \quad 1 \leqq j \leqq n, \end{aligned} \tag{4.11}$$

and

$$E(\alpha, \alpha) = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{4.12}$$

Remark 4.1. The fundamental formulas (4.9)–(4.12) hold for each finite interval $[a, b] \subset R$ such that $a \leqq \alpha \leqq t \leqq b$.

Now, for determining the solution of (4.1) which satisfies the conditions (4.2), we are led to assume that one has $a \leqq \alpha \leqq t \leqq t \leqq b$, i.e., that our method gives us the required solution $(x_i^0(t))_{1 \leqq i \leqq n}$ for $t \geqq t_0$. More precisely, α being arbitrary, this solution is obtained by substituting t_0 for α everywhere into the formulas (4.8)–(4.10). Then we obtain for the solution of the problem (4.1)–(4.2)

$$x_i(t) = \sum_{j=1}^{n} \int_{t_0}^{t} E_{ij}(t, \xi) b_j(\xi)\, d\xi + \sum_{j=1}^{n} E_{ij}(t, t_0) c_j, \qquad 1 \leqq i \leqq n, \quad \text{for } t \geqq t_0 \geqq \alpha. \tag{4.13}$$

The problem which may be stated now is the following: is it possible also to determine, by our algebraic method, the solution for $t \leqq t_0 \leqq \alpha$ with the same initial conditions at the point t_0? More precisely, is it possible to obtain the solution of (4.1) in a neighborhood of t_0?

The answer to this question is affirmative. Indeed, in [1, § 5], we have determined the solution of a differential equation

$$\frac{d^n y}{dt^n} + \sum_{j=1}^{n} a_j(t) \frac{d^{n-j} y}{dt^{n-j}} = f(t), \tag{4.14}$$

which satisfies the initial conditions

$$y^{(k)}(t_0) = c_k, \qquad 0 \leqq k \leqq n-1, \tag{4.15}$$

by means of the composition product in $\mathscr{D}'_{(+\Gamma_t)(-\Gamma_\alpha)}$ defined by

$$S(t, \alpha) \circ T(t, \alpha) = \int_{\alpha}^{t} S(t, \xi) T(\xi, \alpha)\, d\xi, \tag{4.16}$$

and of kernel-functions defined by

$$\{f\}_+ = Y(t-\alpha)f(t,\alpha) = \begin{cases} f(t,\alpha) & \text{for } t \geqq \alpha, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.17}$$

Thus, we obtain the solution for $t \geqq t_0 \geqq \alpha$.

Now let $\mathscr{D}'_{(-\Gamma_t)(+\Gamma_\alpha)}$ be the locally convex space of distributions with support limited to the right with respect to the variable t and to the left with respect to the variable α (for the definition and the properties of this space cf. [4, §2, No. 5]).

Let $\mathscr{D}_{(-\Gamma_t)}$ (resp. $\mathscr{D}_{(+\Gamma_\alpha)}$) be the locally convex space of indefinitely differentiable functions with support limited to the right (resp. to the left) with respect to the variable t (resp. α).

We suppose $\mathscr{D}_{(-\Gamma_t)}$ (resp. $\mathscr{D}_{(+\Gamma_\alpha)}$) equipped with the topology induced by the strong dual $\mathscr{D}'_{(-\Gamma_t)}$ of $\mathscr{D}_{(+\Gamma_t)}$ (resp. by the strong dual $\mathscr{D}'_{(+\Gamma_\alpha)}$ of $\mathscr{D}_{(-\Gamma_\alpha)}$).

Let us consider, on the other hand, the locally convex space $\mathscr{D}_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(+\Gamma_\alpha)}$ (resp. $\mathscr{D}'_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}_{(+\Gamma_\alpha)}$).

Then, for S, T elements of $\mathscr{D}_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(+\Gamma_\alpha)}$ (resp. of $\mathscr{D}'_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}_{(+\Gamma_\alpha)}$), the composition product

$$S \circ T = \int_t^\alpha S(t,\xi)T(\xi,\alpha)\,d\xi \tag{4.18}$$

also belongs to $\mathscr{D}_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(+\Gamma_\alpha)}$ (resp. to $\mathscr{D}'_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}_{(+\Gamma_\alpha)}$) which, therefore, is a composition algebra.

It follows that $\mathscr{D}'_{(-\Gamma_t)(+\Gamma_\alpha)}$ is a composition module over $\mathscr{D}_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}'_{(+\Gamma_\alpha)}$ (resp. $\mathscr{D}'_{(-\Gamma_t)} \hat{\otimes} \mathscr{D}_{(+\Gamma_\alpha)}$).

In order to obtain the solution of (4.14) satisfying the conditions (4.15) for $t \leqq t_0 \leqq \alpha$, we transform the equation (4.14) into the composition bimodule $\mathscr{D}'_{(-\Gamma_t)(+\Gamma_\alpha)}$.

For doing this, we multiply both sides of (4.14) by $Y(\alpha - t)$, and apply after this operation, the same algebraic method as for $\alpha \leqq t_0 \leqq t$ but using the composition product (4.18).

The important result of this method consists in the fact that the solution of (4.14)–(4.15) is given, for $t \leqq t_0 \leqq \alpha$ by the same formula as for $\alpha \leqq t_0 \leqq t$.

In brief, by means of this method we obtain the solution in a neighborhood of the point t_0.

In our next papers we will show that this result is also for linear systems of differential equations and for partial differential equations with variable coefficients.

4.2. Fundamental system of integrals of a system of linear differential equations. The formula (4.13), which gives us the solution of the problem (4.1)–(4.2), holds for any arbitrary functions $(b_i)_{1 \leqq i \leqq n}$ satisfying certain conditions of differentiability. In particular, this is true for $b_i(t) = 0$, $1 \leqq i \leqq n$. In this case, (4.13) is the solution of the *associated* homogeneous system of linear differential equations

$$\frac{dy_i}{dt} = \sum_{k=1}^{n} a_{ij}(t)y_i, \qquad 1 \leqq i \leqq n. \tag{4.20}$$

From (4.13) and (4.19) we obtain

$$\frac{dy_i}{dt} = \sum_{j=1}^{n} c_{ii}\frac{dE_{ij}}{dt} = \sum_{j=1}^{n} a_{ij}(t)\Big(\sum_{k=1}^{n} c_k E_{jk}(t, t_0)\Big). \tag{4.21}$$

Then, if we put

$$Y_j = \begin{pmatrix} E_{1j}(t, t_0) \\ E_{2j}(t, t_0) \\ \vdots \\ E_{nj}(t, t_0) \end{pmatrix}, \qquad 1 \leqq j \leqq n, \tag{4.22}$$

and if we identify the coefficients of c_i, $1 \leqq i \leqq n$, in both sides of the second equality in (4.21), we find that Y_j is a solution of

$$\frac{dY_j}{dt} = MY_i, \qquad 1 \leqq j \leqq n. \tag{4.23}$$

On the other hand, from (3.19) and (4.21) we obtain

$$E(t, t_0) = (Y_1(t, t_0), \cdots, Y_n(t, t_0)), \tag{4.24}$$

Hence, by virtue of (4.12),

$$E(t_0, t_0) = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & & \cdots & & \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{4.25}$$

and

$$\text{Det } E(t_0, t_0) = 1. \tag{4.26}$$

Therefore, (cf. Bourbaki, [5, Chapt. IV, § 2, No. 4, Props. 4, 5]), $(Y_j)_{1 \leqq j \leqq m}$ is a fundamental system of integrals of (4.1).

4.3. Example. Assume $n = 2$, and consider the boundary value problem $x_i(t_0) = c_1$, $1 \leqq i \leqq 2$, for the system

$$\begin{aligned} \frac{dx_1}{dy} &= b_1(t) + a_{12}(t)x_2, \\ \frac{dx_2}{dt} &= b_2(t) + a_{21}(t)x_1. \end{aligned} \tag{4.27}$$

One has

$$I(t-\alpha) = \begin{pmatrix} \delta(t-\alpha) & 0 \\ 0 & \delta(t-\alpha) \end{pmatrix}, \quad M(\alpha) = \begin{pmatrix} 0 & a_{12}(\alpha) \\ a_{21}(\alpha) & 0 \end{pmatrix},$$

$$G_{11}(t, \alpha) = \sum_{p=1}^{\infty} (a_{12}(\alpha) \circ a_{21}(\alpha))^{(p)},$$

$$G_{12}(t, \alpha) = \sum_{p=0}^{\infty} (a_{12}(\alpha) \circ a_{21}(\alpha))^{(p)} \circ a_{12}(\alpha),$$

$$G_{21}(t, \alpha) = \sum_{p=0}^{\infty} (a_{21}(\alpha) \circ a_{12}(\alpha))^{(p)} \circ a_{21}(\alpha),$$

$$G_{22}(t, \alpha) = \sum_{p=1}^{\infty} (a_{21}(\alpha) \circ a_{12}(\alpha))^{(p)},$$

where

$$(a_{12}(\alpha) \circ a_{21}(\alpha))^{(p)} = \int_\alpha^t a_{12}(\xi)\, d\xi \int_\alpha^\xi a_{21}(\xi_1)\, d\xi_1 \cdots \int_\alpha^{\xi_{2p-3}} a_{12}(\xi_{2p-2}) a_{21}(\alpha)\, d\xi_{2p-2}$$

and

$$(a_{12}(\alpha) \circ a_{21}(\alpha))^{(p)} \circ a_{12}(\alpha) = \int_\alpha^t a_{12}(\xi)\, d\xi \int_\alpha^\xi a_{21}(\xi_1)\, d\xi_1 \cdots \int_\alpha^{\xi_{2p-2}} a_{21}(\xi_{2p-1}) a_{12}(\alpha)\, d\xi_{2p-1}.$$

Similar expressions are obtained for the terms G_{21} and G_{22}.

Finally, the required solution is given by

$$\begin{aligned} x_1(t) = c_1\Big(1 + \int_{t_0}^t G_{11}(t, \xi)\, d\xi\Big) + c_2 \int_{t_0}^t G_{12}(t, \xi)\, d\xi \\ + \int_{t_0}^t b_1(\xi)\, d\xi \Big(1 + \int_\xi^t G_{11}(t, \xi_1)\, d\xi_1\Big) \\ + \int_{t_0}^t b_2(\xi)\, d\xi \int_\xi^t G_{12}(t, \xi_1)\, d\xi_1, \end{aligned}$$

and

$$\begin{aligned} x_2(t) = c_1 \int_{t_0}^t G_{21}(t, \xi)\, d\xi + c_2\Big(1 + \int_{t_0}^t G_{22}(t, \xi)\, d\xi\Big) \\ + \int_{t_0}^t b_1(\xi)\, d\xi \int_\xi^t G_{21}(t, \xi_1)\, d\xi_1 \\ + \int_{t_0}^t b_2(\xi)\, d\xi \Big(1 + \int_\xi^t G_{22}(t, \xi_1)\, d\xi_1\Big), \end{aligned}$$

in which

$$c_1 = x_1(t_0) \quad \text{and} \quad c_2 = x_2(t_0).$$

Acknowledgments. The author wishes to thank the referees for numerous helpful suggestions, and Professor Gregers Krabbe, Purdue University, Lafayette, Indiana, who read the final version of the manuscript and gave me good advice that led to simpler presentations of several results.

[1] S. VASILACH, *Algebraic method for solving linear differential equations whose coefficients are functions of one variable*, this Journal, 6 (1975), pp. 295–311.

[2] ———, *Algebraic method for solving linear partial differential equations with variable coefficients, Part I. Basic theory*, this Journal, 10 (1979), pp. 586–602.

[3] ———, *Algebraic method for solving linear partial differential equations with variable coefficients*, this Journal, 10 (1979), pp. 1077–1088.

[4] ———, *Sur une classe d'espaces de distributions*, Boll. Un. Mat. Ital., 5 (1970), pp. 745–760.

[5] ———, *Calcul opérationnel Algébrique des distributions à support dans* R_+^n, $n \geqq 1$, Rev. Math. Pures et Appl. I, 4 (1959), pp. 185–219; II, 5 (1960), pp. 495–531; III, 6 (1961), pp. 69–100; IV, V, 8 (1963), pp. 19–66.

[6] N. BOURBAKI, *Fonctions d'une variable réelle*, Hermann, Paris, 1951, Chap. IV–VII.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0004 $01.00/0

FEEDBACK REDUCING SENSITIVITY BY A FACTOR $\lambda<1$*

VACLAV DOLEZAL†

Abstract. Necessary and sufficient conditions for a plant and controller are established under which there exists a feedback reducing the incremental sensitivity by a given factor $\lambda < 1$. It is understood that the reduction occurs for all prescribed inputs and plant perturbations.

Introduction. The problem of reducing sensitivity by feedback is old and well researched. Modern approaches to this area use operator theory for the description of a system, and the sensitivity is defined in terms of norms. In this context, let us mention works [1], [5], [6], and more recent papers [7], [8]. Basically, the known results concern the case in which the error of an equivalent closed-loop system does not exceed in norm the error of the open-loop system.

In this paper we discuss a stronger requirement—reducing the incremental sensitivity by a factor $\lambda < 1$. To explain the problem under consideration, assume that an open-loop system $\{G_0, P_0\}$ is given (see Fig. 1), where G_0, P_0 are bounded, linear (possibly causal) operators on a Banach space. Moreover, suppose that a number

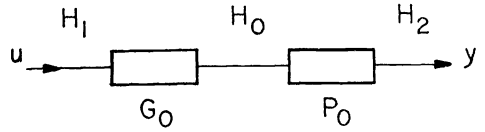


FIG. 1.

$0<\lambda<1$, a set U of inputs and a set $\mathcal{W}$ of plant perturbations of interest is prescribed. The question is whether there exists a feedback and controller described by a bounded, linear (possibly causal) operator K and G, respectively, such that:

(a) The closed-loop system $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$; i.e. both systems have the same input-output operator, (see Fig. 2).

(b) For every $u \in U$ and $W \in \mathcal{W}$ we have $\|\delta_c\| \leqq \lambda \|\delta_0\|$, where δ_c and δ_0 are the errors in the output of $\{G, P_0, K\}$ and $\{G_0, P_0\}$, respectively, determined by the input u and the perturbation W of the nominal plant P_0. It is understood that the errors δ_c, δ_0 are given in terms of the Fréchet derivatives of the corresponding input–output operators [2].

If the sets U and $\mathcal{W}$ are not too slim, i.e., if a certain set Ω depending on G_0, U and $\mathcal{W}$ is dense, we can give a necessary and sufficient condition for the existence of such a feedback K and controller G. It turns out that this condition is the existence of a bounded, linear (possibly causal) right-inverse of the operator P_0G_0.

Undoubtedly, this is quite a strong requirement. However, since this condition is both necessary and sufficient, it clearly gives the limits for sensitivity reduction.

On the other hand, in the second part of the paper, we examine assumptions on U, $\mathcal{W}$ and G_0 which guarantee the density of the set Ω. In particular, we study the case where $L_2[0, \infty)$ is the underlying space. Surprisingly enough, the sets U, $\mathcal{W}$ need not be too large for Ω to be dense. For example, Theorem 4 shows that if U consists of a single exponential $ae^{-\alpha t}$, $\alpha > 0$, $a \neq 0$, and $\mathcal{W}$ contains input–output operators of all stable systems with constant, lumped elements, then Ω is dense.

* Received by the editors May 15, 1979, and in revised form June 10, 1980. This research was supported by the National Science Foundation under grant MPS 7505268.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, New York 11794.

In our setting, the causality does not play any essential role, but can be easily incorporated into the considerations. As a result, our main Theorem 2 consists in fact of two assertions: one ignoring causality, and the other taking it into account.

1. Results. We begin with some notations and definitions.

If H_1, H_2 are Banach spaces over the same field of scalars, we let $[H_1, H_2]$ be the Banach space of all linear bounded operators $A: H_1 \to H_2$ equipped with the customary norm $\|A\| = \sup\{\|Ax\|: x \in H_1, \|x\| = 1\}$.

If, in particular, H_i is a Hilbert space endowed with a resolution of identity $\{P_T^{(i)}: T \in R^1\}$, $i = 1, 2$, and $A \in [H_1, H_2]$, then A will be called causal [1], if

$$P_T^{(2)} A = P_T^{(2)} A P_T^{(1)}$$

for every $T \in R^1$.

Let L be a Banach space and let $X^0 \in L$, $r > 0$; then we put $B_r(X^0) = \{X: X \in L, \|X - X^0\| < r\}$.

We introduce the (incremental) sensitivity as follows, [2]:

DEFINITION 1. Let H_0, H_1, H_2 be Banach spaces having the same set of scalars, and let $X^0 \in [H_0, H_2]$. For every fixed $X \in B_r(X^0)$, $r > 0$, let S_X be a system whose input–output operator $F(X)$ is in $[H_1, H_2]$; i.e., S_X carries each input $u \in H_1$ into an output $y \in H_2$ and $y = F(X)u$. If the Fréchet derivative $\partial F(\cdot)$ of F exists at X^0, then it will be called the (incremental) sensitivity of the (nominal) system S_{X^0} at X^0.

As it was indicated in [2], for $W \in [H_0, H_2]$ with $\|W\|$ small, the operator $\partial F(W)$ approximates $F(X^0 + W) - F(X^0)$. Consequently, for any fixed input $u_0 \in H_1$, the element $\partial F(W)u_0 \in H_2$ is an approximation to the actual error $\delta = F(X^0 + W)u_0 - F(X^0)u_0$. This fact justifies our definition.

In order to compare sensitivities of different systems, we introduce the following definition.

DEFINITION 2. Let $\partial F_i(\cdot)$ be the sensitivity of the (nominal) system $S^i_{X^0}$, $i = 1, 2$ at X^0, and let $\lambda > 0$. Moreover, let $U \subset H_1$ and $\mathcal{W} \subset [H_0, H_2]$ be nonempty subsets. We will write $\partial F_1(\cdot) \leqq \lambda \partial F_2(\cdot)$ with respect to $(U, \mathcal{W})$ if

$$\|\partial F_1(W)u\| \leqq \lambda \|\partial F_2(W)u\| \quad \text{for all } u \in U \text{ and } W \in \mathcal{W}. \tag{1}$$

This definition expresses precisely what we have said above about reducing sensitivity by a factor λ. Indeed, assume that in both systems $S^1_{X^0}$ and $S^2_{X^0}$ the nominal operator X^0 is perturbed by an increment W which is in the set $\mathcal{W}$ of all perturbations of interest. If both systems $S^1_{X^0+W}$ and $S^2_{X^0+W}$ have the same input $u \in U$ (set of all inputs of interest), then $\delta_1 = \partial F_1(W)u$ and $\delta_2 = \partial F_2(W)u$ are the respective errors provided second-order quantities are neglected. Thus, (1) means that $\|\delta_1\| \leqq \lambda \|\delta_2\|$.

Turning now to our proper objective, let H_0, H_1, H_2 be Banach spaces having the same system of scalars and let $G \in [H_1, H_0]$, $P_0 \in [H_0, H_2]$, $K \in [H_2, H_1]$. It is well known [3] that any input $u \in H_1$ of the closed-loop system $\{G, P_0, K\}$ (Fig. 2) produces a unique output $y \in H_2 \Leftrightarrow$ the operator $N = I + KP_0G \in [H_1, H_1]$ is invertible. In this case, $y = F_c^0 u$, where

$$F_c^0 = P_0 G (I + KP_0 G)^{-1}. \tag{2}$$

Referring to Figs. 1 and 2, we introduce the following terminology:

DEFINITION 3. Let $G_0, G \in [H_1, H_0]$, $P_0 \in [H_0, H_2]$ and $K \in [H_2, H_1]$.

(i) The closed-loop system $\{G, P_0, K\}$ is called normal, if the operator $N = I + KP_0G$ is invertible. If, in addition, H_0, H_1, H_2 are Hilbert resolution spaces and the operators G, P_0, K, N^{-1} are causal, then $\{G, P_0, K\}$ is called causal.

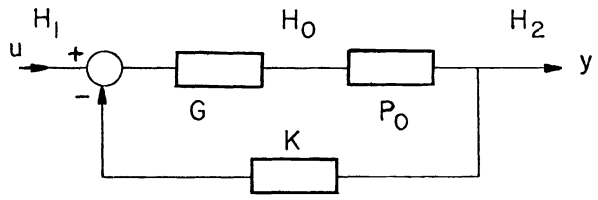


FIG. 2.

(ii) The closed-loop system $\{G, P_0, K\}$ is called equivalent to the open-loop system $\{G_0, P_0\}$, if $\{G, P_0, K\}$ is normal and

$$G_0 = G(I + KP_0G)^{-1}. \tag{3}$$

Several comments are in order. First, observe that in our setting N^{-1} is bounded as a consequence of the open mapping theorem. Hence, the operator F_c^0 is bounded; i.e., a normal system $\{G, P_0, K\}$ is stable. Second, if $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$, then (2) and (3) show that $F_c^0 = F_0^0$, where $F_0^0 = P_0G_0$ is the input–output operator of $\{G_0, P_0\}$. Hence, these systems are indistinguishable if considered as black-boxes.

If an open-loop system $\{G_0, P_0\}$ is given, we can readily single out all closed-loop systems $\{G, P_0, K\}$ that are equivalent to $\{G_0, P_0\}$. To do this, we will need the following well-known result (see [1], p. 64):

LEMMA 1. *Let h, h' be linear spaces having the same system of scalars, and let $A : h' \to h$, $B : h \to h'$ be linear operators. If the operator $N = I + AB : h \to h$ is invertible, then $M = I + BA : h' \to h'$ is also invertible, and we have*

$$M^{-1} = I - BN^{-1}A, \tag{4}$$

$$BN^{-1} = M^{-1}B, \qquad AM^{-1} = N^{-1}A. \tag{5}$$

If, in addition, h, h' are Hilbert resolution spaces and A, B, N^{-1} are causal, then M^{-1} is also causal.

LEMMA 2. *Let $P_0 \in [H_0, H_2]$ and $G_0 \in [H_1, H_0]$. If $G \in [H_1, H_0]$ and $K \in [H_2, H_1]$, then the closed-loop system is equivalent to the open-loop system $\{G_0, P_0\}$, iff*

(i) *the operator $Q = I - P_0G_0K \in [H_2, H_2]$ is invertible, and*

(ii) $G = (I - G_0KP_0)^{-1}G_0$.

Proof. (a) Assume first that conditions (i) and (ii) are satisfied. Note that (ii) is meaningful; i.e., the operator $R = I - G_0KP_0 \in [H_0, H_0]$ is invertible by virtue of (i) and Lemma 1.

Next, by (ii),

$$N = I + KP_0G = I + KP_0(I - G_0KP_0)^{-1}G_0.$$

Since R is invertible, the operator $L = I - KP_0G_0$ is also invertible by Lemma 1, and (4) yields

$$L^{-1} = I + KP_0(I - G_0KP_0)^{-1}G_0 = N.$$

Hence, N is invertible, and

$$N^{-1} = I - KP_0G_0; \tag{6}$$

i.e., the closed-loop system $\{G, P_0, K\}$ is normal.

Moreover, from (ii) we have $G_0(I + KP_0G) = G$. Since $I + KP_0G$ is invertible, (3) follows. Thus, $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$.

(b) Conversely, assume that $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$; i.e., N is invertible and (3) holds. Then, $Q = I - P_0 G_0 K = I - P_0 G(I + KP_0 G)^{-1} K$. Since $I + KP_0 G$ is invertible, $S = I + P_0 GK$ is also invertible, and by (4),

$$S^{-1} = I - P_0 G(I + KP_0 G)^{-1} K = Q.$$

Hence, Q is invertible and

$$Q^{-1} = I + P_0 GK_3; \tag{7}$$

i.e., condition (i) is satisfied.

Moreover, (3) yields $G_0 = (I - G_0 KP_0)G$. However, since Q is invertible, Lemma 1 shows that $I - G_0 KP_0$ is invertible. Thus, condition (ii) follows and the proof is complete.

REMARK 1. If we had defined the equivalence by equation $F_c^0 = F_0^0$, i.e., via equality of the corresponding input–output operators, then a result like Lemma 2 is still true. The only difference is that condition (ii) has to be replaced by

(ii)* $G = (I - G_0' KP_0)^{-1} G_0'$, where $G_0' = G_0 + U_0$ with $U_0 \in [H_1, H_0]$ being an operator satisfying the equation $P_0 U_0 = 0$.

The next assertion is of crucial importance for our considerations.

LEMMA 3. *Let $G_0, G \in [H_1, H_0]$, $P_0 \in [H_0, H_2]$, $K \in [H_2, H_1]$, and assume that the closed-loop system $\{G, P_0, K\}$ is equivalent to the open-loop system $\{G_0, P_0\}$. Then:*

(i) *There exists $r > 0$ such that the operator $I + KXG \in [H_1, H_2]$ is invertible for each $X \in B_r(P_0) = \{X : X \in [H_0, H_2], \|X - P_0\| < r\}$.*

(ii) *The maps $F_0, F_c : B_r(P_0) \to [H_1, H_2]$ defined by*

$$F_0(X) = XG_0, \quad F_c(X) = XG(I + KXG)^{-1} \tag{8}$$

are differentiable at P_0.

(iii) *For the sensitivity $\partial F_0(\cdot)$ and $\partial F_c(\cdot)$ of $\{G_0, P_0\}$ and $\{G, P_0, K\}$ at P_0, respectively, we have*

$$\partial F_c(W) = Q \partial F_0(W) \tag{9}$$

for all $W \in [H_0, H_2]$, where $Q = I - P_0 G_0 K$.

Proof. By Definition 3, the operator $N = I + KP_0 G$ is invertible, and (3) holds. Next, invoking the results on Fréchet derivatives discussed in [2] we obtain that the maps F_0 and F_c are well defined on a ball $B_r(P_0)$ with some $r > 0$, and they are differentiable at P_0. Also, by (8) we have, for each $W \in [H_0, H_2]$,

$$\partial F_0(W) = WG_0, \tag{10}$$

and

$$\begin{aligned} \partial F_c(W) &= \partial[XG](W)(I + KP_0 G)^{-1} + P_0 G \partial[(I + KXG)^{-1}](W) \\ &= WG(I + KP_0 G)^{-1} - P_0 G(I + KP_0 G)^{-1} KWG(I + KP_0 G)^{-1} \\ &= [I - P_0 G(I + KP_0 G)^{-1} K] WG(I + KP_0 G)^{-1}. \end{aligned} \tag{11}$$

However, invoking (3) and (10), we get

$$\partial F_c(W) = (I - P_0 G_0 K) WG_0 = Q \partial F_0(W),$$

which confirms (9).

If $U \subset H_1$ and $\mathcal{W} \subset [H_0, H_2]$ are given nonempty sets of inputs and plant perturbations of interest, respectively, and if $G_0 \in [H_1, H_0]$, we denote

$$\Omega_{G_0}(U, \mathcal{W}) = \{WG_0 u : W \in \mathcal{W}, u \in U\} \subset H_2. \tag{12}$$

Then we have the following result:

THEOREM 1. *Let G_0, $G \in [H_1, H_0]$, $P_0 \in [H_0, H_2]$, $K \in [H_2, H_1]$, and let the closed-loop system $\{G, P_0, K\}$ be equivalent to the open-loop system $\{G_0, P_0\}$. Moreover, let $U \subset H_1$, $\mathcal{W} \subset [H_0, H_2]$ be nonempty, and let $\lambda > 0$. If $\partial F_c(\cdot)$ and $\partial F_0(\cdot)$ are the sensitivities at P_0 of $\{G, P_0, K\}$ and $\{G_0, P_0\}$, respectively, then $\partial F_c(\cdot) \leqq \lambda \partial F_0(\cdot)$ with respect to $(U, \mathcal{W}) \Leftrightarrow$*

$$\sigma = \sup \{\|Qx\| \cdot \|x\|^{-1} : x \in \Omega_{G_0}(U, \mathcal{W}), x \neq 0\} \leqq \lambda, \tag{13}$$

where $Q = I - P_0 G_0 K$.

(The proof follows trivially from (9), (10), (12) and Definition 2.)

Let us consider now the following problem. Suppose that $\{G_0, P_0\}$ and U, $\mathcal{W}$ are given, and that $0 < \lambda < 1$. What conditions must be imposed on P_0, G_0 under which there exist G and K, such that $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$, and for the corresponding sensitivities we have $\partial F_c(\cdot) \leqq \lambda \partial F_0(\cdot)$ with respect to $(U, \mathcal{W})$? It turns out that such a condition can be found easily provided $\Omega_{G_0}(U, \mathcal{W})$ is dense in H_2. Also, in the next paragraph we will show that the density requirement is satisfied for quite meager but reasonably selected sets U and $\mathcal{W}$.

To pave the way towards the main result, we have to prove a few auxiliary claims.

LEMMA 4. *Let $\Omega_{G_0}(U, \mathcal{W})$ be dense in H_2; then $\sigma = \|Q\|$, where σ is defined by* (13).

Proof. The function $\phi(x) = \|Qx\| \cdot \|x\|^{-1}$ is continuous on $H_2 - \{0\}$.

It is convenient to introduce the following two sets $\mathcal{K}_\lambda(T)$ and $\mathcal{K}_\lambda^*(T)$.

DEFINITION 4. If $T \in [H_1, H_2]$ [is causal] and $0 < \lambda < 1$, let $\mathcal{K}_\lambda(T)[\mathcal{K}_\lambda^*(T)]$ be the set of all [causal] operators $K \in [H_2, H_1]$ such that

(i) $Q = I - TK \in [H_2, H_2]$ is invertible [and Q^{-1} is causal],

(ii) $\|Q\| \leqq \lambda$.

We will need the following assertion:

LEMMA 5. *Let h, h' be Banach spaces over the same field of scalars, let $A_0 \in [h, h']$ be invertible, and let $0 < \alpha < 1$. If $A \in [h, h']$ and*

$$\|A - A_0\| \leqq \alpha \|A_0^{-1}\|^{-1},$$

then A is invertible, $A^{-1} \in [h', h]$ and

$$\|A^{-1}\| \leqq (1 - \alpha)^{-1} \|A_0^{-1}\|.$$

If, in addition, h, h' are Hilbert resolution spaces and A_0^{-1}, $A - A_0$ are causal, then A^{-1} is also causal.

This is a standard result except, perhaps, for the second claim which follows easily by realizing the fact that $(A - A_0)A_0^{-1}$ is a contraction.

LEMMA 6. *Let $T \in [H_1, H_2]$ [be causal] and let $0 < \lambda < 1$. Then $\mathcal{K}_\lambda(T) \neq \emptyset$ $[\mathcal{K}_\lambda^*(T) \neq \emptyset] \Leftrightarrow T$ possesses a linear, bounded [causal] right-inverse; i.e., there exists [a causal] $M \in [H_2, H_1]$ such that $TM = I$.*

Proof. (a) Assume that there exists $M \in [H_2, H_1]$ such that $TM = I$. Choose a number μ with $1 - \lambda \leqq \mu < 1$ and put $K = \mu M$. Then $K \in [H_2, H_1]$ and $Q = I - TK = (1 - \mu)I$. Since $0 < 1 - \mu \leqq \lambda$, Q is invertible and $\|Q\| = 1 - \mu \leqq \lambda$, so that $K \in \mathcal{K}_\lambda(T) \neq \emptyset$. If, in addition, T is causal and the existing M is causal, then the above K is causal and so is $Q^{-1} = (1 - \mu)^{-1} I$. Hence, $\mathcal{K}_\lambda^*(T) = \emptyset$.

(b) Conversely, let $\mathcal{K}_\lambda(T) \neq \emptyset$. Choosing some $K \in \mathcal{K}_\lambda(T)$, then $Q = I - TK \in [H_2, H_2]$ satisfies the condition $\|Q\| \leqq \lambda < 1$, and consequently, by Lemma 5, the operator $I - Q = TK$ possesses a bounded inverse $S \in [H_2, H_2]$. Hence, $TKS = I$; i.e. $KS \in [H_2, H_1]$ is a bounded right-inverse of T.

Finally, let T be causal and let $\mathcal{K}_\lambda^*(T) \neq \emptyset$. If we choose again some $K \in \mathcal{K}_\lambda^*(T)$, the operator K will be causal; consequently, $Q = I - TK$ is causal. From Lemma 5 it

follows that $(I-Q)^{-1}$ is causal, i.e., that TK possesses a causal inverse $S \in [H_2, H_2]$. Hence, KS is a causal right-inverse of T and the proof is complete.

Now we can state the main theorem.

THEOREM 2. *Let H_0, H_1, H_2 be Banach [Hilbert resolution] spaces over the same field of scalars, let $P_0 \in [H_0, H_2]$, $G_0 \in [H_1, H_0]$ [be causal], and let $0 < \lambda < 1$. Also, let $U \subset H_1$ and $\mathcal{W} \subset [H_0, H_2]$ be such that the set $\Omega_{G_0}(U, \mathcal{W})$ defined by* (12) *is dense in H_2. Then the following two conditions are equivalent*:

(i) *There exist [causal] operators $G \in [H_1, H_0]$ and $K \in [H_2, H_1]$ such that the closed-loop system $\{G, P_0, K\}$ is [causal and] equivalent to the open-loop system $\{G_0, P_0\}$, and*

$$\partial F_c(\cdot) \leqq \lambda \, \partial F_0(\cdot) \tag{14}$$

with respect to $(U, \mathcal{W})$, where $\partial F_c(\cdot)$ and $\partial F_0(\cdot)$ is the sensitivity of $\{G, P_0, K\}$ and $\{G_0, P_0\}$ at P_0, respectively.

(ii) *$P_0 G_0$ possesses a linear, bounded [and causal] right-inverse; i.e., there exists a [causal] $M \in [H_2, H_1]$ such that $P_0 G_0 M = I$.*

Proof. (a) Assume that condition (ii) is satisfied. Then by Lemma 6, $\mathcal{K}_\lambda(P_0 G_0) \neq \varnothing$, or $\mathcal{K}_\lambda^*(P_0 G_0) \neq \varnothing$ provided P_0, G_0 are causal. Thus, choose a $K \in \mathcal{K}_\lambda(P_0 G_0)$, $[K \in \mathcal{K}_\lambda^*(P_0 G_0)]$, and put

$$Q = I - P_0 G_0 K. \tag{15}$$

Then, Q is invertible, $[Q^{-1}$ is causal] and $\|Q\| \leqq \lambda$ by Definition 4. Since Q is invertible, the operator $I - G_0 K P_0$ is also invertible by Lemma 1. Now, define G by

$$G = (I - G_0 K P_0)^{-1} G_0. \tag{16}$$

Clearly, G is bounded since $(I - G_0 K P_0)^{-1}$ is bounded by the open mapping theorem.

Also, note that G is causal provided $K \in \mathcal{K}_\lambda^*(P_0 G_0)$. Indeed, by (4) in Lemma 1,

$$(I - G_0 K P_0)^{-1} = I + G_0 K (I - P_0 G_0 K)^{-1} P_0,$$

so that

$$G = (I + G_0 K Q^{-1} P_0) G_0. \tag{17}$$

However, since P_0, G_0, K, Q^{-1} are all causal, (17) shows that G is causal. Moreover, by identity (6), $(I + K P_0 G)^{-1} = I - K P_0 G_0$, so that $(I + K P_0 G)^{-1}$ is causal. Hence, by Definition 3, the closed-loop system $\{G, P_0, K\}$ is causal.

On the other hand, if we recall Lemma 2, (15) and (16) show that $\{G, P_0, K\}$ is equivalent to $\{G_0, P_0\}$.

Next, since $\Omega_{G_0}(U, \mathcal{W})$ is dense in H_2, Lemma 4 shows that $\sigma = \|Q\| \leqq \lambda$; consequently, by Theorem 1, $\partial F_c(\cdot) \leqq \lambda \, \partial F_0(\cdot)$ with respect to $(U, \mathcal{W})$. Hence, the closed-loop system $\{G, P_0, K\}$ we constructed satisfies condition (i).

(b) Assume now that condition (i) is satisfied. Then (14) implies by Theorem 1 and Lemma 4 that, with $Q = I - P_0 G_0 K$, we have $\|Q\| \leqq \lambda$. Also, by Lemma 2, Q is invertible. Hence, by Definition 4, $K \in \mathcal{K}_\lambda(P_0 G_0)$. Consequently, by Lemma 6, $P_0 G_0$ possesses a bounded right-inverse.

If, in addition, P_0, G_0, G, K are causal, then identity (7) shows that Q^{-1} is also causal. Thus, $K \in \mathcal{K}_\lambda^*(P_0 G_0)$ so that $P_0 G_0$ has a bounded, causal right-inverse by Lemma 6. Hence, condition (ii) is satisfied and the proof is complete.

2. Discussion. Condition (ii) in Theorem 2 is a rather severe requirement. Indeed, if $\{G_0, P_0\}$ is a "practical" open-loop system, then $P_0 G_0$ is usually not an onto map and

thus (ii) is not met. Consequently, by virtue of Theorem 2, for such a system $\{G_0, P_0\}$ with sets U, $\mathcal{W}$ satisfying the density assumption, it is impossible to find a feedback K and a controller G such that the equivalent closed-loop system $\{G, P_0, K\}$ would reduce the sensitivity of $\{G_0, P_0\}$ with respect to $(U, \mathcal{W})$ by a factor $\lambda < 1$.

On the other hand, it is clear that if the selected sets U and $\mathcal{W}$ are not too large, i.e., the set $\Omega_{G_0}(U, \mathcal{W})$ is not dense in H_2, then such K and G might exist. However, as we shall show below, U and $\mathcal{W}$ can be quite small yet $\Omega_{G_0}(U, \mathcal{W})$ can still be dense.

To begin our considerations note first the fact that $\Omega_{G_0}(U, \mathcal{W})$ is the set of all errors in the output of $\{G_0, P_0\}$ that are determined by the sets U and $\mathcal{W}$, i.e., by all inputs and plant perturbations of interest, respectively.

LEMMA 7. *Let $U \subset H_1$, $\mathcal{W} \subset [H_0, H_2]$, and let $G_0 \in [H_1, H_0]$. If* (i) *U is dense in H_1,* (ii) *$G_0 H_1$ is dense in H_0,* (iii) *$\bigcup_{W \in \mathcal{W}} WH_0$ is dense in H_2, then $\Omega_{G_0}(U, \mathcal{W})$ is dense in H_2.*

Proof. Without loss of generality we can assume that $0 \notin \mathcal{W}$. Choose $x \in H_2$, and let $\varepsilon > 0$. By (iii) there exist $W \in \mathcal{W}$ and $y \in H_0$ such that $\|x - Wy\| < \varepsilon/3$. By (ii) there exists $z \in H_1$ such that $\|y - G_0 z\| < \varepsilon/3\|W\|$. Finally, by (i) there is a $u \in U$ with $\|z - u\| < \varepsilon/3\|W\| \cdot \|G_0\|$. Then $x' = WG_0 u \in \Omega_{G_0}(U, \mathcal{W})$, and the triangular law yields $\|x - x'\| < \varepsilon$.

LEMMA 8. *Let H be a Banach space, and let A_n, $B_n \in [H, H]$, $n = 1, 2, \cdots$. If $A_n x \to x$ and $B_n x \to x$ for each $x \in H$, then $A_n B_n x \to x$ for every $x \in H$.*

Proof. Since $\{A_n x : n = 1, 2, \cdots\}$ is a bounded set for each $x \in H$, the Banach–Steinhaus theorem shows that $\|A_n\| \leqq a$ for all n and some $a > 0$. Thus, for any $x \in H$ we have

$$\|A_n B_n x - x\| \leqq \|A_n(B_n x - x)\| + \|A_n x - x\| \leqq a\|B_n x - x\| + \|A_n x - x\| \to 0.$$

In the following considerations we will assume that $H_0 = H_1 = H_2 = L_2[0, \infty)$, since this is the most important case in our setting.

Let $\alpha > 0$ be a fixed number; for $n = 1, 2, \cdots$, and $t \geqq 0$, let $e_n(t) = \alpha n e^{-\alpha n t}$. Also, if r is a positive integer, we let e_n^{*r} be the r-fold convolution of e_n.

It is clear that $e_n^{*r} \in L_2[0, \infty) \cap L_1[0, \infty)$; moreover, we have

LEMMA 9. *Let $r \geqq 1$ be an integer, and let $x \in L_2[0, \infty)$; then $e_n^{*r} * x \in L_2[0, \infty)$ and $e_n^{*r} * x \to x$ as $n \to \infty$.*

Proof. Since $e_n \in L_2 \cap L_1$, the Fourier–Plancherel transform $\hat{e}_n$ of e_n coincides with the L_1-Fourier transform of e_n, and we have $\hat{e}_n(i\omega) = \alpha n(\alpha n + i\omega)^{-1}$ for $\omega \in R^1$. By Parseval's equality,

$$\|e_n * x - x\|^2 = (2\pi)^{-1}\|\hat{e}_n \hat{x} - \hat{x}\|^2 = (2\pi)^{-1} \int_{-\infty}^{\infty} |\hat{e}_n - 1|^2 \cdot |\hat{x}|^2 \, d\omega \tag{18}$$

for every $x \in L_2[0, \infty)$. However, $\hat{e}_n(i\omega) - 1 \to 0$ pointwise, and $|\hat{e}_n - 1|^2 \cdot |\hat{x}|^2 \leqq |\hat{x}|^2 \in L_1(-\infty, \infty)$. Thus, by the dominated convergence theorem,

$$\int_{-\infty}^{\infty} |\hat{e}_n - 1|^2 \cdot |\hat{x}|^2 \, d\omega \to 0$$

as $n \to \infty$. Hence, by (18), $e_n * x \to x$ as $n \to \infty$.

Finally, since each operator $e_n *$ is bounded, Lemma 8 and the associativity of convolution conclude the proof.

LEMMA 10. *The set $\{e_{2q+1} : q = 0, 1, 2, \cdots\}$ is fundamental in $L_2[0, \infty)$.*

Proof. Using the substitution $\xi = e^{-2\alpha t}$ we confirm easily that $x(t) \in L_2[0, \infty) \Leftrightarrow \xi^{-1/2} x(-1/2\alpha \ln \xi) \in L_2[0, 1]$. Thus, choose $x \in L_2[0, \infty)$, and let $\varepsilon > 0$. By the

Weierstrass theorem for the space $L_2[0, 1]$, there exists a polynomial p such that

$$\int_0^1 \left| \xi^{-1/2} x\left(-\frac{1}{2\alpha} \ln \xi\right) - p(\xi) \right|^2 d\xi < 2\alpha\varepsilon,$$

i.e.,

$$\int_0^1 \left| x\left(-\frac{1}{2\alpha} \ln \xi\right) - \xi^{1/2} p(\xi) \right|^2 \xi^{-1}\, d\xi < 2\alpha\varepsilon. \tag{19}$$

With the above substitution, (19) yields

$$\int_0^\infty |x(t) - e^{-\alpha t} p(e^{-2\alpha t})|^2\, dt < \varepsilon,$$

which proves our claim.

To simplify the wording of the theorems, let us introduce the following notation:

(i) Let $\mathcal{P}$ be the set of all functions $\sum_{i=1}^m P_i(t)\, e^{-\lambda_i t}$, where P_i are polynomials and $\operatorname{Re} \lambda_i > 0$ for all i.

(ii) Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ consisting of all functions k such that the Laplace transform $\hat{k}(s)$ of k has all zeros in the open right half-plane.

Note that, by virtue of Lemma 10, $\mathcal{P}$ is a dense linear subspace of $L_2[0, \infty)$.

THEOREM 3. *Assume that*

(i) *U is a linear subspace of $L_2[0, \infty)$ containing the elements e_{2q+1}, $q = 0, 1, 2, \cdots$,*

(ii) *$G_0 = k_0 *$ with $k_0 \in \mathcal{P}_0$, $k_0 \neq 0$,*

(iii) *$\mathcal{W}$ contains the operators $e_n *$ for $n = 1, 2, \cdots$.*

Then the set $\Omega_{G_0}(U, \mathcal{W})$ is dense in $L_2[0, \infty)$.

Proof. By Lemma 10, U is dense in $L_2[0, \infty)$, i.e., condition (i) in Lemma 7 holds. Next, since the Laplace transform $\hat{k}_0(s)$ of k_0 is a rational function of s, denote $\hat{k}_0 = P/Q$, where P and Q are polynomials without a common factor. Clearly, for degrees of P and Q we have $\partial P < \partial Q$. Now, choose $f \in L_2[0, \infty)$ and let $\hat{f}$ be the Fourier–Plancherel transform of f. Choose a fixed integer $r > 1$ so large that $\partial Q < \partial P + r$, and consider the function

$$\hat{x}_n = \frac{Q}{P}\left(\frac{\alpha n}{\alpha n + i\omega}\right)^r \hat{f}.$$

Using the theorem on a transform of a convolution it follows readily that $\hat{x}_n$ is a Fourier–Plancherel transform of some $x_n \in L_2[0, \infty)$, and that $k_0 * x_n = e_n^{*r} * f$. Thus, by Lemma 9, $k_0 * x_n = G_0 x_n \to f$. Hence, $G_0 L_2[0, \infty)$ is dense in $L_2[0, \infty)$; i.e., condition (ii) in Lemma 7 is satisfied.

Finally, the assumption (iii) implies by Lemma 9 that condition (iii) in Lemma 7 is fulfilled. Hence, $\Omega_{G_0}(U, \mathcal{W})$ is dense in $L_2[0, \infty)$ as claimed.

Theorem 3 clearly shows that if G_0 is a convolution operator with kernel $k_0 \in \mathcal{P}_0$, if the set of inputs of interest U is as narrow as the collection of "polynomials" $\sum_{j=0}^m c_j e_{2j+1}$, and if the set of perturbations of interest $\mathcal{W}$ consists only of the operators $e_n *$, then the set of all errors $\Omega_{G_0}(U, \mathcal{W})$ is already dense in $L_2[0, \infty)$.

It is interesting to note that if $\mathcal{W}$ is suitably enlarged, then U can be as meager as a singleton. Indeed, we have the following assertion:

THEOREM 4. Assume that

(i) *U contains the function $ae^{-\alpha t}$ with $\alpha > 0$, $a \neq 0$,*

(ii) *$G_0 = k_0 *$ with $k_0 \in \mathcal{P}_0$, $k_0 \neq 0$,*

(iii) *$\mathcal{W}$ contains the collection* $\{k*: k \in \mathcal{P}\}$.

Then the set $\Omega_{G_0}(U, \mathcal{W})$ *is dense in* $L_2[0, \infty)$.

Proof. Choose $x \in L_2[0, \infty)$ and let $\varepsilon > 0$. By Lemma 10 there exists a linear combination f of $e_1, e_3, e_5, \cdots$ such that $\|x - f\| < \varepsilon/2$. Denote $\hat{f}$ and $\hat{k}_0$ the Fourier transform of f and k_0, respectively. Since these are rational functions, put $\hat{k}_0 = P/Q$ and $\hat{f} = \tilde{P}/\tilde{Q}$, where P, Q, $\tilde{P}$, $\tilde{Q}$ are polynomials.

Next, choose integer $r \geqq 1$ so large that $1 + \partial Q + \partial \tilde{P} < r + \partial P + \partial \tilde{Q}$, and put

$$\hat{k}_n = \left(\frac{\alpha n}{\alpha n + i\omega}\right)^r \frac{Q}{P} \frac{\alpha + i\omega}{a} \frac{\tilde{P}}{\tilde{Q}}. \tag{20}$$

From (20) it follows easily that $\hat{k}_n$ is the Fourier transform of some $k_n \in \mathcal{P}$, and, by properties of a transform of a convolution, that

$$k_n * (G_0 u_0) = k_n * (k_0 * u_0) = e_n^{*r} * f, \tag{21}$$

where $u_0 = a e^{-\alpha t}$. Thus, by Lemma 9, $k_n * (G_0 u_0) \to f$; i.e., there is an n such that $\|f - k_n * (G_0 u_0)\| < \varepsilon/2$. Hence, by the above, $\|x - k_n * (G_0 u_0)\| < \varepsilon$, which completes the proof.

Note that alternatives of Theorems 3 and 4 dealing with more general families of exponentials can easily be obtained by using Müntz's theorem, (see [4, p. 197]), but we omit the rather obvious details.

REFERENCES

[1] R. SAEKS, *Resolution Space, Operators and Systems*, Springer-Verlag, New York, 1973.

[2] V. DOLEZAL, *A sensitivity analysis of systems consisting of linear blocks*, this Journal, 10 (1979), pp. 1121–1137.

[3] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input–Output Properties*, Academic Press, New York, 1975.

[4] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[5] W. A. PORTER AND C. L. ZAHM, *Basic Concepts in System Theory*, University of Michigan, SEL Tech. Report #44, 1969.

[6] R. W. NEWCOMB AND B. D. O. ANDERSON, *A distributional approach to time-varying sensitivity*, SIAM J. Appl. Math., 15 (1967), pp. 1001–1010.

[7] R. M. DESANTIS AND W. A. PORTER, *A generalized Nyquist plot and its use in sensitivity analysis*, Int. J. Systems Sci., 5 (1974), pp. 1143–1153.

[8] W. A. PORTER AND R. M. DESANTIS, *Sensitivity analysis in multilinear systems*, Int. J. Systems Sci., 7 (1976), pp. 191–205.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0005 $01.00/0

PERISTALTIC TRANSPORT OF A FLUID-PARTICLE MIXTURE*

M. C. SHEN,† K. C. LIN† AND S. M. SHIH‡

Abstract. An asymptotic method is developed for the solution of the mathematical problem of peristaltic transport of a viscous fluid with solid particles in a flexible tube of arbitrary cross-section. Under long wave approximation, the coupled nonlinear equations governing the fluid-particle mixture are reduced to a sequence of two-dimensional, linear boundary-value problems. The asymptotic method is justified rigorously and the existence and uniqueness of the exact solution are proved.

1. Introduction. The problem of fluid transport through a flexible tube by peristaltic motion of the tube wall has attracted a great deal of attention in recent years. It has played an important role in many physiological processes, biomechanical devices and engineering applications. The early mathematical models for peristaltic transport are based upon the Navier-Stokes equations subject to a prescribed transverse displacement at the wall, and a survey of the research work on this problem up to 1971 was given in [3]. Some recent work may be found in [7], [8], [9], [15] and others.

The early models become inadequate in applications when more than a fluid phase is transported by peristalsis, and refined models have to be used. Hung and Brown [2] made an empirical study of the peristaltic transport of a fluid with solid particles based upon a two-dimensional model. An axisymmetric model of this problem was analyzed by Kaimal [5]. In this paper, we shall develop an asymptotic method for the three-dimensional peristaltic transport of an incompressible viscous fluid with solid particles. The governing equations for a liquid-particle mixture have been discussed by several authors [10], [11], [12], [14]. The model we shall adopt is simply based upon the two-fluid theory, assuming that constitutive equations for both fluid and particulate phases are of similar forms and that the interaction forces between the two phases are proportional to the difference of their velocities. Applications of this model may be found in [2] and [5]. For simplicity, we shall neglect the pressure of the particulate phase. However, our method applies equally well to the case in which this pressure is present. To neglect the particulate pressure only simplifies the asymptotic scheme, but has no effect at all upon the variational formulation of the problem. Some discussions on omitting the particulate pressure based upon physical grounds may be found in [10] and [11].

The basic ideas used in our asymptotic method are motivated by the study of surface waves on a viscous fluid [13]. In general, there are three length scales at our disposal; the amplitude A, the wavelength L of the prescribed peristaltic motion, and the maximum diameter d of the tube. Within the framework of long wave approximation, $\alpha = A/L$ is assumed to be a small parameter, but A/d is assumed to be of order unity. In the problem, two Reynolds numbers, R_1, R_2, corresponding to the fluid and particulate phases, respectively, will appear, which are also assumed to be of order unity. We note that our results do not apply when $1/R_2 = 0$, as is the case in [12]. However, by formulating the transport problem as a time-dependent problem, it can be shown that under certain conditions the solution of the problem tends to the solution of the governing equations with $1/R_2 = 0$ as $R_2 \to \infty$. We defer the details to a subsequent study. Furthermore, as shown in [3], the proper Reynolds number correctly describing the ratio of inertia to viscous terms is αR_1; thus restriction of R_1 to $O(1)$ really restricts

* Received by the editors December 3, 1979, and in final form June 5, 1980. This research was supported by the National Science Foundation under grant MCS 77-00097.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

‡ Institute of Mathematics, Academia Sinica, Taiwan, R.O.C.

the proper Reynolds number to $O(\alpha)$. We formally expand the solution of our problem in a power series of α and the successive approximations are determined by a sequence of linear, two-dimensional, elliptic boundary value problems. Our contributions here are to justify rigorously the asymptotic method and to prove the existence and uniqueness of the exact solution of the problem. The methods of proof are extensions of those given in [6], [13].

We formulate the problem in § 2. In § 3, a formal asymptotic method is developed under the long-wave approximation and is rigorously justified in § 4 under some restriction on the two Reynolds numbers. Finally, in § 5, we prove the existence and uniqueness of the exact solution.

2. Formulation of the problem. We consider the motion of an incompressible fluid of constant density ρ_1 with particles of constant density ρ_2 in a long flexible tube. A transverse displacement in the form of a progressive wave of period T moving at constant speed c in the axial direction is prescribed on the tube wall, and there is no longitudinal displacement. Furthermore, we assume that the minimum radius of the tube does not vanish. In reference to a coordinate system (x^*, y^*, z^*) moving with the wave, the boundary of the tube is stationary and the equations governing the steady flow of the fluid-particle mixture are assumed to be the following, for $i, j = 1, 2$, $i \neq j$:

$$\nabla^* \cdot \mathbf{q}^{i*} = 0, \tag{1}$$

$$\mathbf{q}^{i*} \cdot \nabla^* \mathbf{q}^{i*} = -\delta_{i1} \nabla^* p^* / \rho_i + \upsilon_i \Delta^* \mathbf{q}^{i*} + M_i^* (\mathbf{q}^{*j} - \mathbf{q}^{*i}), \tag{2}$$

$$\mathbf{q}^{*i} = \mathbf{q}_b^* = (-c, cf^*/L, cg^*/L) \quad \text{at the boundary } H^*(x^*, y^*, z^*,) = 0. \tag{3}$$

Here the scripts 1 and 2 denote the quantities pertaining to the fluid and particulate phases respectively, $\mathbf{q}^{i*}$ are the velocities, p^* is the pressure, υ_i are the kinematic viscosities, M_i^* are positive constants, f^* and g^* are two given functions of x^*, y^*, and z^*, $\nabla^* = (\partial/\partial x^*, \partial/\partial y^*, \partial/\partial z^*)$, and $\delta_{i1} = 1$ for $i = 1$, $\delta_{i1} = 0$ for $i \neq 1$. We now measure x^*, y^*, z^*, f^* and g^* in units of A, $\mathbf{q}^{i*}$ in units of c, p^* in units of $\rho_1 c^2$, and define $R_i = cA/\upsilon_i$, $M_i = M_i^* A/c$. In terms of the nondimensional variables without a star, (1) to (3) become

$$\nabla \cdot \mathbf{q}^i = 0, \tag{4}$$

$$(\mathbf{q}^i - \mathbf{i}) \cdot \nabla \mathbf{q}^i = -\delta_{i1} \nabla p + R_i^{-1} \Delta \mathbf{q}^i + M_i (\mathbf{q}^j - \mathbf{q}^i), \tag{5}$$

$$\boldsymbol{q}^i = (O, \alpha f, \alpha g) \quad \text{at } H = 0, \tag{6}$$

where $\mathbf{q}^i = \mathbf{q}^{i*}/c + \mathbf{i}$ and $\mathbf{i}$ is the unit vector in the x-direction. We note here that the boundary condition given in (3) follows from the assumptions that there is no longitudinal displacement and the fluid and particles satisfy the no-slip condition at the boundary.

Let Ω be the domain defined by

$$\Omega = \{(x, y, z) \,|\, O \leqq x \leqq 1/\alpha, (y, z) \in D\},$$

where $1/\alpha$ is the period of the progressive wave by nondimensionalization and D is any open cross-section of the tube. We shall look for a solution of (4)–(6) with the same period in some suitable function space. Let $\mathring{J}(\Omega)$ be the completion of $\dot{J}(\Omega)$, the space of solenoidal c^∞-functions of compact support in Ω and period T in x, with the scalar product

$$(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \cdot \mathbf{v} \, d\Omega.$$

Let $H(\Omega)$ be the completion of $\dot{J}(\Omega)$ with the scalar product

$$(\mathbf{u}, \mathbf{v})_H = \int_\Omega \sum_{i=1} \sum_{j=1} (\partial u_i/\partial x_j)(\partial v_i/\partial x_j)\, d\Omega,$$

$$= \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v}\, d\Omega,$$

where u_i, v_i are respectively the components of $\mathbf{u}$ and $\mathbf{v}$. The space $H \times H$ is the direct product of H and H with the scalar product

$$(\mathbf{U}, \mathbf{V})_{H\times H} = (\mathbf{u}_1, \mathbf{v}_1)_H + (\mathbf{u}_2, \mathbf{v}_2)_H,$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2)$, $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2)$. A generalized solution of (4)–(6) is defined as a pair of functions $(\mathbf{q}^1, \mathbf{q}^2)$, satisfying the integral identities

$$\int_\Omega [-\mathbf{q}^i_x + \mathbf{q}^i \cdot \nabla \mathbf{q}^i + M^i(\mathbf{q}^i - \mathbf{q}^j)] \cdot \boldsymbol{\phi}\, d\Omega + \int_\Omega R_i^{-1} \nabla \mathbf{q}^i \cdot \nabla \boldsymbol{\phi}\, d\Omega = 0, \tag{7}$$

$$i, j = 1, 2, \quad i \neq j,$$

for any $\boldsymbol{\phi} \in H$ and for smooth functions $\mathbf{a}^i$, $\mathbf{q}^i - \mathbf{a}^i \in H$, where $\mathbf{a}^i = \mathbf{q}_b$ on $\partial\Omega$, the boundary of Ω. In what follows, we shall develop and justify an asymptotic method for finding an approximation to the generalized solution. The existence and uniqueness of the generalized solution are proved in the last section.

3. Formal asymptotic expansions. In this section we formally set up an asymptotic scheme to solve the equations (4) to (6). In doing this, we assume that α is a small parameter, $\partial/\partial x = O(\alpha)$, and $\mathbf{q}^i$, p have asymptotic expansions

$$\mathbf{q}^i = \mathbf{q}^i_0 + \mathbf{q}^i_1 + \mathbf{q}^i_2 + \cdots,$$

$$p = p_{-1} + p_0 + p_1 + \cdots,$$

where $\mathbf{q}^i_k = (u^i_k, v^i_k, w^i_k) = O(\alpha^k)$, $\mathbf{q}^i_0 = (u^i_0, 0, 0)$, for $i = 1, 2$, and $p_k = O(\alpha^k)$. Substitution of the series for $\mathbf{q}^i$ and p in (4)–(6) yields a sequence of equations and boundary conditions for the successive approximations. The equations for the first approximations are

$$u^i_{0x} + v^i_{1y} + w^i_{1z} = 0, \tag{8}$$

$$R_i^{-1} \Delta_2 u^i_0 = \delta_{i1} p_{-1x} + M_i(u^i_0 - u^j_0), \tag{9}$$

$$p_{-1y} = p_{-1z} = 0, \tag{10}$$

$$u^i_0 = 0, \quad v^i_1 = \alpha f, \quad w^i_1 = \alpha g \quad \text{on } \partial\Omega, \tag{11}$$

where $\Delta_2 = \partial^2/\partial y^2 + \partial^2/\partial z^2$. By dividing (9) by M_i, setting $i, j = 1, 2$, and adding the equations, we obtain

$$\Delta_2[(M_1R_1)^{-1}u^1_0 + (M_2R_2)^{-1}u^2_0] = p_{-1x}/M_1. \tag{12}$$

As seen from (10), p_{-1} is a function of x only, and we may set

$$(M_1R_1)^{-1}u^1_0 + (M_2R_2)^{-1}u^2_0 = -u_0 p_{-1x}/M_1. \tag{13}$$

It follows from (11) and (12) that

$$\Delta_2 u_0 = -1 \quad \text{in } D, \tag{14}$$

$$u_0 = 0 \quad \text{on } \partial D, \tag{15}$$

where ∂D is the boundary of D.

Let $i=2$ in (9), and eliminate u_0^1 in the resulting equation by (13) to obtain

$$R_2^{-1}\Delta_2 u_0^2=[M_2+M_1R_1/(M_2R_2)]u_0^2=u_0R_1p_{-1x}/M_1. \tag{16}$$

Set

$$u_0=-u_{01}p_{-1x}, \tag{17}$$

and from (11), (13), and (16) it follows that

$$R_2^{-1}\Delta_2 u_{01}^2-[M_2+M_1R_1/(M_2R_2)]u_{01}^2=-u_0R_1/M_1 \quad \text{in } D, \tag{18}$$

$$u_{01}^2=0 \quad \text{on } \partial D, \tag{19}$$

$$u_{01}^1=M_1R_1[u_0/M_1-(M_2R_2)^{-1}u_{01}^2]. \tag{20}$$

Since the flow is steady, it is easily shown that

$$\int_D (u_0^i-1)\,dA=c_0^i, \tag{21}$$

where c_0^i are constants assumed to be given. From (17), we have

$$u_0^1+u_0^2=-p_{-1x}(u_{01}^1+u_{02}^2).$$

By integrating the above equation over D, and making use of (21), we obtain

$$p_{-1x}=-(c_0+2A)\left[\int_D (u_{01}^1+u_{02}^2)\,dA\right]^{-1}, \tag{22}$$

where $c_0=c_0^1+c_0^2$, and $A=\int_D dA$.

The equations for the second approximations are

$$u_{1x}^i+v_{2y}^i+w_{2z}^i=0, \tag{23}$$

$$R_i^{-1}u_1^i=\delta_{i1}p_{0x}+M_i(u_1^i-u_1^j)+(u_0^i-1)u_{0x}^i+v_1^iu_{0y}^i+w_1^iu_{0z}^i, \tag{24}$$

$$p_{0y}=p_{0z}=0, \tag{25}$$

$$u_1^i=v_2^i=w_2^i=0 \quad \text{on } \partial D. \tag{26}$$

Assuming v_1^i, w_1^i have been found, by (24) and (25) we may set

$$u_1^i=-u_{01}^ip_{0x}+u_{11}^i, \tag{27}$$

where u_{11}^i satisfy

$$R_i^{-1}\Delta_2 u_{11}^i=M_i(u_{11}^i-u_{11}^j)+(u_0^i-1)u_{0x}^i+v_1^iu_{0y}^i+w_1^iu_{0z}^i, \tag{28}$$

$$u_{11}^i=0 \quad \text{on } \partial D, \tag{29}$$

and u_{01}^i are determined by (18)–(20); (28) and (29) can be reduced in the same way to a Poisson equation and an inhomogeneous Helmholtz equation as before. To determine p_{0x} we make use of the integral invariants

$$\int_D u_1^i\,dA=c_1^i, \tag{30}$$

as a consequence of (23) and (26), where c_1^i are given constants. By adding the two

equations in (27) and integrating the resulting equation over D, it is obtained that

$$p_{0x} = -\left[c_1 - \int_D (u_{11}^1 + u_{11}^2)\, dA\right]\left[\int_D (u_{01}^1 + u_{01}^2)\, dA\right]^{-1}, \tag{31}$$

where $c_1 = c_1^1 + c_1^2$.

The equations for the third and fourth approximations are

$$u_{2x}^i + v_{1y}^i + w_{1z}^i = 0, \tag{32}$$

$$\begin{aligned} R_i^{-1}\Delta_2 u_2^i = {} & \delta_{i1} p_{1x} + M_i(u_2^i - u_2^j) + (u_0^i - 1)u_{1x}^i + u_1^i u_{0x}^i \\ & + v_1^i u_{1y}^i + v_2^i u_{0y}^i + w_1^i u_{1z}^i + w_2^i u_{0z}^i, \end{aligned} \tag{33}$$

$$R_i^{-1}\Delta_2 v_1^i = \delta_{i1} p_{1y} + M_i(v_1^i - v_1^j), \tag{34}$$

$$R_i^{-1}\Delta_2 w_1^i = \delta_{i1} p_{1z} + M_i(w_1^i - w_1^j), \tag{35}$$

$$R_i^{-1}\Delta_2 v_2^i = \delta_{i1} p_{2y} + M_i(v_2^i - v_2^j) + (u_0^i - 1)v_{1x}^i + v_1^i v_{1y}^i + w_1^i v_{1z}^i, \tag{36}$$

$$R_i^{-1}\Delta_2 w_2^i = \delta_{i1} p_{2z} + M_i(w_2^i - w_2^j) + (u_0^i - 1)w_{1x}^i + v_1^i w_{1y}^i + w_1^i w_{1z}^i, \tag{37}$$

$$u_2^i = v_k^i = w_k^i = 0 \quad \text{on } \partial D \quad \text{for } i, j, k = 1, 2, \quad i \neq j. \tag{38}$$

We now introduce functions Q_k^i $i, k = 1, 2$, satisfying

$$\Delta_2 Q_k^i = -u_{(k-1)x}^i \quad \text{in } D, \tag{39}$$

$$\begin{aligned} Q_{kn}^i &= \alpha(fn_1 + gn_2) \quad \text{on } \partial D \quad \text{for } k = 1, \\ &= 0 \qquad\qquad\quad \text{on } \partial D \quad \text{for } k = 2, \end{aligned} \tag{40}$$

where the subscript n denotes differentiation in the direction of an outward normal (n_1, n_2) to ∂D. Equations (39) and (40) are solvable because of (8), (11), (23), (26), (32) and (38). It follows from (8), (23) and (32) that

$$(v_k^i - Q_{ky}^i)_y + (w_k^i - Q_{kz}^i)_z = 0. \tag{41}$$

As seen from (41), we may introduce functions Φ_k^i, $i, k = 1, 2$, such that

$$v_k^i = \Phi_{kz}^i + Q_{ky}^i, \qquad w_k^i = -\Phi_{ky}^i + Q_{kz}^i, \tag{42}$$

and replace v_k^i, w_k^i in (34) to (37) by (42). Upon cross-differentiation of (34) and (35), we obtain

$$R_i^{-1}\Delta_2^2 \Phi_1^i = M_i \Delta_2(\Phi_1^i - \Phi_1^j). \tag{43}$$

Let

$$\Phi_1 = \Phi_1^2/M_1R_1 + \Phi_1^2/M_2R_2. \tag{44}$$

Then it follows from (11), (42) and (43) that

$$\Delta_2^2 \Phi_1 = 0, \tag{45}$$

subject to the boundary conditions on ∂D,

$$\Phi_{1y} = (Q_{1z}^1 - \alpha g)/M_1R_1 + (Q_{1z}^2 - \alpha g)/M_2R_2, \tag{46}$$

$$\Phi_{1z} = (-Q_{1y}^1 + \alpha f)/M_1R_1 + (-Q_{1y}^2 + \alpha f)/M_2R_2. \tag{47}$$

By using (44) to eliminate Φ_1^1 in (43) for $i = 2$, we have

$$R_2^{-1}\Delta_2^2 \Phi_1^2 - (M_2 + R_2^{-1})\Delta_2 \Phi_1^2 = -M_2 \Delta_2 \Phi_1, \tag{48}$$

$$\Phi^2_{1y} = Q^2_{1z} - \alpha g, \qquad \Phi^2_{1z} = -Q^2_{1y} + \alpha f \quad \text{on } \partial D. \tag{49}$$

By the same token, it is easily obtained from (36) and (37) that Φ^i_2 satisfy

$$\begin{aligned} R_i^{-1}\Delta_2\Phi^i_2 = M_i\Delta_2(\Phi^i_2 - \Phi^i_2) + [(u^i_0 - 1)v^i_{1x} + v^i_1 v^i_{1y} + w^i_1 v^i_{1z}]_z \\ - [(u^i_0 - 1)w^i_{1x} + v^i_1 w^i_{1y} + w^i_1 w^i_{1z}]_y, \end{aligned} \tag{50}$$

subject to

$$\Phi^i_{2y} = Q^i_{2z}, \Phi^i_{2z} = -Q^i_{2y} \quad \text{on } \partial D. \tag{51}$$

Here, (50) and (51) may be reduced to equations similar to (45) and (48) with additional terms. If Q^i_k, Φ^i_k are obtained from (39), (40) and (44) to (50), then v^i_k will be determined from (42) and p_1, p_2, from (34) to (37) by integration. Higher order approximations can be found in the same manner and we shall not proceed any further.

4. Justification of the asymptotic method. Assuming that a generalized solution of (4)–(6) exists, we shall show that the asymptotic solution found in § 3 is indeed an asymptotic approximation to the generalized solution. Let

$$\mathbf{q}^i = \mathbf{q}^i_0 + \mathbf{q}^i_1 + \mathbf{q}^i_*, \qquad i = 1, 2, \tag{52}$$

where

$$\begin{aligned} \mathbf{q}^i_k &= (u^i_k, v^i_k, w^i_k) = (q^i_{k1}, q^i_{k2}, q^i_{k3}), \\ \mathbf{q}^i_* &= (q^i_{*1}, q^i_{*2}, q^i_{*3}). \end{aligned} \tag{53}$$

Since the components of $\mathbf{q}^i_0$ and $\mathbf{q}^i_1$ are determined by solutions of elliptical boundary-value problems and $\mathbf{q}^i_0 + \mathbf{q}^i_1 = \mathbf{q}_b$ on $\partial\Omega$, they are sufficiently smooth if the prescribed displacements on $\partial\Omega$ are sufficiently smooth. Hence, by definition, $\mathbf{q}^i_* \in H$. Substituting (52) for $\mathbf{q}^i$ in (7) and rearranging the terms, we obtain

$$\begin{aligned} \int_\Omega [-\mathbf{q}^i_{*x} + (\mathbf{q}^i_0 + \mathbf{q}^i_1 + \mathbf{q}^i_*)\cdot\nabla\mathbf{q}^i_* - \mathbf{G}^i + \mathbf{q}^i_*\cdot\nabla(\mathbf{q}^i_0 + \mathbf{q}^i_1) + M_i(\mathbf{q}^i_* - \mathbf{q}^j_*)]\cdot\boldsymbol{\phi}\, d\Omega \\ + R_i^{-1}\int_\Omega \nabla\mathbf{q}^i_*\cdot\nabla\boldsymbol{\phi}\, d\Omega = 0, \end{aligned} \tag{54}$$

where

$$\mathbf{G}^i = R_i^{-1}\Delta(\mathbf{q}^i_0 + \mathbf{q}^i_1) - [(\mathbf{q}^i_0 + \mathbf{q}^i_1)\cdot\nabla](\mathbf{q}^i_0 + \mathbf{q}^i_1) + (\mathbf{q}^i_0 + \mathbf{q}^i_1)_x + M_i(\mathbf{q}^j_0 + \mathbf{q}^j_1 - \mathbf{q}^i_0 - \mathbf{q}^i_1). \tag{55}$$

By choosing $\boldsymbol{\phi} = \mathbf{q}^i_*$ in (54), performing integration by parts and making use of the periodicity of q^i_*, (54) becomes

$$\begin{aligned} R_i^{-1}\|\mathbf{q}^i_*\|^2_H = \int_\Omega \mathbf{G}^i\cdot\mathbf{q}^i_*\, d\Omega + \int_\Omega \sum_{l=1}^{3}\sum_{k=1}^{3} q^i_{*k}(\partial q^i_{*l}/\partial x_k)(q^i_{0l} + q^i_{1l})\, d\Omega \\ + M_i\int_\Omega (\mathbf{q}^j_* - \mathbf{q}^i_*)\cdot\mathbf{q}^i_*\, d\Omega, \end{aligned} \tag{56}$$

where $\|\cdot\|_H$ is the norm on H. We shall estimate the right-hand side of (56) to get an upper bound for $\|\mathbf{q}^i_*\|^2_H$. First, we establish several lemmas.

LEMMA 1.

$$\|\boldsymbol{\phi}\| \leqq K\|\boldsymbol{\phi}\|_H \quad \textit{for any } \boldsymbol{\phi} \in H,$$

where $K \leqq 1/\sqrt{2}$, *and* $\|\cdot\|$ *is the* L_2*-norm.*

Lemma 1 is the well-known Poincaré inequality, and a proof may be found in [4].

LEMMA 2. $u_0^i = O(1)$, u_k^i, v_k^i, $w_k^i = O(\alpha^k)$ *for* $k = 1, 2$, $p_{kx} = O(\alpha^{k+1})$ *for* $k = -1, 0, 1$, *and* $\partial/\partial x = O(\alpha)$.

Proof. The proof of Lemma 2 follows directly from the regularity of the solutions of elliptical equations up to the boundary if $\mathbf{q}_b$, the prescribed velocity on $\partial\Omega$, is sufficiently smooth [1]. For example, u_0, u_{01}^2 are solutions of (14), (15), (18) and (19). They are regular up to the boundary and $O(1)$ for a fixed x in $0 \leqq x \leqq 1/\alpha$. By differentiating (14), (15), (18) and (19) with respect to x, and letting $\xi = \alpha x$, we obtain

$$\Delta_2 u_{0\xi} = 0 \quad \text{in } D \quad \text{for } 0 \leqq \xi \leqq 1,$$

$$u_{0\xi} = u_{0y} f + u_{0z} g \quad \text{on } \partial D,$$

$$R_2^{-1} \Delta_2 u_{01\xi}^2 - [M_2 + M_1 R_1/(M_2 R_2)] u_{01\xi}^2 = -u_{0\xi} R_1 / M_1 \quad \text{in } D \quad \text{for } 0 \leqq \xi \leqq 1,$$

$$u_{01\xi}^2 = u_{01y}^2 f + u_{01z}^2 g \quad \text{on } \partial D,$$

where $\partial y/\partial x = -\alpha f$, $\partial z/\partial x = -\alpha g$ on ∂D. Hence, $u_{0\xi} = \alpha^{-1} u_{0x}$. $u_{01\xi}^2 = \alpha^{-1} u_{01x}^2$ are regular and $O(1)$, and u_{0x}, u_{01x}^2 are regular and $O(\alpha)$ for a fixed x in $0 \leqq x \leqq 1/\alpha$. This shows that both u_0 and u_{01}^2 are regular and $O(1)$ in $\mathbf{\Omega} = \Omega + \partial\Omega$, and $\partial/\partial x$ is $O(\alpha)$ as applied to u_0 and u_{01}. To show p_{-1x} is regular, we need a different expression for p_{-1x}, although (22) is more convenient for applications. From (13) and (21) it is obtained that

$$p_{-1x} = -M_1 \{(M_1 R_1)^{-1} c_0^1 + (M_2 R_2)^{-1} c_0^2 + [(M_1 R_1)^{-1} + (M_2 R_2)^{-1}] A\} \Big(\int_D u_0 \, dA \Big)^{-1},$$

where $\int_D u_0 \, dA = -\int_D u_0 \Delta_2 u_0 \, dA = \int_D (\nabla_2 u_0)^2 \, dA > 0$, $\nabla_2 = (\partial/\partial y, \partial/\partial z)$ since the minimum radius of the tube never vanishes. Therefore, p_{-1x} is regular and $O(1)$. It follows from (17) that u_0^i are regular and $O(1)$. The rest of the lemma may be proved similarly, and we shall not go into details.

Remark. If a function $f(x, y, z)$ is regular in $\mathbf{\Omega}$ and $O(\alpha^n)$, then

$$\|f\| = \Big[\int_0^{\alpha^{-1}} \Big(\int_D |f|^2 \, dA \Big) \, dx \Big]^{1/2} = O(\alpha^{n-1/2}).$$

Hence, by the lemma,

$$\|u_0^i\| = O(\alpha^{1/2}), \qquad \|u_k^i\|, \|v_k^i\|, \|w_k^i\| = O(\alpha^{k-1/2})$$

for $k = 1, 2$.

LEMMA 3.

$$\Big| \int_\Omega^{\mathbf{G}^i} \cdot \mathbf{q}_*^i \, d\Omega \Big| \leqq c_i \alpha^{3/2} \|\mathbf{q}_*^i\|_H,$$

where c_i *are constants independent of* α.

Proof. First we consider $i = 1$. Since $\mathbf{q}_*^1 \in H$, we can add $-\nabla(p_{-1} + p_0 + p_1)$ to $\mathbf{G}^1$ without changing the value of $\int_\Omega \mathbf{G}^1 \cdot \mathbf{q}_*^1 \, d\Omega$. By (9), (24), (33) to (37) and (55), we have

$$\begin{aligned}
\mathbf{G}_*^1 &= \mathbf{G}^1 - \nabla(p_{-1} + p_0 + p_1), \\
&= \mathbf{i}[R_1^{-1} \Delta_2 u_2^1 + R_1^{-1} (u_0^1 + u_1^1)_{xx} - \mathbf{q}_1^1 \cdot \nabla u_1^1] \\
&\quad + \mathbf{j}[R_1^{-1} v_{1xx}^1 + R_1^{-1} \Delta v_2^1 + (v_1^1 + v_2^1)_x + (\mathbf{q}_0^1 + \mathbf{q}_1^1) \cdot \nabla(v_1^1 + v_2^1) + M_1(v_2^2 - v_2^1)] \\
&\quad + \mathbf{k}[R_1^{-1} w_{1xx}^1 + R_1^{-1} \Delta w_2^1 + (w_1^1 + w_2^1)_x + (\mathbf{q}_0^1 + \mathbf{q}_1^1) \cdot \nabla(w_1^1 + w_2^1) + M_1(w_2^2 - w_2^1)],
\end{aligned}$$

where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are respectively unit vectors in the x, y and z directions.

From Lemma 2 it is easily seen that $|\mathbf{G}_*^1| = O(\alpha^2)$ and $\|\mathbf{G}_*^1\| = O(\alpha^{3/2})$. Similarly,

$$\mathbf{G}^2 = \mathbf{G}_*^2 = \mathbf{i}(R_2^{-1}\Delta_2 u_2^2 + R_2^{-1}(u_0^2 + u_1^2)_{xx} - \mathbf{q}_1^2 \cdot \nabla u_1^2)$$
$$+ \mathbf{j}[(R_2^{-1}v_{1xx}^2 + R_2^{-1}\Delta v_2^2 + (v_2^2 + v_2^2)_x + (q_0^{-2} + q_1^{-2}) \cdot \nabla(v_1^2 + v_2^2) + M_2(v_2^1 - v_2^2)]$$
$$+ \mathbf{R}[R_2^{-1}w_{1xx}^2 + R_2^{-1}\Delta w_2^2 + (w_1^2 + w_2^2)_x + (\mathbf{q}_0^2 + \mathbf{q}_1^2) \cdot \nabla(w_1^2 + w_2^2) + M_2(w_2^1 - w_2^2)],$$

and by Lemma 2, again, $|\mathbf{G}_x^2| = O(\alpha^2)$ and $\|\mathbf{G}_*^2\| = O(\alpha^{3/2})$. Hence,

$$\left|\int_\Omega^{\mathbf{G}^i} \cdot \mathbf{q}_*^i \, d\Omega\right| = \left|\int_\Omega^{\mathbf{G}_*^i} \cdot \mathbf{q}_*^i \, d\Omega\right|$$
$$\leqq \|\mathbf{G}_*^i\| \, \|\mathbf{q}_*^i\| \leqq c_i \alpha^{3/2} \|\mathbf{q}_*^i\|_H,$$

by Schwarz's inequality and Lemma 1.

LEMMA 4.

$$\left|\sum_k \sum_l \int_\Omega q_{*k}^i (\partial q_{*l}^i / \partial xk)(q_{0l}^i + q_{1l}^i) \, d\Omega\right| \leqq (1/\sqrt{2}) \|\mathbf{q}_0^i + \mathbf{q}_1^i\|_\infty \|\mathbf{q}_x^i\|_H^2,$$

where

$$\|\mathbf{q}_0^i + \mathbf{q}_1^i\|_\infty = \sup_\Omega \left[\sum_{k=1}^3 (q_{0k}^i + q_{1k}^i)^2\right]^{1/2}.$$

Proof. This lemma follows from Schwarz's inequality and Lemma 1.

Now we are in a position to prove the following:

THEOREM 1. *If*

$$1 - (R_1 + R_2)/2 - (R_1\|\mathbf{q}_0^1 + \mathbf{q}_1^1\|_\infty + R_2\|\mathbf{q}_0^2 + \mathbf{q}_1^2\|_\infty)/\sqrt{2} - (M_1R_1 + M_2R_2)/4 > 0,$$

then

$$\|\mathbf{q}^i - \mathbf{q}_0^i - \mathbf{q}_1^i\| = O(\alpha^{3/2}).$$

Proof. It follows from (56) and Lemmas 2–4 that

$$R_1^{-1}\|\mathbf{q}_*^i\|_H^2 \leqq c_i\alpha^{3/2}\|\mathbf{q}_*^i\|_H + \|\mathbf{q}_0^i + \mathbf{q}_1^i\|_\infty \|\mathbf{q}_*^i\|_H^2/\sqrt{2}$$
$$+ M_i\left|\int_\Omega \mathbf{q}_*^j \cdot \mathbf{q}_*^i \, d\Omega\right| - M_i\|\mathbf{q}_*^i\|^2,$$

and by Schwarz's inequality, Lemma 1 and $ab \leqq (a^2 + b^2)/2$,

$$\|\mathbf{q}_*^i\|_H^2 \leqq R_i c_i^2 \alpha^3/2 + R_i\|\mathbf{q}_*^i\|_H^2/2 + R_i\|\mathbf{q}_0^i + \mathbf{q}_1^i\|_\infty \|\mathbf{q}_*^i\|_H^2/\sqrt{2}$$
$$+ M_iR_i(\|\mathbf{q}_*^j\|_H^2 + \|\mathbf{q}_*^i\|_H^2)/4. \tag{57}$$

Let $\mathbf{Q}_* = (\mathbf{q}_*^1, \mathbf{q}_*^2)$. Then, by adding the two equations in (57), we have

$$\|\mathbf{Q}_*\|_{H\times H}^2 = \|\mathbf{q}_*^1\|_H^2 + \|\mathbf{q}_*^2\|_H^2 \leqq (R_1 + R_2)C^2\alpha^3/2 + (R_1 + R_2)\|\mathbf{Q}_*\|_{H\times H}^2/2$$
$$+ (R_1\|\mathbf{q}_0^1 + \mathbf{q}_1^1\|_\infty + R_2\|\mathbf{q}_i^2 + \mathbf{q}_1^2\|_\infty)\|\mathbf{Q}_*\|_{H\times H}^2/\sqrt{2} \tag{58}$$
$$+ (M_1R_1 + M_2R_2)\|\mathbf{Q}_*\|_{H\times H}^2/4,$$

where $C^2 = \sum_{i=1}^{2} c_i^2$. If

$$1-(R_1+R_2)/2-(R_1\|\mathbf{q}_0^1+\mathbf{q}_1^1\|_\infty+R_2\|\mathbf{q}_0^2+\mathbf{q}_1^2\|_\infty)/\sqrt{2}-(M_1R_1+M_2R_2)/4>0,$$
$$\|\mathbf{Q}_*\|_{H\times H}=O(\alpha^{3/2}),$$

which, by Lemma 1 implies

$$\|\mathbf{q}^i-\mathbf{q}_0^i-\mathbf{q}_1^i\|=\|\mathbf{q}_*^i\|\leqq\|\mathbf{q}_*^i\|_H/\sqrt{2}\leqq\|\mathbf{Q}_*\|_{H\times H}/\sqrt{2}=O(\alpha^{3/2}).$$

This completes the proof of the theorem.

COROLLARY.

$$\|u^i-u_0^i\|=O(\alpha^{1/2}),\qquad \|v^i-v_0^i\|,\|w^i-w_0^i\|=O(\alpha^{3/2}).$$

Proof. Since

$$\|u^i-u_0-u_1^i\|,\|v^i-v_0^i-v_1^i\|,\|w^i-w_0^i-w_1^i\|\leqq\|q^i-q_0^i-q_1^i\|=O(\alpha^{3/2}),$$

we have

$$\|u^i-u_0^i\|\leqq\|u^i-u_0^i-u_1^i\|+\|u_1^i\|=O(\alpha^{1/2}),$$
$$\|v^i-v_0^i\|\leqq\|v^i-v_0^i-v_1^i\|+\|v_1^i\|=O(\alpha^{3/2}),$$
$$\|w-w_0^i\|\leqq\|w^i-w_0^i-w_1^i\|+\|w_1^i\|=O(\alpha^{3/2}),$$

by Lemma 2.

Remarks. (1) Under the same condition as in the theorem, error estimates for higher order approximations can also be obtained, if the boundary condition is sufficiently smooth.

(2) If $R_1, R_2 = O(\alpha)$ and α is sufficiently small, then the condition on R_1, R_2 is automatically satisfied and better estimates can be obtained. An asymptotic scheme may be carried out by assuming

$$u^i=u_0^i+\alpha^2u_1^i+\alpha^4u_2^i+\cdots,$$
$$v^i=\alpha v_0^i+\alpha^3v_1^i+\alpha^5v_2^i+\cdots,$$
$$w^i=\alpha w_0^i+\alpha^3w_1^i+\alpha^5w_2^i+\cdots,$$
$$p=\alpha^{-2}p_0+p_1+\alpha^2p_2+\cdots.$$

5. Existence and uniqueness of a generalized solution. In the following, it is shown that there exists a unique solution $\mathbf{Q}_*$ for (54) in $H\times H$. To this end, we shall reduce (54) to an operator equation on $H\times H$. Let $\mathbf{a}^i=\mathbf{q}_0^i+\mathbf{q}_1^i$; then (54) becomes

$$R_1^{-1}(\mathbf{q}_*^i,\boldsymbol{\phi})_H=\int_\Omega(\mathbf{q}_{*x}^i-\mathbf{a}^i\cdot\nabla\mathbf{q}_*^i-\mathbf{q}_*^i\cdot\nabla\mathbf{q}_*^i-\mathbf{q}_*^i\cdot\nabla\mathbf{a}^i$$
$$-M_i\mathbf{q}_*^i)\cdot\boldsymbol{\phi}\,d\Omega+\int_\Omega\mathbf{G}^i\cdot\boldsymbol{\phi}\,d\Omega+\int_\Omega M_i\mathbf{q}_*^i\cdot\boldsymbol{\phi}\,d\Omega. \tag{59}$$

By integration by parts, Schwarz's inequality and Lemma 1, we obtain

$$\left|\int_\Omega (\mathbf{q}^i_{*x} - \mathbf{a}^i \nabla \mathbf{q}^i_* - \mathbf{q}^i_* \cdot \nabla \mathbf{q}^i_* - \mathbf{q}^i_* \cdot \nabla \mathbf{a}^i - M_i \mathbf{q}^i_*) \cdot \boldsymbol{\phi}\, d\Omega\right|$$

$$< (\|\mathbf{q}^i_*\| + 2\|\mathbf{a}^i\|\,\|\mathbf{q}^i_*\| + \|\mathbf{q}^i_*\|^2 + M_i \|\mathbf{q}^i_*\|)\|\boldsymbol{\phi}\|_H,$$

$$\left|\int_\Omega \mathbf{G}^i \cdot \boldsymbol{\phi}\, d\Omega\right| \leqq \|\mathbf{G}^i\|\,\|\boldsymbol{\phi}\|_H/\sqrt{2},$$

$$\left|\int_\Omega \mathbf{q}^j_* \cdot \boldsymbol{\phi}\, d\Omega\right| \leqq \|\mathbf{q}^j_*\|\,\|\boldsymbol{\phi}\|_H/\sqrt{2}.$$

By Reisz's representation theorem, there exist operators A^i, B^j on H and a fixed element $\mathbf{f}^i \in H$, such that

$$\int_\Omega (\mathbf{q}^i_{*x} - \mathbf{a}^i \cdot \nabla \mathbf{q}^i_* - \mathbf{q}^i_* \cdot \nabla \mathbf{q}^i_* - \mathbf{q}^i_* \cdot \nabla \mathbf{a}^i - M_i \mathbf{q}^i_*) \cdot \boldsymbol{\phi}\, d\Omega = (A^i \mathbf{q}^i_*, \boldsymbol{\phi})_H,$$

$$\int_\Omega \mathbf{G}^i \cdot \boldsymbol{\phi}\, d\Omega = (\mathbf{f}^i, \boldsymbol{\phi})_H, \qquad \int_\Omega \mathbf{q}^j_* \cdot \boldsymbol{\phi}\, d\Omega = (B^j \mathbf{q}^j_*, \boldsymbol{\phi})_H.$$

It follows from (59) that

$$(R_i^{-1} \mathbf{q}^i_*, \boldsymbol{\phi})_H = (A^i \mathbf{q}^i_* + B^j \mathbf{q}^j_* + \mathbf{f}^i, \boldsymbol{\phi})_H,$$

for any $\boldsymbol{\phi} \in H$. Hence,

$$\mathbf{q}^i_* = R_i(A^i \mathbf{q}^i_* + B^j \mathbf{q}^j_* + \mathbf{f}^i). \tag{60}$$

Let

$$M = \begin{bmatrix} R_1 A^1, & R_1 B^2 \\ R_2 A^2, & R_2 B^1 \end{bmatrix}, \quad F = \begin{bmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \end{bmatrix}, \quad \mathbf{Q}_* = \begin{bmatrix} \mathbf{q}^1_* \\ \mathbf{q}^2_* \end{bmatrix},$$

and (60) may be expressed as an equation on $H \times H$,

$$\boldsymbol{Q}_* = M\mathbf{Q}_* + \mathbf{F}. \tag{61}$$

First we show that M is completely continuous on $H \times H$. The proof is essentially the same as given in [6]; some of the derivations are omitted. We consider a weakly convergent sequence $\{\mathbf{Q}_n\}$ in $H \times H$. Let $\mathbf{Q}_n = (\mathbf{q}^1_n, \mathbf{q}^2_n)$. Then $\{\mathbf{q}^1_n\}$, $\{\mathbf{q}^2_n\}$ are weakly convergent in H, and strongly convergent in $L_4(\Omega)$ [6]. Furthermore, for any element $\mathbf{q} \in L_2(\Omega)$, $\|\mathbf{q}\| \leqq C_1\|\mathbf{q}\|_{L4}$ where C_1 is a constant. By integration by parts, Schwarz's inequality and

$$\left|\int_\Omega (\mathbf{u} \cdot \nabla)\mathbf{v} \cdot \mathbf{w}\, d\Omega\right| \leqq C_2 \|\mathbf{u}\|_{L4}\|\mathbf{v}\|_{L4}\|\mathbf{w}\|_H,$$

where C_2 is a constant, it is shown that, for any $\boldsymbol{\Phi} \in H \times H$,

$$|(M\mathbf{Q}_m - M\mathbf{Q}_n, \boldsymbol{\Phi})_{H\times H}| < C_3(\|\mathbf{q}^1_m - \mathbf{q}^1_n\|_{L4} + \|\mathbf{q}^2_m - \mathbf{q}^2_n\|_4)\|\boldsymbol{\Phi}\|_{H\times H},$$

where C_3 is a constant. It follows that $\{M\mathbf{Q}_n\}$ is strongly convergent in $H \times H$ and M is completely continuous. Next we show that all solutions of

$$\mathbf{Q}_* - \lambda(M\mathbf{Q}_* + \mathbf{F}) = 0, \tag{62}$$

are uniformly bounded for $\lambda \in [0, 1]$ if R_1, R_2 satisfy the condition in Theorem 1. However, this follows from the inequality in (58) to estimate $\|\mathbf{Q}_*\|^2_{H\times H}$. We take the

scalar product of (62) with $\mathbf{Q}_*$, and follow the same derivations as in Theorem 1 to obtain

$$\{1-\lambda[(R_1+R_2)/2+(R_1\|\mathbf{q}_0^1+\mathbf{q}_1^1\|_\infty+R_2\|\mathbf{q}_0^2+\mathbf{q}_1^2\|_\infty)/\sqrt{2}+(M_1R_1+M_2R_2)/4]\}\|\mathbf{Q}_*\|_{H\times H}^2$$

$$\leqq (R_1+R_2)C^2\alpha^3/2. \tag{63}$$

Therefore, $\mathbf{Q}_*$ is uniformly bounded in $H\times H$ if the same condition on R_1, R_2 as in Theorem 1 is satisfied. By the Leray–Schauder fixed point theorem [6], (61) has a solution $\mathbf{Q}_*$ in $H\times H$. Since $\mathbf{q}^i=\mathbf{q}_0^i+\mathbf{q}_1^i+\mathbf{q}_*^i$, this also implies the existence of $\mathbf{q}^i$.

Suppose now that there are two solutions $\mathbf{q}_{(1)}^i$, $\mathbf{q}_{(2)}^i$ of our problem. Let

$$\mathbf{q}_*^i=\mathbf{q}_{(1)}^i-\mathbf{q}_{(2)}^i.$$

Then, $\mathbf{q}_*^i\in H$ and satisfies the same equation (54) if we replace $\mathbf{q}_0^i+\mathbf{q}_1^i$ by $\mathbf{q}_{(2)}^i$. Since $\mathbf{q}_{(2)}^i$ satisfies (7), $\mathbf{G}^i=0$, and (63) still holds for $\mathbf{Q}_*=(\mathbf{q}_*^1,\mathbf{q}_*^2)$ except that the right side is replaced by zero. Hence, under the same condition on R_1, R_2, $\mathbf{Q}_*=0$ and the generalized solution is unique. In summary, we state our results as:

THEOREM 2. *Under the same condition as in Theorem* 1, *there exists a unique generalized solution* $(\mathbf{q}^1,\mathbf{q}^2)$ *of* (4)–(6), *such that* $\mathbf{q}^i-\mathbf{q}_0^i-\mathbf{q}_1^i\in H$.

Acknowledgment. The research reported here was supported by the National Science Foundation under grant MCS 77-00097.

REFERENCES

[1] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
[2] T. K. HUNG AND T. D. BROWN, *Solid-particle motion in two-dimensional peristaltic flows*, J. Fluid Mech., 73 (1976), pp. 77–96.
[3] M. Y. JAFFRIN AND A. H. SHAPIRO, *Peristaltic pumping*, Ann. Rev. Fluid Mech., 3 (1971), pp. 13–36.
[4] D. D. JOSEPH, *Stability of Fluid Motion I*, Springer-Verlag, New York, 1976.
[5] M. R. KAIMAL, *Peristaltic pumping of a Newtonian fluid with particles suspended in it at low Reynolds number under long wavelength approximations*, Trans. ASME, 45 (1978), pp. 32–36.
[6] O. A. LADYZHENSKAYA, *The mathematical theory of viscous incompressible flow*, Gordon and Breach, New York, 1969.
[7] N. LIRON, *On peristaltic flow and its efficiency*, Bull. Math. Biol., 38 (1976), pp. 573–596.
[8] M. J. MANTON, *Long wavelength peristaltic pumping at low Reynolds number*, J. Fluid Mech., 68 (1975), pp. 467–476.
[9] T. K. MITTRA AND S. N. PRASAD, *Interaction of peristaltic motion with Poiseuille flow*, Bull. Math. Biol., 36 (1974), pp. 127–144.
[10] J. D. MURRAY, *On the mathematics of fluidization I*, J. Fluid Mech., 21 (1965), pp. 465–493.
[11] S. I. PAI, *Two-Phase Flows*, Vieweg, Braunschweig, West Germany, 1977.
[12] P. G. SAFFMAN, *On the stability of laminar flow of a dusty gas*, J. Fluid Mech., 13 (1962), pp. 120–128.
[13] S. M. SHIH AND M. C. SHEN, *Uniform asymptotic approximation for viscous fluid flows down an inclined plane*, this Journal, 6 (1975), pp. 560–581.
[14] S. L. SOO, *Fluid Dynamics of Multiphase Systems*, Blaisdell, Waltham MA, 1967.
[15] M. C. SHEN, S. M. SHIH AND A. M. WU, *Asymptotic method for peristaltic transport*, Bull. Math. Biol., 42 (1980), pp. 305–325.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0006 $01.00/0

STABILITY CONDITIONS FOR LINEAR HAMILTONIAN SYSTEMS WITH PERIODIC COEFFICIENTS*

EARL R. BARNES†

Abstract. We begin with a system of k uncoupled harmonic oscillators $\dot{x}_i = -\omega_i x_{k+i}$, $\dot{x}_{k+i} = \omega_i x_i$, $i = 1, \cdots, k$. We then couple the oscillators according to the equations $\dot{x}_i = -\omega_i x_{k+i} - \sum_{j=1}^{2k} h_{k+i,j}(t)x_j$, $\dot{x}_{k+i} = \omega_i x_i + \sum_{j=1}^{2k} h_{ij}(t)x_j$, where $H(t) = (h_{ij}(t))$ is a $2k \times 2k$ symmetric matrix, periodic of period $T > 0$. The uncoupled system is clearly stable. We show that the coupled system is also stable if the nonresonant condition $\omega_\nu + \omega_\mu \neq 2n\pi/T$, $\nu, \mu = 1, \cdots, k$, $n = 0, \pm 1, \pm 2, \cdots$ is satisfied, and if the interaction between each pair of oscillators ν, μ, and all other oscillators in the system, is so weak that $\sum_{i=\nu,\mu} \int_0^T \{\sum_{j=1}^{2k} |h_{ij}| + \sum_{j\neq i} |h_{k+i,j}|\}\, dt < T \min_n |\omega_\nu + \omega_\mu - 2n\pi/T|$. This condition is sharp. For certain ω_i's, the theorem becomes false if the inequality $<$ is replaced by $\leqq$. For other ω_i's, the theorem can be weakened slightly, but not significantly.

1. Introduction.

Let J denote the nonsingular skew-symmetric matrix

$$J = \begin{pmatrix} 0 & I_k \\ -I_k & 0 \end{pmatrix},$$

and let $H(t)$ be a real symmetric $2k \times 2k$ periodic matrix function of period $T > 0$. That is, $H(t+T) = H(t) = H^T(t)$. I_k denotes the $k \times k$ identity matrix. We shall give conditions on H which will guarantee stability of the trivial solution of the differential equation

$$J\dot{x} = H(t)x, \qquad -\infty < t < \infty. \tag{1.1}$$

Let $h_{ij}(t)$ denote the ijth element of $H(t)$ and introduce the Hamiltonian

$$\mathcal{H} = -\frac{1}{2} \sum_{i,j=1}^{2k} h_{ij}(t)x_i x_j.$$

We can then write (1.1) as

$$\dot{x}_i = \frac{\partial \mathcal{H}}{\partial x_{k+i}} = -\sum_{j=1}^{2k} h_{k+i,j}x_j, \qquad i = 1, \cdots, k,$$

$$\dot{x}_{k+i} = -\frac{\partial \mathcal{H}}{\partial x_i} = \sum_{j=1}^{2k} h_{ij}x_j, \qquad i = 1, \cdots, k.$$

Thus, (1.1) is a Hamiltonian system. For brevity, we shall call the matrix function $H(t)$ a Hamiltonian. We restrict our attention to the space M of Hamiltonians whose entries $h_{ij}(t)$ are Lebesgue integrable over the interval $[0, T]$. We define a norm on M by

$$\|H\| = \int_0^T \sum_{1 \leqq i \leqq j \leqq 2k} |h_{ij}|\, dt. \tag{1.2}$$

For second-order systems, our stability condition can be stated in terms of this norm.

DEFINITION 1.1. The system (1.1) is said to be *stable* if all its solutions are bounded on the entire line $\infty < t < \infty$.

DEFINITION 1.2. The system (1.1) is said to be *strongly stable* if there exists an $\varepsilon > 0$ such that the equation $J\dot{x} = Lx$ is stable for all Hamiltonians $L \in M$ satisfying $\|L - H\| < \varepsilon$. That is, (1.1) is strongly stable if it is stable and remains stable under small perturbations of H within the space M.

* Received by the editors September 27, 1979, and in revised form June 20, 1980.
† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10958.

In some cases, we say that the Hamiltonian H is stable, or strongly stable, to mean that the system (1.1) is stable, or strongly stable, respectively. The set of strongly stable Hamiltonians will be denoted by O. O is open by definition. Moreover, it is known, cf. [1], [4] or [2, Chap. 3], that O is the union of a countable collection of simply connected domains $O_n^{(\sigma)}$, $n = 0, \pm 1, \pm 2, \cdots$. The symbol (σ) denotes a signature. It is a sequence of length k, made up of the symbols + and −. Thus, for each index n, O has 2^k simply connected components $O_n^{(\sigma)}$.

In studying the stability of the system (1.1) we assume that $H(t)$ can be written as $H(t) = C + L(t)$, where $C, L \in M$ and C is a constant strongly stable Hamiltonian. When this is the case we can assume without loss of generality that C is a diagonal matrix of the form diag $(\omega_1, \cdots, \omega_k, \omega_1, \cdots, \omega_k)$ where the ω_i's, satisfy

$$\omega_\nu + \omega_\mu \neq \frac{2n\pi}{T}, \qquad \nu, \mu = 1, \cdots, k, \quad n = 0, \pm 1, \pm 2, \cdots. \tag{1.3}$$

This is shown in [2, p. 263]. Thus, without loss of generality, we shall replace (1.1) by

$$J\dot{x} = (\Omega + H(t))x, \tag{1.4}$$

where

$$\Omega = \text{diag}\,(\omega_1, \cdots, \omega_k, \omega_1, \cdots, \omega_k).$$

This equation is to be thought of as a perturbation of the equation

$$J\dot{x} = \Omega x, \tag{1.5}$$

describing the motion of k uncoupled harmonic oscillators $\ddot{x}_\nu + \omega_\nu^2 x_\nu = 0$, $\nu = 1, \cdots, k$. When the condition

$$\omega_\nu + \omega_\nu = \frac{2n\pi}{T}$$

is satisfied for two oscillators ν and μ, and for some integer n, we say the oscillators ν and μ are in resonance.

If the nonresonant condition (1.3) is satisfied, the system (1.5) is strongly stable. By definition, the system (1.4) is also strongly stable for $\|H\|$ sufficiently small. The purpose of this paper is to determine exactly how small H must be to guarantee that (1.4) is strongly stable, given that (1.5) is strongly stable. The answer is provided by the following theorem.

THEOREM 1.1. *Let* $\Omega = \text{diag}\,(\omega_1, \cdots, \omega_k, \omega_1, \cdots, \omega_k)$ *and let* $H \in M$ *be a Hamiltonian of period* $T > 0$. *If* (1.3) *is satisfied, the differential equation* (1.4) *is strongly stable if for every pair* $\nu, \mu \in \{1, \cdots, k\}$

$$\sum_{i=\nu,\mu} \int_0^T \left\{ \sum_{j=1}^{2k} |h_{ij}| + \sum_{j \neq i} |h_{k+i,j}| \right\} dt < T \min \left| \omega_\nu + \omega_\mu - \frac{2n\pi}{T} \right|, \tag{1.6}$$

where the minimum on the right is taken over $n = 0, \pm 1, \pm 2, \cdots$.

Note that the coefficients $h_{ij}, h_{k+i,j}$, $i = \nu, \mu$, $j = 1, \cdots, 2k$ appearing in the left side of (1.6) are precisely the coefficients which couple the oscillators ν and μ with all other oscillators in the system (1.5). The theorem gives a measure of how strong these couplings can be, for each pair of oscillators ν, μ, without destroying the strong stability of the unperturbed system.

2. Preliminaries. In order to prove Theorem 1.1 we need a few facts concerning the multipliers of the Hamiltonian system (1.1). Let $X(t)$ denote the fundamental matrix for the system (1.1). Let R_{2k} denote complex $2k$-space, with the inner product of two vectors $x, y \in R_{2k}$ defined by $(x, y) = y^*x$. The symbol $*$ denotes complex conjugate transpose. Introduce a new inner product $[\cdot, \cdot]$ in R_{2k} by the definition

$$[x, y] = i(Jx, y) = iy^*Jx.$$

It is easy to see that $[x, x]$ is a real number for each $x \in R_{2k}$. Unlike the inner product (x, x) the inner product $[x, x]$ may be negative for certain x's.

DEFINITION 2.1.

(a) Let $\rho(|\rho| = 1)$ be an r-fold eigenvalue of $X(T)$ lying on the unit circle. Let $G\rho$ denote the subspace of R_{2k} spanned by the eigenvectors corresponding to ρ. If $[x, x] > 0$ for each $x \neq 0$ in $G\rho$, we say that ρ is an r-fold eigenvalue of the *first kind*, and if $[x, x] < 0$ for each $x \neq 0$ in $G\rho$ we say ρ is an r-fold eigenvalue of the *second kind.*

(b) Let $\rho(|\rho| = 1)$ be an r-fold eigenvalue of $X(T)$ on the unit circle such that $[x, x]$ is not of fixed sign on the eigenspace $G\rho$. Then we say that ρ is an r-fold eigenvalue of *mixed kind.*

(c) Let $\rho(|\rho| \neq 1)$ be an r-fold eigenvalue of $X(T)$. Then ρ is called an r-fold eigenvalue of the *first kind* if $|\rho| < 1$ and of the *second kind* if $|\rho| > 1$.

The eigenvalues of $X(T)$ are called multipliers of the system (1.1). Multipliers of the first and second kind are said to be definite. It turns out that definite multipliers depend continuously on H (cf. [2, p. 191]). Thus, under small perturbations of H, a multiplier of the first kind will not suddenly change to one of the second kind, and vice-versa. When a multiplier of the first kind coincides with one of the second kind, a multiplier of mixed (indefinite) kind is formed.

The following theorem from [2, p. 196] expresses stability conditions for (1.1) in terms of its multipliers.

THEOREM 2.1. *The Hamiltonian H is strongly stable if and only if all the multipliers of* (1.1) *lie on the unit circle and are definite.*

It is shown in [2, p. 258] that the multipliers of (1.5) are

$$\rho_\nu = \rho^{i\omega_\nu T} \quad \text{and} \quad \bar{\rho}_\nu = e^{-i\omega_\nu T}, \qquad \nu = 1, \cdots, k,$$

and that ρ_ν is a multiplier of the first kind and $\bar{\rho}_\nu$ a multiplier of the second kind. Condition (1.3) states that no two multipliers $\rho_\nu = e^{i\omega_\nu T}$ and $\bar{\rho}_\mu = e^{-i\omega_\mu T}$ of different kinds coincide.

3. Second-order systems. In this section, we prove Theorem 1.1 for second-order systems. For this class of systems the structure of the stability domains $O_n^{(\sigma)}$ is much simpler than in the general case. These domains have been described in fine detail in [2, Chap. 8]. In particular, a three-dimensional model of the space M_2 of 2×2 Lebesgue integrable Hamiltonians is given in [2, Chap. 8]. This model provides a useful geometric interpretation of Theorem 1.1 for second-order systems. We begin with a description of the model.

In order to avoid the use of subscripts, let us agree to denote a typical Hamiltonian in M_2 by

$$H(t) = \begin{pmatrix} \alpha(t) & \beta(t) \\ \beta(t) & \gamma(t) \end{pmatrix}. \tag{3.1}$$

The differential equation (1.1) then becomes

$$\begin{aligned} \dot{x}_1 &= -\beta(t)x_1 - \gamma(t)x_2, \\ \dot{x}_2 &= \alpha(t)x_1 + \beta(t)x_2. \end{aligned} \tag{3.2}$$

The three-dimensional model of M_2 will be denoted by R^3. It consists of open connected sets O_n, H_n, $n = 0, \pm 1, \pm 2, \cdots$, and closed connected sets π_n^{*-}, π_n^{*+}, π_n^{**}, $n = 0, \pm 1, \pm 2, \cdots$, and may be obtained by rotating Fig. 3.1 about a line through the points π_n^{**}, $n = 0, \pm 1, \pm 2, \cdots$. The sets O_n are domains of strong stability for 2×2 Hamiltonians. Here we have labeled these sets differently than they were labeled in [2, Chap. 3]. The sets here labeled O_{2n} would be labeled $O_n^{(+)}$ by the labeling convention of [2, Chap. 3], and the ones here labeled O_{2n+1} would be labeled $O_n^{(-)}$. The sets H_n are domains of instability. Each Hamiltonian in one of these sets is unstable.

If $H \in \pi_n^{*-} \cup \pi_n^{*+}$ for some n, then (3.2) has exactly one linearly independent periodic solution ($x(t+T) = x(t)$) of period T if n is even, and exactly one linearly independent antiperiodic solution ($x(t+T) = -x(t)$) of period T if n is odd. If $H \in \pi_n^{**}$, then all solutions of (3.2) are periodic of period T if n is even, and all solutions are antiperiodic of period T if n is odd.

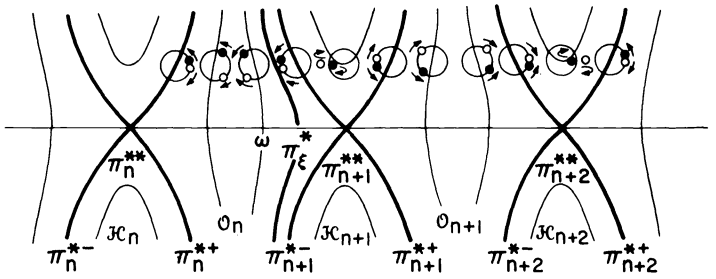


FIG. 3.1. R^3.

Consider the curved vertical lines in Fig. 3.1. The Hamiltonians on each of these lines have the same multipliers. The positions of these multipliers on the unit circle are indicated for each line in Fig. 3.1. A multiplier of the first kind is indicated by a black dot and one of the second kind is indicated by a white dot. The arrows indicate the motion of the multipliers as one moves through R^3 from left to right. The sets π_n^{*-}, π_n^{*+} contain the Hamiltonians having multipliers $e^{\pm in\pi}$, $n = 0, \pm 1, \pm 2, \cdots$. In particular, the set π_n^{**} contains the Hamiltonian $(n\pi/T)I_2$ where I_2 denotes the 2×2 identity matrix. Moreover, if ω satisfies

$$\frac{n\pi}{T} < \omega < \frac{(n+1)\pi}{T}, \tag{3.3}$$

then ωI_2 lies on the line segment connecting π_n^{**} and π_{n+1}^{**}. Thus, the Hamiltonian ωI_2 can be associated with the point ω measuring Euclidean distance along the horizontal line in Fig. 3.1. When this is done, the distance between a Hamiltonian $\omega I_2 \in O_n$ and a vertical line such as π_ξ^* in Fig. 3.1 can be obtained by measuring the Euclidean distance between ω and π_ξ^* along the ω axis. This result is the content of the following theorem.

THEOREM 3.1. *If ω satisfies* (3.3), *then $\omega I_2 \in O_n$, and if H is any Hamiltonian in the left boundary π_n^{*+} of O_n, then*

$$\|H - \omega I_2\| \geqq T\left(\omega - \frac{n\pi}{T}\right) = \omega T - n\pi. \tag{3.4}$$

Similarly, if H is any Hamiltonian in the right boundary π^{*-}_{n+1} of O_n, then

$$\|H-\omega I_2\| \geqq T\left[\frac{(n+1)\pi}{T}-\omega\right] = (n+1)\pi-\omega T. \tag{3.5}$$

More generally, if H is any Hamiltonian on the curved vertical line π^*_ξ containing Hamiltonians with multipliers $e^{\pm i\xi T}$, $n\pi/T<\xi<(n+1)\pi/T$, then

$$\|H-\omega I_2\| \geqq |\xi T-\omega T|. \tag{3.6}$$

Proof. The fact that $\omega I_2 \in O_n$ is shown in [2, p. 659]. We shall prove (3.4) and (3.5) for $n \geqq 1$. The proof for $n<1$ is similar.

Let ρ^+_n denote the distance from ωI_2 to the left boundary π^{*+}_n of O_n. Let ρ^-_{n+1} denote the distance to the right boundary π^-_{n+1}. For $n \geqq 1$, it is shown in [6] that

$$\rho^+_n = \omega T - n\pi, \tag{3.7}$$

and

$$\rho^-_{n+1} = 2(n+1)\log\left[\frac{1+\cos\dfrac{\omega T}{4(n+1)}}{1+\sin\dfrac{\omega T}{4(n+1)}}\right]+4(n+1)\left[\frac{\cos\dfrac{\omega T}{4(n+1)}-\sin\omega T}{1+\cos\dfrac{\omega T}{4(n+1)}+\sin\omega T}\right]. \tag{3.8}$$

It is also shown in [6, Lemma 6.1] that for $\omega = n\pi/T$, $\rho^-_{n+1}>\pi$ and $\lim_{n\to\infty}\rho^-_{n+1}=\pi$. The same technique of proof can be used to show that for any ω satisfying (3.3), $\rho^-_{n+1}>(n+1)\pi-\omega T$ and $\rho^-_{n+1}=(n+1)\pi-\omega T+O(1/(n+1)^2)$; cf. [7, p. 20]. Thus, if $H\in\pi^{*+}_n$ we have

$$\|H-\omega I_2\| \geqq \rho^+_n = \omega T - n\pi,$$

and if $H\in\pi^{*-}_{n+1}$, we have

$$\|H-\omega I_2\| \geqq \rho^-_{n+1} > (n+1)\pi-\omega T.$$

This proves (3.4) and (3.5) for $n\geqq 1$.

Now consider (3.6). First assume that $\xi T=(n+p/q)\pi$ where p and q are integers satisfying $0<p<q$. If $H\in\pi^*_\xi$ the differential equation (3.2) has a solution $x(t)$ satisfying

$$x(t+T)=e^{i\xi T}x(t).$$

It follows that

$$x(t+qT)=(-1)^{nq+p}x(t).$$

Thus, (3.2) has a periodic solution of period $T_1=qT$ if $nq+p$ is even, and an antiperiodic solution of period T_1 if $nq+p$ is odd.

Let ψ_x denote the angle through which the vector $x(t)$ rotates as t goes from 0 to T_1. Since $x(T)=e^{i\xi T}x(0)$, $x(2T)=e^{i2\xi T}x(0)$, etc., we have

$$\psi_x=(nq+p)\pi.$$

Now consider the infimization problem,

$$\inf \int_0^{T_1} \{|\alpha(t)-\omega|+|\beta(t)|+|\gamma(t)-\omega|\}\,dt, \tag{3.9}$$

where the infimum is taken over the class of Hamiltonians (3.1) for which (3.2) has a solution satisfying $\psi_x = (nq+p)\pi$. This class includes, in particular, the set π_ξ^*. Problem (3.9) is precisely the type of problem solved in [6, § 4] in determining the distances from ωI_2 to the boundaries of O_n. If $\xi T > \omega T$, problem (3.9) corresponds to the problem of finding the distance from ωI_2 to the right boundary of O_n. If $\xi T < \omega T$, (3.9) corresponds to finding the distance from ωI_2 to the left boundary of O_n. In either case, the value of (3.9) can be obtained by replacing T by T_1 and n or $n+1$ by $nq+p$ in the appropriate formula for ρ_n^+ or ρ_{n+1}^-. Let δ denote the infimum in (3.9). Then, if $n \geqq 1$ and $\omega T > \xi T$, we obtain, by substituting in (3.7)

$$\delta = 2q\left(n+\frac{p}{q}\right)\log\left[\frac{1+\cos\dfrac{\omega qT}{4(nq+p)}}{1+\sin\dfrac{\omega qT}{4(nq+p)}}\right]$$

$$+4q\left(n+\frac{p}{q}\right)\left[\frac{\cos\dfrac{\omega qT}{4(nq+p)}-\sin\dfrac{\omega qT}{4(nq+p)}}{1+\cos\dfrac{\omega qT}{4(nq+p)}+\sin\dfrac{\omega qTY}{4(nq+p)}}\right]$$

$$\geqq q\left[\left(n+\frac{p}{q}\right)\pi-\omega T\right] = q(\xi T-\omega T).$$

Similarly, if $n<1$ it can be shown that $\delta \geqq q|\omega T-\xi T|$. See [7] for details. Now if $H \in \pi_\xi^*$ and is denoted by (3.1), we have

$$q\|H-\omega I_2\| = \int_0^{T_1} \{|\alpha(t)-\omega|+|\beta(t)|+|\gamma(t)-\omega|\}\,dt$$

$$\geqq \delta \geqq q|\omega T-\xi T|.$$

This proves (3.6) when ξT is a rational multiple of π. The proof for general ξT follows by an easy continuity argument.

COROLLARY. *Theorem* 1.1 *holds for* $k=1$.

Proof. When $k=1$, we must take $\nu=\mu$ and $\omega_\nu = \omega_\mu = \omega$ in (1.3) and (1.6). Then, if H is given by (3.1), condition (1.6) becomes

$$\|H\| = \int_0^T \{|\alpha(t)|+|\beta(t)|+|\gamma(t)|\}\,dt < T \min_m \left|\omega-\frac{m\pi}{T}\right|.$$

If ω satisfies (3.3), this can be written as

$$\|H\| < T \min_m \left|\omega-\frac{m\pi}{T}\right| = T \min\left\{\omega-\frac{n\pi}{T}, \frac{(n+1)\pi}{T}-\omega\right\} \leqq \min\{\rho_n^+, \rho_{n+1}^-\}.$$

This says that the Hamiltonian $\omega I_2 + H$ lies in the interior of a sphere about ωI_2 of radius $\min\{\rho_n^+, \rho_{n+1}^-\}$. This implies that $\omega I_2 + H \in O_n$ and is consequently strongly stable.

Remark. If $n \geqq 1$ and if $n\pi/T < \omega < (n+\frac{1}{2})\pi/T$, then the distance from ωI_2 to the boundary of O_n is $\rho_n^+ = \omega T - n\pi$. Moreover, it is shown in [6] that there is a Hamiltonian

$H \in \pi_n^{*+}$, such that $\|H - \omega I_2\| = \omega T - n\pi$. The Hamiltonian H is not stable. Thus, for certain ω's the inequality in Theorem 1.1 is sharp. On the other hand, if the difference $(n+1)\pi/T - \omega$ is sufficiently small and positive, the distance from ωI_2 to the boundary of O_n is $\rho_{n+1}^- > (n+1)\pi - \omega T = T \min_m |\omega - m\pi/T|$. In this case, the inequality in Theorem 1.1 can be improved. But the improvement will be very small for large values of n, since $\rho_{n+1}^- = (n+1)\pi - \omega T + O(1/(n+1)^2)$.

A similar remark holds for $n < 1$. Also, it will be seen from the proof of Theorem 1.1 that a similar remark holds for inequality (1.6) for $k > 1$.

4. Proof of Theorem 1.1. For the proof of Theorem 1.1 for $k > 1$, we need one more theorem from [2, Chap. 3]; see also [4]. It has to do with the concept of directional wideness.

Let K_1, K_2 be any two Hamiltonians. We shall write $K_1 \leqq K_2$ if $(K_1(t)x, x) \leqq (K_2(t)x, x)$ for any $2k$-vector x, and for $0 \leqq t \leqq T$.

DEFINITION. A set $\mathcal{M} \subset M$ said to be *directionally wide* if, for any $K_1, K_2 \in \mathcal{M}$ such that $K_1 \leqq K_2$ and

$$K_1 + \eta[K_2 - K_1] \in \mathcal{M},$$

for $0 \leqq \eta \leqq 1$, we have $H \in \mathcal{M}$ for any Hamiltonian H satisfying the inequalities

$$K_1 \leqq H \leqq K_2. \tag{4.1}$$

Thus, $\mathcal{M}$ is a *directionally wide* set if the fact that the entire line segment with endpoints K_1 and K_2 belongs to $\mathcal{M}$ implies the Hamiltonians satisfying (4.1) belong to $\mathcal{M}$.

The following theorem appears in [2, p. 239] and in [4].

THEOREM 4.1. *The stability domains $O_n^{(\sigma)}$ are directionally wide.*

Proof of Theorem 1.1. As a first step to proving Theorem 1.1 for $k > 1$, we consider Hamiltonians of the form

$$H(t) = \begin{bmatrix}
h_{11}(t) & 0 & \cdots & 0 & h_{1,k+1}(t) & 0 & \cdots & 0 \\
0 & h_{22}(t) & 0 \cdots & 0 & 0 & h_{2,k+2}(t) & \cdots & 0 \\
\cdot & & \cdot & \cdot & \cdot & & \cdot & \\
\cdot & & & \cdot & \cdot & & & \cdot \\
\cdot & & & \cdot & 0 & & & 0 \\
0 & & \cdots & 0 & h_{kk}(t) & 0 & \cdots & h_{k,2k}(t) \\
h_{k+1,1}(t) & 0 & \cdots & 0 & h_{k+1,k+1}(t) & 0 & \cdots & 0 \\
0 & h_{k+1,1}(t) & 0 \cdots & 0 & 0 & h_{k+1,k+1}(t) & \cdots & 0 \\
\cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot \\
\cdot & & & \cdot & \cdot & & & \cdot \\
\cdot & & & \cdot & 0 & & & 0 \\
0 & \cdots & & 0 & h_{2k,k}(t) & 0 & \cdots & h_{2k,2k}(t)
\end{bmatrix}. \tag{4.2}$$

We denote this class of Hamiltonians by M'. The Hamiltonian $\Omega = \operatorname{diag}(\omega_1, \cdots, \omega_k, \omega_1, \cdots, \omega_k)$ defined in Theorem 1.1 lies in M', and since (1.3) is satisfied, it lies in some stability domain $O_n^{(\sigma)}$. We must show that condition (1.6) guarantees that $\Omega + H \in O_n^{(\sigma)}$.

For $H \in M'$ the differential equation $J\dot{x} = (\Omega + H(t))x$ separates into the k second-order systems

$$\dot{x}_\nu = -h_{k+\nu,\nu}x_\nu - (\omega_\nu + h_{k+\nu,k+\nu})x_{k+\nu},$$

$$\dot{x}_{k+\nu} = (\omega_\nu + h_{\nu\nu})x_\nu + h_{\nu,k+\nu}x_{k+\nu},$$

of the form

$$J_2 y_\nu = (\Omega_2 + H_2(t))y_\nu, \qquad \nu = 1, 2, \cdots, k, \tag{4.3}$$

which we treat as perturbations of the systems $J_2\dot{y}_\nu = \Omega_2 y_\nu$. J_2, Ω_2 and H_2 are 2×2 matrices.

Recall that the multipliers of the unperturbed system $J\dot{x} = \Omega x$ are

$$\rho_\nu = e^{i\omega_\nu T} \quad \text{and} \quad \bar{\rho}_\nu = e^{-i\omega_\nu T}, \qquad \nu = 1, \cdots, k,$$

and that ρ_ν is a multiplier of the first kind while $\bar{\rho}_\nu$ is a multiplier of the second kind. Represent the multipliers of the first kind by a black dot on the unit circle and those of the second kind by a white dot. The nonresonant condition $\omega_\nu + \omega_\mu \neq 2n\pi/T$, $\nu, \mu = 1, \cdots, k$, $n = 0, \pm 1, \pm 1, \cdots$, says that there are no multipliers of mixed kind. Assume now that the system $J\dot{x} = \Omega x$ is perturbed to the system

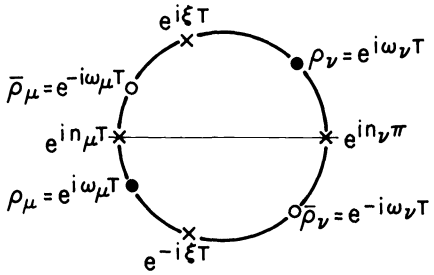


FIG. 4.1.

$$J\dot{x} = (\Omega + H(t))x, \tag{4.4}$$

within the class M'. Assume further that $\Omega + H(t)$ lies in the boundary of $O_n^{(\sigma)}$ and that (1.6) is satisfied. This will lead to a contradiction. Assume that during the perturbation none of the intermediate Hamiltonians is on the boundary of $O_n^{(\sigma)}$. It is always possible to arrange for this to be true since $O_n^{(\sigma)}$ is open and connected. During the perturbation of Ω to $\Omega + H$, some multiplier ρ_ν of the first kind and some multiplier $\bar{\rho}_\mu$ of the second kind move along the unit circle until they collide at some point $e^{i\xi T}$, on the unit circle. Assume for the moment that $e^{i\xi T} \neq \pm 1$. Then $\nu \neq \mu$. During the perturbation of Ω to $\Omega + H$, the multipliers ρ_ν and $\bar{\rho}_\mu$ remain on the same upper or lower half of the unit circle as $e^{i\xi T}$. That is, during the perturbation no one of the subsystems (4.3) is allowed to go unstable at an intermediate point.

Let H_ν and H_μ denote the Hamiltonians defined by

$$H_j = \begin{pmatrix} \omega_j + h_{jj} & h_{k+j,j} \\ h_{k+j,j} & \omega_j + h_{k+j,k+j} \end{pmatrix}, \qquad j = \nu, \mu.$$

Since (1.3) holds, there exist integers n_ν and n_μ such that

$$n_\nu\pi < \omega_\nu T < (n_\nu + 1)\pi \quad \text{and} \quad n_\mu\pi < \omega_\mu T < (n_\mu + 1)\pi.$$

Since the point $e^{i\xi T}$ lies on the same upper or lower half of the unit circle as $\rho_\nu = e^{i\omega_\nu T}$ and is unchanged if we add any even multiple of π to its argument, we may assume that

$$n_\nu\pi < \xi T < (n_\nu + 1)\pi.$$

The multipliers corresponding to the Hamiltonians H_ν, H_μ are $e^{\pm i\xi T}$. It follows from (3.6) that

$$\begin{aligned}\|H_\nu - \omega_\nu I_2\| &= \int_0^T \{|h_{\nu\nu}| + |h_{k+\nu}| + |h_{k+\nu,k+\nu}|\}\, dt \\ &\geqq |\xi T - \omega_\nu T|.\end{aligned} \tag{4.5}$$

Similarly, there exists an integer m such that the point

$$e^{-i\xi T} = e^{i(2m\pi - \xi T)}$$

lies between $e^{in_\mu T}$ and $e^{i(n_\mu+1)T}$ and

$$n_\mu \pi < 2m\pi - \xi T < (n_\mu + 1)\pi.$$

Since $e^{\pm i(2m\pi - \xi T)}$ are the multipliers of H_μ it follows from (3.6) that

$$\begin{aligned}\|H_\mu - \omega_\mu I_2\| &= \int_0^T \{|h_{\mu\mu}| + |h_{k+\mu,\mu}| + |h_{k+\mu,k+\mu}|\}\, dt \\ &\geqq |(2m\pi - \xi T) - \omega_\nu T|.\end{aligned} \tag{4.6}$$

Combining (4.5) and (4.6) we obtain

$$\begin{aligned}\sum_{j=\nu,\mu} \|H_j - \omega_j I_2\| &= \sum_{j=\nu,\mu} \int_0^T \{|h_{jj}| + |h_{k+j,k}| + |h_{k+j,k+j}|\}\, dt \\ &\geqq |\xi T - \omega_\nu T| + |2m\pi - \xi T - \omega_\mu T| \\ &\geqq |2m\pi - (\omega_\nu + \omega_\mu)T| \\ &\geqq T \min_n \left|\omega_\nu + \omega_\mu - \frac{2n\pi}{T}\right|.\end{aligned}$$

This contradicts (1.6). Suppose now that $e^{i\xi T} = \pm 1$. In this case, we can take $\mu = \nu$ in the above argument. The perturbed second-order Hamiltonian H_ν then has multipliers $\rho_\nu = \bar{\rho}_\nu = \pm 1$. H_ν therefore, lies in the boundary of O_{n_ν}. It therefore follows from Theorem 3.1 that

$$\int_0^T \{|h_{\nu\nu}| + |h_{k+\nu,\nu}| + |h_{k+\nu,k+\nu}|\}\, dt \geqq T \min_n \left|\omega_\nu - \frac{n\pi}{T}\right|,$$

and again we have a contradiction of (1.6). It follows that if (1.6) holds, then the system (4.4) can have no multipliers of mixed kind. This completes the proof of Theorem 1.1 for Hamiltonians of the form (4.2).

Suppose now that H is any Hamiltonian in M for which condition (1.6) holds. We shall show that $\Omega + H \in O_n^{(\sigma)}$, thus completing the proof of Theorem 1.1.

Let H_s denote the portion of H given by (4.2) and let H_r denote the remaining portion of H. Both H_s and H_r are symmetric and we have $H = H_s + H_r$.

Let $A = (a_{ij})$ be any real symmetric matrix of order N and let x be an N-vector. We then have

$$\begin{aligned}(x, Ax) = \sum_{i,j=1}^N a_{ij}\bar{x}_i x_j &\leqq \sum_{i,j=1}^N |a_{ij}|\,|x_i|\,|x_j| \\ &\leqq \frac{1}{2}\sum_{i,j=1}^N |a_{ij}|(|x_i|^2 + |x_j|^2) = \sum_{i,j=1}^N |a_{ij}|\,|x_j|^2,\end{aligned}$$

by the symmetry of A. This says that A satisfies $A \leqq A_\Delta$, where A_Δ is the diagonal matrix

$$A_\Delta = \operatorname{diag}\Big(\sum_{j=1}^{N} |a_{1j}|, \sum_{j=1}^{N} |a_{2j}|, \cdots, \sum_{j=1}^{N} |a_{Nj}|\Big).$$

Similarly, one can show that $-A_\Delta \leqq A$.

If we apply this result to H_r, we obtain

$$-\Delta \leqq H_r \leqq \Delta,$$

where Δ is the diagonal matrix

$$\Delta = \operatorname{diag}\Big(\sum_{\substack{j=1 \\ j\neq 1,k+1}}^{2k} |h_{1j}|, \sum_{\substack{j=1 \\ j\neq 2,k+2}}^{2k} |h_{2j}|, \cdots, \sum_{\substack{j=1 \\ j\neq k,2k}}^{2k} |h_{2k,j}|\Big).$$

Let $K_1 = H_s - \Delta$ and $K_2 = H_s + \Delta$. Then, since $H = H_s + H_r$ we have

$$K_1 \leqq H \leqq K_2. \tag{4.7}$$

Moreover, K_1 and K_2 are of the form (4.2).

Let H_ν and H_μ denote the 2×2 Hamiltonians formed from the nonzero elements in rows ν, $k+\nu$, and μ, $k+\mu$ respectively, of K_2. We then have

$$\begin{aligned}\sum_{j=\nu,\mu} \|H_j\| &= \sum_{j=\nu,\mu} \int_0^T \Big\{ \Big| h_{jj} + \sum_{l\neq j,k+j} |h_{j,l}| \Big| \\ &\qquad + |h_{j,k+j}| + \Big| h_{k+j,k+j} + \sum_{l\neq j,k+j} |h_{k+j,l}| \Big| \Big\}\, dt \\ &\leqq \sum_{j=\nu,\mu} \int_0^T \Big\{ \sum_{l=1}^{2k} |h_{j,l}| + \sum_{l\neq j} |h_{k+j,l}| \Big\}\, dt \\ &< T \min_n \Big| \omega_\nu + \omega_\mu - \frac{2n\pi}{T} \Big|. \end{aligned} \tag{4.8}$$

Similarly, if H_ν and H_μ are formed from rows of K_1 we have

$$\begin{aligned}\sum_{j=\nu,\mu} \|H_j\| &= \sum_{j=\nu,\mu} \int_0^T \Big\{ \Big| h_{jj} - \sum_{l\neq j,k+j} |h_{j,l}| \Big| + |h_{j,k+j}| \\ &\qquad + \Big| h_{k+j,k+j} - \sum_{l\neq j,k+j} |h_{k+j,l}| \Big| \Big\}\, dt \\ &< T \min_n \Big| \omega_\nu + \omega_\mu - \frac{2n\pi}{T} \Big|. \end{aligned} \tag{4.9}$$

Since (4.8) and (4.9) hold for any pair $\nu, \mu \in \{1, \cdots, k\}$, the Hamiltonians $\Omega + K_1$ and $\Omega + K_2$ are strongly stable and lie in $O_n^{(\sigma)}$ by the portion of Theorem 1.1 that applies to systems of the form (4.2).

In the same way, one can show that the Hamiltonians

$$\eta[\Omega + K_1] + (1-\eta)[\Omega + K_2] = \Omega + \eta K_1 + (1-\eta)K_2$$

lie in $O_n^{(\sigma)}$ for $0 \leqq \eta \leqq 1$. In particular, these Hamiltonians lie in M'. Moreover, we have

$$\Omega + K_1 \leqq \Omega + H \leqq \Omega + K_2,$$

by (4.7). It therefore follows from Theorem 4.1 that $\Omega + H \in O_n^{(\sigma)}$. This completes the proof of Theorem 1.1.

5. Classical results for Hill's equation. The equation

$$y''+p(t)y=0, \qquad p(t+T)=p(t) \tag{5.1}$$

is known as Hill's equation. It can be written as a special case of (1.1) by taking $x_1=y'$, $x_2=y$. A classical result due to Lyapunov states that all solutions of (5.1) are bounded on $(-\infty,\infty)$ if

$$p(t)\begin{cases}\geqq 0\\ \not\equiv 0\end{cases} \quad \text{and} \quad T\int_0^T p(t)\,dt<4.$$

This result was generalized by Krein in [8]. He showed that all solutions of (5.1) are bounded if for some integer $n\geqq 1$,

$$p(t)\geqq\frac{n^2\pi^2}{T^2} \quad \text{and} \quad T\int_0^T p(t)\,dt<n^2\pi^2+2n\pi(n+1)\tan\frac{\pi}{2(n+1)}. \tag{5.2}$$

Since

$$\tan\frac{\pi}{2(n+1)}=\frac{\pi}{2(n+1)}+\frac{\pi^3}{24(n+1)^3}+\cdots>\frac{\pi}{2(n+1)},$$

it follows that all solutions of (5.1) are bounded if for some integer $n\geqq 1$ we have

$$p(t)\geqq\frac{n^2\pi^2}{T^2} \quad \text{and} \quad T\int_0^T p(t)\,dt<n^2\pi^2+n\pi^2. \tag{5.3}$$

Moreover, this condition is only slightly stronger than (5.2) for large values of n. We shall now deduce the stability condition (5.3) from results we obtained in § 3. For this purpose, we need a comparison theorem for second-order systems from [2, p. 682].

THEOREM 5.1. *Let $x^1(t)$ and $x^2(t)$ be solutions of* (3.2) *with Hamiltonians $H_1(t)$ and $H_2(t)$ respectively, and let $x^1(0)=x^2(0)\neq 0$. Let* Arg $(x^j(t))$ *be a continuous branch of the argument of $x^j(t)$ satisfying* Arg $(x^1(0))=$ Arg $(x^2(0))$. *If $H_1(t)\geqq H_2(t)$, $0\leqq t\leqq T$, then $x_1(t)$ rotates "ahead" of $x_2(t)$ in the sense that* Arg $(x^1(t))\geqq$ Arg $(x^2(t))$, $0\leqq t\leqq T$. *If there is a set of positive measure in the interval $(0,t_0)$ on which $H_1(t)>H_2(t)$, then* Arg $(x^1(t))>$ Arg $(x^2(t))$ *for $t\geqq t_0$.*

From this theorem, it foltows that if H_2 lies in one of the sets π_n^{**} in Fig. 3.1, and $H_1(t)\geqq H_2(t)$, $0\leqq t\leqq T$, then H_1 lies in one of the sets π_n^{**}, O_n, π_{n+1}^{*-}, π_{n+1}^{**}, H_{n+1}, $\cdots$ to the right of π_n^{**} in Fig. 3.1.

Now consider Hill's equation (5.1). Let $p(t)=(n^2\pi^2/T^2)+h(t)$, where $h(t)\geqq 0$. (5.1) is then equivalent to the Hamiltonian system.

$$J_2\begin{pmatrix}x_1\\ x_2\end{pmatrix}'=\left[\begin{pmatrix}\frac{n\pi}{T} & 0\\ 0 & \frac{n\pi}{T}\end{pmatrix}+\begin{pmatrix}\frac{T}{n\pi}h(t) & 0\\ 0 & 0\end{pmatrix}\right]\begin{pmatrix}x_1\\ x_2\end{pmatrix}. \tag{5.4}$$

Since the Hamiltonian

$$\begin{pmatrix}\frac{n\pi}{T} & 0\\ 0 & \frac{n\pi}{T}\end{pmatrix}=\frac{n\pi}{T}I_2$$

lies in π_n^{**}, and $h(t)\geqq 0$, the Hamiltonian in (5.4) lies in one of the sets π_n^{**}, O_n, π_{n+1}^{*-}, π_{n+1}^{**}, H_{n+1}, $\cdots$ to the right of π_n^{**} in Fig. 3.1. Thus, the Hamiltonian in (5.4) lies in π_n^{**} or in O_n if the norm of the Hamiltonian

$$\begin{pmatrix}\frac{T}{n\pi}h(t) & 0\\ 0 & 0\end{pmatrix}$$

is less than the distance from $(n\pi/T)I_2$ to the right boundary of O_n. It follows from (3.5) that the distance from $(n\pi/T)I_2$ to the right boundary of O_n is $\geqq \pi$. Thus, the system (5.4) lies in $\pi_n^{**} \cup O_n$, and hence, all its solutions are bounded, if

$$\frac{T}{n\pi}\int_0^T h(t)\,dt < \pi.$$

Since $p(t) = (n^2\pi^2/T^2) + h(t)$, this condition is equivalent to

$$T\int_0^T p(t)\,dt < n^2\pi^2 + n\pi^2.$$

Thus, we have shown that conditions (5.3) imply stability of (5.1).

There are several similar stability conditions for Hill's equation in [9, pp. 61–63]. These results are also very closely related to the conditions we derived in § 3. Also, results similar to ours are contained in [10], [11], [12] and [13].

REFERENCES

[1] I. M. GELFAND AND V. B. LIDSKII, *On the structure of the domain of stability of linear canonical systems of differential equations with periodic coefficients*, Uspehi Mat. Nauk (N.S.), 10 (1955), pp. 3–40, in Russian; Amer. Math. Soc. Transl., 2, 8 (1958), pp. 143–181.

[2] V. A. YAKUBOVICH AND V. M. STARZHINKII, *Linear Differential Equations with Periodic Coefficients* (2 Vols.), John Wiley, New York, 1975.

[3] W. A. COPPEL AND A. HOWE, *On the stability of linear canonical systems with periodic coefficients*, J. Austral. Math. Soc., 5 (1965), pp. 169–198.

[4] M. LEVY, *Stability of Linear Hamiltonian Systems with Periodic Coefficients*, IBM Research Report RC 610 (#28482), June, 1977, IBM Watson Research Center, Yorktown Heights, NY 10598.

[5] J. MOSER, *New aspects in the theory of stability of Hamiltonian systems*, Comm. Pure Appl. Math., 11 (1958), pp. 81–114.

[6] E. R. BARNES, *Stability conditions for second ordinary differential equations with periodic coefficients*, this Journal, 10 (1979), pp. 985–1001.

[7] ———, *Stability Conditions for Linear Hamiltonian Systems with Periodic Coefficients*, IBM Research Report RC 7865 (#34027), September 1979, IBM Watson Research Center, Yorktown Heights, NY 10598.

[8] M. G. KREIN, *On certain problems on the maximum and minimum characteristic values on Liapunov zones of stability*, Prikl. Math. Meh., 15 (1951) (in Russian); Amer. Math. Soc. Transl., 2, 1 (1955), pp. 163–187.

[9] L. CESARI, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Academic Press, New York, 1963.

[10] K. L. CHIOU, *The geometry of indefinite J-space and strong stability criteria of canonical differential equations with periodic coefficients*, this Journal, 8 (1977), pp. 118–126.

[11] ———, *The geometry of indefinite J-space and stability behavior of linear systems*, J. Differential Equations, 28 (1978), pp. 104–123.

[12] YU. L. DALECKII AND M. G. KREIN, *Stability of Solutions of Differential Equations in Banach Space*, Translations of Mathematical Monographs, Vol. 43, American Mathematical Society, Providence RI.

[13] R. W. BROCKETT, *Stability of Periodic Linear Systems and the Geometry of Lie Groups*, in Dynamical Systems, L. Cesari, J. Hale, J. P. LaSalle, eds., Academic Press, New York, 1976.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0007 $01.00/0

AN ORDERING OF OSCILLATION TYPES FOR $y^{(n)}+py=0$*

GARY D. JONES†

Abstract. Assuming p is continuous and sign definite, we give an ordering of oscillation types for

$$y^{(n)}+py=0.$$

1. Introduction.

We will consider the differential equation

$$y^{(n)}+py=0, \tag{1}$$

where p will be assumed continuous and sign definite.

We will say that (a, b) is a $(k, n-k)$ interval of oscillation provided there is a solution of (1) which is positive on (a, b) with zeros of order not less than k and $n-k$ at a and b respectively. If for every $M>0$ there is a $(k, n-k)$ interval of oscillation in $[M, \infty)$, we will say that (1) has $(k, n-k)$ oscillation type. If no such interval exists on $[c, \infty)$, we will say that (1) is $(k, n-k)$ disconjugate there. Equation (1) will be said to be eventually $(k, n-k)$ disconjugate if it is $(k, n-k)$ disconjugate on $[b, \infty)$ for some large b.

Levin [6] stated and Nehari [7] proved that if (1) has a $(k, n-k)$ interval of oscillation then $n-k$ must be even or odd according as to whether p is negative or positive, respectively.

If p is a constant, then (1) has every possible oscillation type subject to the restrictions mentioned above. However, by considering equations (1) of the Euler type, it can be shown that (1) can have some oscillation types but not others [3]. It is the purpose of this paper to show that there is an ordering of the oscillation types for the equations (1). That is, we will show that the existence of certain oscillation types implies the existence of others.

Our method will be to use a characterization of eventual $(k, n-k)$ disconjugacy in terms of the existence of certain monotone solutions of (1) due to Nehari [7] to develop theorems comparing (1) to lower order equations. Our main theorems concerning the ordering of oscillation types will then follow.

More comparison theorems of the type used here are contained in [2]. As a matter of fact, Theorems 2 and 4 use the same idea as is used in [2, Theorem 6]. Also, Theorems 1 and 3 can be proved using [2, Theorem 8].[1]

2. Preliminary results.

To prove our main result, we will need the following inequality, which follows from an inequality due to Kiguradze [4].

LEMMA. *Let* $y \in C^{k+1}[0, b)$ *with* $y^{(i)}(x)>0$, $y^{(k+1)}(x) \leqq 0$ *on* $(0, b)$ *for* $i=0, 1, \cdots, k$ *and* $y(0)=y'(0)=\cdots=y^{(k-1)}(0)=0, y^{(k)}(0)=1$. *Then*

$$y(x)/y^{(j)}(x) \geqq (k-j)!\,x^j/k! \tag{2}$$

on $(0, b)$ *for* $j=1, 2, \cdots, k-1$ *and* $k=2, 3, \cdots$.

Proof. For $k=2$ the result is due to Lazer [5]. Suppose (2) holds for $k<m$.

Let

$$H(x)=xy(x)-x^m y^{(m-1)}(x)/m!.$$

---

* Received by the editors February 15, 1980, and in revised form June 6, 1980.

† Mathematics Department, Murray State University, Murray, Kentucky 42071.

[1] The author is indebted to the referee for pointing out reference [2].

Then

$$H'(x)=xy'(x)+y(x)-x^{m-1}y^{(m-1)}(x)/(m-1)!-x^m y^{(m)}(x)/m!.$$

By the inductive hypothesis (applying (2) to $y'(x)$), $xy'(x)-x^{m-1}y^{(m-1)}(x)/(m-1)!\geqq 0$ and by Taylor's formula $y(x)-x^m y^{(m)}(x)/m!\geqq 0$. Since $H(0)=0$ and H is nondecreasing, (2) holds for $k=m$ and $j=m-1$.

Now, assume that (2) holds for $k<m$ and for $k=m$ and $j>r$. Let

$$R(x)=x^{m-r}y(x)/(m-r)!-x^m y^{(r)}(x)/m!$$

Then,

$$R'(x)=x^{m-r-1}y(x)/(m-r-1)!+x^{m-r}y'(x)/(m-r)!$$
$$-x^{m-1}y^{(r)}(x)/(m-1)!-x^m y^{(r+1)}(x)/m!.$$

But $x^{m-r-1}y(x)/(m-r-1)!-x^m y^{(r+1)}(x)/m!\geqq 0$, by applying (2) to y for $k=m$ and $j=r+1$. And $x^{m-r}y'(x)/(m-r)!-x^{m-1}y^{(r)}(x)/(m-1)!\geqq 0$, by applying (2) to y' with $k=m-1$ and $j=r-1$. Again, since $R(0)=0$ and R is nondecreasing, (2) holds for $k=m$ and $j=r$. Hence, (2) is valid for all k.

3. Comparison theorems. Our main result will follow from the following four comparison theorems.

THEOREM 1. *If*

$$y^{(n-2)}+(x^2p(x)/(k-1)(k-2))y=0, \tag{3}$$

is eventually $(n-k, k-2)$ *disconjugate (with k odd for p positive and k even for p negative), then* (1) *is eventually* $(n-k, k)$ *disconjugate.*

Proof. If (3) is $(n-k, k-2)$ disconjugate, then there is a solution y of (3) such that

$$y^{(i)}(x)>0 \quad \text{for } i=0, 1, \cdots, n-k-1,$$

and

$$(-1)^i y^{(i+n-k)}(x)>0 \quad \text{for } i=0, 1, \cdots, k-2$$

for all large x [7]. Since

$$\lim_{x\to\infty} y^{(n-k+j)}(x)=0 \quad \text{for } j\geqq 1,$$

and

$$\lim_{x\to\infty} y^{(n-k)}(x)\geqq 0$$

exist, we have upon integrating (3),

$$\begin{aligned} y^{(n-k)}(x) &\geqq \frac{1}{(k-3)!}\int_x^\infty (t-x)^{k-3}\frac{t^2|p(t)|}{(k-2)(k-1)}y(t)\,dt \\ &\geqq \frac{1}{(k-1)!}\int_x^\infty (t-x)^{k-1}|p(t)|\,y(t)\,dt. \end{aligned} \tag{4}$$

Now integrating (4) from some fixed b to x, we have

$$\begin{aligned} y(x)\geqq y(b)+y'(b)(x-b)+\cdots+y^{(n-k-1)}(x)\frac{(x-b)^{n-k-1}}{(n-k-1)!} \\ +\int_b^x \frac{(x-t)^{n-k-1}}{(n-k-1)!}\int_t^\infty \frac{(s-t)^{k-1}}{(k-1)!}|p(s)|\,y(s)\,ds\,dt. \end{aligned} \tag{5}$$

By the monotone convergence theorem there is a function $z \geqq 0$ that satisfies the equality (5). Differentiating, we have

$$z^{(i)}(x) > 0 \qquad \text{for } i = 0, 1, \cdots, n-k,$$

and

$$(-1)^i z^{(n-k+i)}(x) > 0 \quad \text{for } i = 1, \cdots, k,$$

where z satisfies (1). Thus, by [1] (also see [3]) the result follows.

THEOREM 2. *If* (1) *is* $(n-k, k)$ *disconjugate on* $[0, \infty)$ (*with k odd for p positive and k even for p negative*), *then*

$$y^{(n-2)} + \frac{x^2}{(n-k)(n-k-1)} p(x) y = 0, \tag{6}$$

is eventually $(n-k-2, k)$ *disconjugate.*

Proof. Since (1) is $(n-k, k)$ disconjugate on $[0, \infty)$, there is a solution y of (1) such that

$$y^{(i)}(x) > 0 \qquad \text{for } i = 0, 1, \cdots, n-k,$$
$$(-1)^i y^{(n-k+i)}(x) > 0 \quad \text{for } i = 1, \cdots, k,$$

for $x > 0$ with $y^{(i)}(0) = 0$ for $i = 0, 1, \cdots, n-k-1$, and $y^{(n-k)}(0) = 1$ [7]. Since

$$y^{(n)} + py = y^{(n)} + (py/y'')y'',$$

it follows that

$$z^{(n-2)} + (py/y'')z = 0 \tag{7}$$

has a solution z such that

$$z^{(i)}(x) > 0 \qquad \text{for } i = 0, 1, \cdots, n-k-2,$$
$$(-1)^i z^{(n-k-2+i)}(x) > 0 \quad \text{for } i = 1, \cdots, k,$$

for $x > 0$. It then follows [1] that (7) is eventually $(n-k-2, k)$ disconjugate. But by the lemma,

$$y(x)/y''(x) \geqq x^2/(n-k)(n-k-1) \quad \text{for } x > 0.$$

Hence, by known comparison theorems [7], it follows that

$$y^{(n-2)} + \frac{x^2}{(n-k)(n-k-1)} p(x) y = 0$$

is eventually $(n-k-2, k)$ disconjugate.

The following two theorems will be used in the case where (1) is of odd order. Since their proofs are essentially the same as the proofs of Theorems 1 and 2, they will be omitted.

THEOREM 3. *If*

$$y^{(n-1)} + \frac{xp(x)}{k-1} y = 0$$

is eventually $(n-k, k-1)$ *disconjugate* (*with k odd for p positive and k even for p negative*), *then*

$$y^{(n)} - py = 0$$

is eventually $(n-k, k)$ *disconjugate.*

THEOREM 4. *If* (1) *is* $(n-k, k)$ *disconjugate on* $[0, \infty)$ (*with k odd for p positive and k even for p negative*), *then*

$$y^{(n-1)}+\frac{x}{n-k}py=0$$

is eventually $(n-k-1, k)$ *disconjugate.*

4. Case where *n* is even. The following theorem is valid whether n is even or odd, but we will use it to give an ordering of the oscillation types in the case where n is even.

THEOREM 5. *If* $k \leqq (n+1)/2$ *and* (1) *is eventually* $(k, n-k)$ *disconjugate* (*with* $n-k$ *odd for p positive and even for p negative*), *then it is eventually* $(k-2, n-k+2)$ *disconjugate.*

Proof. Suppose (1) is $(k, n-k)$ disconjugate on $[c, \infty)$. Then

$$y^{(n)}+p(t)y=0 \tag{8}$$

is disconjugate on $[0, \infty)$ where $t=x-c$. By Theorem 2,

$$y^{(n-2)}+\frac{t^2}{(k)(k-1)}p(t)y=0$$

is eventually $(k-2, n-k)$ disconjugate. Since $k \leqq (n+1)/2$, $k \leqq n-k+1$ and $k-1 \leqq n-k$. Thus, by known comparison theorems [5],

$$y^{(n-2)}+\frac{t^2}{(n-k+1)(n-k)}p(t)y=0$$

is eventually $(k-2, n-k)$ disconjugate. But, by Theorem 1, (8) is eventually $(k-2, n-k+2)$ disconjugate, and it follows that (1) is also.

Since (1) is self-adjoint when n is even, it is $(k, n-k)$ disconjugate if and only if it is $(n-k, k)$ disconjugate. Thus, applying Theorem 5, we obtain the following ordering on the oscillation types.

COROLLARY 1. *If p is positive and n is even but not divisible by* 4, *then eventual* $(n/2+2i, n/2-2i)=(n/2-2i, n/2+2i)$ *disconjugacy implies eventual* $(n/2+2(i+1), n/2-2(i+1))=(n/2-2(i+1), n/2+2(i+1))$ *disconjugacy for* $i=0, \cdots, (n-6)/4$.

COROLLARY 2. *If p is negative and n is divisible by* 4, *then the same conclusion as in Corollary* 1 *holds, except* $i=0, \cdots, (n-8)/4$.

COROLLARY 3. *If p is positive and n is divisible by* 4, *then eventual* $(n/2+1+2i, n/2-1-2i)=(n/2-1-2i, n/2+1+2i)$ *disconjugacy implies eventual* $(n/2+3+2i, n/2-3-2i)=(n/2-3-2i, n/2+3+2i)$ *disconjugacy for* $i=0, \cdots, (n-8)/4$.

COROLLARY 4. *If p is negative, n even but not divisible by* 4, *then the same conclusion holds as in Corollary* 3 *except* $i=0, 1, \cdots, (n-10)/4$.

As an immediate consequence of Corollaries 1–4, we have the following theorems:

THEOREM 6. *If p is positive and n is even but not divisible by* 4 *or negative and n is divisible by* 4, *then* (1) *is eventually disconjugate provided it is eventually* $(n/2, n/2)$ *disconjugate.*

THEOREM 7. *If p is positive and n is divisible by* 4 *or p is negative and n is even but not divisible by* 4, *then* (1) *is eventually disconjugate provided it is eventually* $((n+1)/2, (n-1)/2)$ *disconjugate.*

5. Case where *n* is odd. In order to give an ordering of the oscillation types for the odd order case, we will need the following theorem.

THEOREM 8. *If n is odd, $k \geqq n/2$ and* (1) *is eventually $(n-k, k)$ disconjugate (with k odd for p positive and even for p negative), then it is eventually $(k+1, n-k-1)$ disconjugate. If $k < n/2$, then eventual $(n-k, k)$ disconjugacy implies eventual $(k-1, n-k+1)$ disconjugacy.*

Proof. Suppose (1) is $(n-k, k)$ disconjugate on $[c, \infty)$. Then, (8) is disconjugate on $[0, \infty)$. Thus, by Theorem 4,

$$y^{(n-1)} + \frac{t}{(n-k)} p(t) y = 0,$$

is eventually $(n-k-1, k)$ disconjugate. Since $k \geqq n-k$, we have, by [5], that

$$y^{(n-1)} + \frac{t}{k} p(t) y = 0$$

is also eventually $(n-k-1, k)$ disconjugate. Thus, by Theorem 3,

$$y^{(n)} - p(t) y = 0$$

is eventually $(n-k-1, k+1)$ disconjugate. Thus, (1) is $(k+1, n-k-1)$ disconjugate.

The second half of the theorem can be obtained from the first by applying it to the adjoint of (1).

COROLLARY 1. *If p is positive, n and $(n+1)/2$ are odd, then eventual $((n-1)/2-2i, (n+1)/2+2i)$ disconjugacy implies eventual $((n+3)/2+2i\ (n-3)/2-2i)$ disconjugacy for $i = 0, \cdots, (n-5)/4$. Also, eventual $((n-1)/2+2i, (n+1)/2-2i)$ disconjugacy implies eventual $((n-1)/2-2i, (n+1)/2+2i)$ disconjugacy for $i = 1, \cdots, (n-5)/4$.*

COROLLARY 2. *If p is negative, n odd and $(n+1)/2$ is even, then the same conclusions hold as in Corollary* 1, *except that $i = 0, \cdots, (n-7)/4$ in the first case and $i = 1, \cdots, (n-3)/4$ in the second.*

COROLLARY 3. *If p is positive and $(n+1)/2$ is even, then $((n+1)/2+2i, (n-1)/2-2i)$ eventual disconjugacy implies eventual $((n-3)/2-2i, (n+3)/2+2i)$ disconjugacy for $i = 0, \cdots, (n-7)/4$. Also, eventual $((n+1)/2-2i, (n-1)/2+2i)$ disconjugacy implies eventual $((n+1)/2+2i, (n-1)/2-2i)$ disconjugacy for $i = 1, \cdots, (n-3)/4$.*

COROLLARY 4. *If p is negative and $(n+1)/2$ is odd, then the same conclusions hold as in Corollary* 3, *except $i = 0, \cdots, (n-5)/4$ in the first case and $i = 1, \cdots, (n-5)/4$ in the second.*

As for the case when n is even, we obtain the following theorems about eventual disconjugacy.

THEOREM 9. *If p is positive and $(n+1)/2$ is odd or if p is negative and $(n+1)/2$ is even, then* (1) *is eventually disconjugate if it is eventually $((n-1)/2, (n+1)/2)$ disconjugate.*

THEOREM 10. *If p is positive and $(n+1)/2$ is even or if p is negative and $(n+1)/2$ is odd, then* (1) *is eventually disconjugate if it is eventually $((n+1)/2, (n-1)/2)$ disconjugate.*

That none of the implications of Corollaries 1–4 of Theorems 5 or 8 are reversible in general can be seen from examples like those given in [3].

REFERENCES

[1] U. ELIAS, *Nonoscillation and eventual disconjugacy*, Proc. Amer. Math. Soc., 66 (1977), pp. 269–275.

[2] ——— *Necessary conditions and sufficient conditions for disfocality and disconjugacy of differential equations*, Pacific J. Math., 81 (1979), pp. 379–397.

[3] G. JONES, *Oscillation properties of* $y^n + py = 0$, Proc. Amer. Math. Soc., to appear.
[4] I. T. KIGURADZE, *Oscillation properties of solutions of certain ordinary differential equations*, Soviet Math. Dokl., 3 (1963), pp. 649–652.
[5] A. C. LAZER, *The behavior of solutions of the differential equation* $y''' + p(x)y' + q(x)y = 0$, Pacific J. Math., 17 (1966), pp. 435–466.
[6] A. J. LEVIN, *Some questions on the oscillation of solutions of a linear differential equation*, Dokl. Akad. Nauk., 148 (1963), pp. 512–515.
[7] Z. NEHARI, *Green's function and disconjugacy*, Arch. Rational Mech. Anal., 62 (1976), pp. 53–76.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0008 $01.00/0

ON THE INTERVAL OF DISCONJUGACY OF LINEAR, AUTONOMOUS DIFFERENTIAL EQUATIONS*

I. TROCH†

Abstract. A lower and an upper bound for the maximum length of an interval of disconjugacy are given which depend only on the maximum eigenfrequency of the differential equation. These bounds are shown to be best possible in case no further information on the zeros of the characteristic equation is available. Further, a procedure is given which allows one to calculate the length of a maximal interval of disconjugacy from determinants of order $\leqq n/2$, where n denotes the order of the differential equation or system. Applications of these results are to be found in control theory, mathematical system theory and approximation theory.

1. Introduction. Optimal control problems often result in bang-bang controls, i.e. controls which are, roughly speaking, piecewise constant and of maximum value. The switching times are determined as the zeros of certain functional relations. In the case of linear autonomous systems these relations are of the form $z(t) = b^T \cdot \exp(At) \cdot c$, where A is a constant square matrix and b and c denote constant vectors (for further details see [2], [6], [8]). It is not only of theoretical interest to have some information about the possible number of switchings, but of importance, also, for the practical computation of such controls.

Further, in mathematical system theory the questions whether a linear system can be controlled by impulses or piecewise constant functions or, conversely, can be observed by sampling the output, lead to similar problems [10]–[13].

These remarks may illustrate the practical importance of the availability of good estimates on the number N of zeros in an interval I of an arbitrary solution of a linear autonomous differential equation

$$M[z] = z^{(n)} + a_{n-1}z^{(n-1)} + \cdots + a_1\dot{z} + a_0 z = 0, \tag{1}$$

with characteristic polynomial (equation)

$$m(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0, \tag{2}$$

where $a_0, \cdots, a_{n-1}$ are real constants. One possibility of getting the desired estimates is by determining the possible intervals of disconjugacy of (1): in such an interval, any solution of (1) has at most $n-1$ zeros.

For systems (1) having only real characteristic roots (these are the zeros of $m(\lambda) = 0$), it is well known, [8], [4], that $(-\infty, +\infty)$ is an interval of disconjugacy. Therefore, only the case where (2) has at least one pair of nonreal characteristic roots will be investigated further. In the following, the maximum imaginary part of these characteristic roots, i.e., the maximum eigenfrequency of (1), will be denoted by ω. Equation (1) is autonomous, and therefore the greatest possible length η_0 of an interval of disconjugacy does not depend on the location of that interval on the real axis. Such a maximum interval will be denoted by I_0.

In the following, estimates for η_0 will be given which use only the maximum eigenfrequency ω of the system described by (1). They can be applied also in case only bounds for ω are available. Results in [1] that offer a possibility to determine η_0 exactly are used for this purpose as well as for the derivation of results which are of more theoretical interest, such as continuous dependence of η_0 on parameters.

* Received by the editors April 2, 1980, and in revised form June 2, 1980.
† Technische Universität Wien, Karlsplatz 13, A-1040 Vienna, Austria.

Finally, the dependence on the number r of real roots of (2) is investigated. It turns out that r has only little influence on η_0. Nevertheless, knowledge of r can be of great importance and help when bounds on the number of zeros within an interval of given length greater than η_0 are to be given.

2. Main results. As has been pointed out, the maximum length η_0 of intervals of disconjugacy for (1) will be investigated. The estimate

$$\pi/\omega \leqq \eta_0 \leqq (n-1)\pi/\omega$$

has been known for a long time, and Hájek made in [3] the conjecture that this inequality is best possible at this state of generality, i.e., without detailed knowledge about the characteristic roots of (2). This is clarified in Theorems 1 and 2. Further, in Proposition 1 a simplified "test for disconjugacy" is presented, which allows the computation of η_0 from inspection of at most $n/2$ Wronskians of order $\leqq n/2$, instead of n Wronskians up to order η as in the case of nonautonomous equations.

THEOREM 1. *Let ω and I_0 be defined as above. Then*:

(a) *For every equation* (1) *the maximum intervals of disconjugacy are half-open intervals,* $[\alpha, \alpha+\eta_0)$ *or* $(\alpha, \alpha+\eta_0]$ *respectively, with length* η_0 *independent of the real constant* α.

(b) *For every equation* (1) *of second order,*

$$\eta_0 = \pi/\omega. \tag{3}$$

(c) *For equation* (1) *with* $n \geqq 3$ *the maximum length satisfies*

$$\pi/\omega < \eta \leqq (n-1)\pi/\omega. \tag{4}$$

The earlier results [3], [11] cited in the introduction, are improved only slightly by this. In fact, only an equality sign has been removed. But Theorem 2 shows that (3) and (4) are in fact best possible estimates at this state of generality. These results allow us further to consider in the following only intervals with one endpoint equal to zero.

Proof. Assertion (a) is a consequence of Proposition 1 and Lemma 1 in connection with the autonomy of (1). Assertions (b) and (c) are consequences of Propositions 2 and 3.

It should be mentioned that an upper and/or lower bound for ω can be used to derive bounds on η_0 also.

THEOREM 2. *Let the notation be as in Theorem* 1. *Then*:

(a) *For any natural number* $n \geqq 3$ *and every constant* $\hat{\eta}$ *which satisfies* (4), *there exists at least one differential equation* (1) *of order n, such that* $\eta_0 = \hat{\eta}$ *holds for it.*

(b) *Let the coefficients in* (1) *be continuous functions of a real parameter s. Then* $\eta_0 = \eta_0(s)$ *is also a continuous function of s.*

Proof. This is a consequence of Theorem 1 and Propositions 4 and 6.

The statements above imply the continuous dependence of η_0 on the coefficients of (1), or, equivalently, on its characteristic roots. These results may also be used to improve a result of Hájek [3], on the number N of zeros of an arbitrary solution of (1) on an interval of given length T. Let $n \geqq 3$ and denote by r the number of real zeros of (2), which, therefore, can be rewritten as

$$m(\lambda) = q(\lambda)p(\lambda), \tag{5}$$

where $q(\cdot)$ is of degree r and has only real zeros. Consequently, $p(\cdot)$ is of (even) degree $n-r$ and has only nonreal zeros. The differential operator corresponding to $p(\cdot)$ shall be denoted by $P[\cdot]$. Further, let $[x]^*$ be the smallest integer $k \geqq x$. Then the following holds:

THEOREM 3. *Assume that* (5) *holds for* (2). *Then any solution of* (1) *has in an arbitrary nonclosed interval of length* $T>0$ *at most* N *zeros, where*

$$N \leqq r+(n-r-1)[T/\eta_{01}]^*,$$

with η_{01} *being the maximum length of an interval of disconjugacy of the differential equation* $P[z]=0$.

With this the result given already in [3] is a consequence of inequality (4):

COROLLARY. *With the notation from above,*

$$N \leqq r+(n-r-1)[T\omega/\pi]^*.$$

The proof of this theorem as well as those of the propositions cited are given in § 4 in detail. They show that the number of real roots of (2) has a somewhat unexpected influence on the size of η_0. From the knowledge that for $r=n$ the interval I_0 equals the real axis, one might expect that for given n and ω the length η_0 would be, roughly speaking, an increasing function of the number r. But, on the contrary, the lower bound for η_0 is proved by the investigation of differential equations which have only one pair of purely imaginary and $n-2$ real characteristic roots. Moreover, the upper bound in (4) is reached by equations which have only pure imaginary characteristic roots of multiplicity one. Further, it seems that the resulting trigonometric polynomials are the only examples for $\eta_0=(n-1)\pi/\omega$.

Nevertheless, the knowledge of the number of real characteristic roots will be advantageous whenever the number N of zeros in an interval of given length T must be estimated. The straightforward inequality

$$N \leqq (n-1)[T/\eta_0]^* \leqq (n-1)[T\omega/\pi]^*$$

is improved greatly by Theorem 3 and its corollary.

Finally, it will be pointed out that Lemmas 4–6 of § 3 provide some information on the relation between the numbers of zeros of a rather arbitrary function $\phi(\cdot)$ and a differential expression $L[\phi]=\ddot{\phi}+2\delta\dot{\phi}+(\delta^2+\omega^2)\phi$ with real δ and ω. Further, some reflections on the dependence of η_0 on the order n of (1) will be given.

3. Preliminaries. During these investigations the terms "disconjugacy," "Chebyshev-system," etc., will be used as defined in [1]. There, it is also clarified that for nonclosed intervals I the property "(1) is disconjugate on I" does not depend on whether the zeros are counted according to their multiplicity or not. A useful test for disconjugacy based on Wronski-determinants is given in [1, Chapt. 3], and is cited in Lemmas 1, 2. The following notation for the Wronskian will be used:

$$W_k(x(t))=W_k(x)=W(x,\dot{x},\cdots,x^{(k-1)})=\begin{vmatrix} x & \dot{x} & \cdots & x^{(k-1)} \\ \dot{x} & \ddot{x} & \cdots & x^{(k)} \\ \cdots & \cdots & & \cdots \\ x^{(k-1)} & \cdots & & x^{(2k-2)} \end{vmatrix}, \qquad 1 \leqq k \leqq n-1.$$

Further, a nontrivial solution of (1) will be said to have *property* (k, a, b), $1 \leqq k \leqq n-1$, if it has a zero of multiplicity $\geqq k$ at b and a zero of multiplicity $\geqq n-k$ at a.

Now, the solution $z_n(t, t_0)$ ($z_n(t)$ iff $t_0 = 0$) of our standard initial value problem (δ_{ik} is the Kronecker symbol)

$$\text{(SIVP)} \qquad \begin{aligned} &z^{(n)} + a_{n-1}z^{(n-1)} + \cdots + a_0 z = 0, \\ &z^{(k)}(t_0) = \delta_{k,n-1}, \qquad k = 0, 1, \cdots, n-1 \end{aligned}$$

is of great importance:

LEMMA 1 [1, p. 99]. *Equation* (1) *has a solution with property* (k, a, b) *if and only if* $W_k(z_n(b, a)) = 0$.

LEMMA 2 [1]. *The number* η_0 *defined earlier is given by*

$$\eta_0 = \min \{T > 0 \mid W_k(z_n(T)) = 0 \text{ for at least one } k \text{ with } 1 \leqq k \leqq n-1\}.$$

As is well known, [5], solutions of autonomous differential equations are invariant under shifting (i.e., under transformations of time $t' = t - t_0$). By definition, a solution has property (k, a, b) iff it has property $(n-k, b, a)$. Thus, from Lemma 1 follows that

$$W_k(z_n(b, a)) = 0 \quad \text{iff} \quad W_{n-k}(z_n(a, b)) = 0.$$

Making use of the autonomy, we see that this is equivalent to

$$W_k(z_n(b-a, 0)) = 0 \quad \text{iff} \quad W_{n-k}(z_n(a-b, 0)) = 0,$$

or

$$W_k(z_n(T)) = 0 \quad \text{iff} \quad W_{n-k}(z_n(-T)) = 0.$$

Combination of this with Lemma 2 yields a test for disconjugacy which requires only the calculation of Wronskians with order $\leqq n/2$ and, consequently, has considerable advantage.

PROPOSITION 1. *The number* η_0 *is given by*

$$\eta_0 = \min \{|T| \neq 0 \mid W_k(z_n(T)) = 0 \text{ for at least one } k \text{ with } 1 \leqq k \leqq [n/2]\},$$

where $[a]$ *denotes the largest integer* $j \leqq a$.

The next assertion allows important simplifications in the following considerations. It is an immediate consequence of Leibniz' rule for $(uv)^{(j)}$ and the definition of the order of a zero.

LEMMA 3. *Let* $z(t)$ *be an* n *times continuously differentiable function on* I *with a zero of order* $k < n$ *at* $t = t_0 \in I$. *Then, for any real* β,

$$w(t) := z(t) \exp(\beta t)$$

has a zero of order k *at* $t = t_0$.

The following Lemma 4 is a rather trivial but useful reformulation of a result of Hájek [3], whereas Lemma 5 generalizes these results to the case of multiple zeros. Its proof uses rather straightforward modifications of the ideas used in [3].

LEMMA 4. *Let a real twice differentiable function* ϕ *which is not a simple harmonic with frequency* $\omega = \pi/(t_2 - t_1)$, *have two consecutive zeros* $t_1 < t_2$. *Then, at some intermediate point* $t' \in (t_1, t_2)$,

$$\ddot{\phi}(t')/\phi(t') < -\alpha^2$$

holds for all $0 < \alpha \leqq \omega$.

Proof. In [3] the existence of such a t' with $\ddot{\phi}(t')/\phi(t') < -\omega^2$ is proved. Combining this with $-\omega^2 \leqq -\alpha^2 < 0$ gives the assertion.

LEMMA 5. *Let a twice differentiable real function* ϕ *have* N *zeros (counting multiplicities) on* $[0, \pi/\omega] = I$. *Then,* $L[\phi] = \ddot{\phi} + \omega^2\phi$ *has at least* $N-2$ *zeros in* I.

Proof. In the case where ϕ is a simple harmonic with frequency Ω on a subinterval of I with length $\leqq \pi/\Omega$, the assertion is trivial. Therefore, this can be excluded. Assume that there are M distinct zeros t_k with multiplicities m_k. According to Lemma 4 there exist intermediate points s_k,

$$0 \leqq t_1 < s_1 < t_2 < \cdots < s_{M-1} < t_{M-1} \leqq \pi/\omega,$$

at which $\phi L[\phi] < 0$. Note, that for $m_k \geqq 2$, $L[\phi]$ has a zero of multiplicity $m_k - 2$ at t_k. The case $m_k = 1$ has been treated in [3]. Observe that $\ddot{\phi}/\phi$ and $\phi L[\phi]$ have the same sign whenever $\phi \neq 0$. Therefore, $\phi L[\phi] < 0$ at $t_k' = s_k$.

Assume first, that $m_k \geqq 2$ is not even. Then ϕ and $L[\phi]$ both change sign at t_k. Differentiating $L[\phi]$ $m_k - 2$ times shows that either both functions are increasing or both are decreasing in a neighborhood of t_k. Now, from $\phi L[\phi] < 0$ at s_{k-1} and s_k we can conclude that $L[\phi]$ has a zero in (s_{k-1}, t_k) and one in (t_k, s_k) and consequently, that it has m_k zeros in (s_{k-1}, s_k).

Secondly, assume m_k to be even. Again, $L[\phi]$ has a zero of order $m_k - 2$ at t_k. Comparing the derivatives of order $m_k - 2$ shows that the functions ϕ and $L[\phi]$ have either both a minimum or both a maximum at t_k. But these two functions are of opposite signs at s_{k-1} and s_k. Again, $L[\phi]$ must have a zero in (s_{k-1}, t_k) and one in (t_k, s_k).

Consequently, $L[\phi]$ has m_k zeros in each interval (s_{k-1}, s_k), (at least!), and $m_1 - 1$ zeros in $[t_1, s_2)$ and $m_M - 1$ zeros in $(s_{M-1}, t_M]$, that are in total $N - 2$ zeros in $[t_1, t_M]$.

COROLLARY 1. *The assertion remains valid for operators* $L_1[\phi] = \ddot{\phi} + \alpha^2\phi$ *with* $0 < \alpha \leqq \omega$.

Proof. Lemma 4 guarantees the existence of the intermediate points s_k where $\phi L_1[\phi] < 0$ holds.

COROLLARY 2. *Let* ϕ, ω, N *be as in the lemma, and* L_2 *the differential operator associated with* $p_2(\lambda) = \lambda^2 - 2\delta\lambda + \delta^2 + \alpha^2$ *with arbitrary real* δ *and* $0 < \alpha \leqq \omega$. *Then* $L_2[\phi]$ *has at least* $N - 2$ *zeros in* I.

Proof. This is an immediate consequence of $L_2[\phi] = \exp(-\delta t) L[\phi \exp(\delta t)]$, Lemma 3 and Corollary 1.

LEMMA 6. *Let a twice differentiable function* ϕ *have* N *zeros (counting multiplicities) in* $I = [a, b]$. *Let* τ *be the maximum distance between two consecutive zeros. Then,* $L[\phi] = \ddot{\phi} + \alpha^2\phi$ *has for any* $\alpha \leqq \pi/\tau$ *at least* $N - 2$ *zeros in* I.

Proof. The same arguments as in the proof of Lemma 5 will be used. Thus, it remains to show the existence of the intermediate points s_k. Lemma 4 states the existence of points t_k' at which

$$\ddot{\phi}/\phi < -(\pi/(t_{k+1} - t_k))^2 \leqq -(\pi/\tau)^2 \leqq -\alpha^2.$$

Application of Lemma 6 requires additional information, but allows the use of relatively high frequencies in the operator L.

COROLLARY. *Let the notation be as in Lemma* 6, *and again* L_2 *the operator associated with* $p_2(\lambda) = \lambda^2 - 2\delta\lambda + \delta^2 + \alpha^2$. *Then for every real* δ *and every* α *with* $0 < \alpha \leqq \pi/\tau$ *the function* $L_2[\phi]$ *has at least* $N - 2$ *zeros in* I.

It is important to note that an upper bound on the frequency of the operator L cannot be avoided. This is demonstrated by the following simple example.

Example 1. The function $\phi = \sin t - \sin 3t$ has four simple zeros in $[0, \pi]$ and seven zeros $(k\pi/4, k = 0, 1, 3, 4, 5, 7, 8)$ in $[0, 2\pi]$. Lemma 6 gives $\alpha \leqq 2$, whereas Lemma 5 only allows conclusions for frequencies $\alpha \leqq \frac{1}{2}$. But the condition $\alpha \leqq \pi/\tau$ of Lemma 6 is also only a sufficient one: the function $\ddot{\phi} + 5\phi = 4(\sin t + \sin 3t)$ has three zeros in $[0, \pi]$ and 5 zeros in $[0, 2\pi]$. In the case $\alpha > \pi/\tau$, the total length of the interval

in question might be of importance: ϕ has ten zeros in $[0, 3\pi]$, but $\ddot{\phi}+5\phi$ has only $7<10-2$.

Take now, as a slight modification, the function $\phi = \sin t - A \cdot \sin 3t$ with $A>1$, and consider the interval $I=[0, 2\pi]$. Each function of this type has exactly seven zeros in I with $\tau \to \pi/3$ as $A \to \infty$. Therefore, Lemma 6 allows for conclusions in the case $\alpha \leqq \pi/\tau \to 3$ as $A \to \infty$. Thus, the maximum frequency of ϕ seems to be of some importance. One might suppose from this that a statement like "for $\alpha \geqq \omega_{\max}$, $\psi = \ddot{\phi}+\alpha^2\phi$ has at most $N-2$ zeros in I" for functions which are solutions of an equation (1) holds. In fact one can find examples where this is true, but it is not so in general, as can be seen from the next example.

Example 2. Equation (1) with characteristic polynomial $p(\lambda)=(\lambda^2+1)^k(\lambda^2+\Omega^2)$ has $z_6=[-\Omega(3-\Omega^2)\sin t+\Omega(1-\Omega^2)t\cos t+2\sin \Omega t]/(2\Omega(1-\Omega^2)^2)$ for $k=2$ and $z_4=\ddot{z}_6+z_6=(-\Omega\sin t+\sin \Omega t)/(\Omega(1-\Omega^2))$ for $k=1$ as solution of the SIVP. The first zero of z_6 (z_4, respectively) determines $\eta_0=\eta_{0,6}$ ($\eta_{0,4}$ respectively), and it is easy to verify that

$$\eta_{0,4}=7.9604>7.775=\eta_{0,6} \quad \text{for } \Omega=0.35,$$

and

$$\eta_{0,4}=2\pi<7.1073=\eta_{0,6} \qquad \text{for } \Omega=0.5.$$

This example demonstrates further that for a fixed set of characteristic numbers, η_0 is in general not a monotone function of the order n of (1).

4. On the first conjugate point of (1). As has been pointed out, one can restrict the consideration to the characterization of the first conjugate point of $t_0=0$. Further, Proposition 1 has $\eta_0=\eta^+(0)=-\eta^-(0)$ as consequence. Here, $\eta^+(\eta^-)$ denotes the first right (left) conjugate point of 0 as defined in [1]. Therefore, the following investigations try to characterize η^+ and to give best possible bounds for it.

PROPOSITION 2. *Every differential equation of second order with nonreal characteristic roots has*

$$I_0=[0, \pi/\omega) \quad \textit{or} \quad (0, \pi/\omega] \textit{ respectively}$$

as maximum interval of disconjugacy.

Proof. Lemma 3 allows reduction to the case $\ddot{z}+\omega^2 z=0$, and from this the result follows trivially.

PROPOSITION 3. *There is no differential equation* (1) *of order* $n \geqq 3$ *that has* $I_0=[0, \pi/\omega)$ *or* $(0, \pi/\omega]$ *respectively, as maximum interval of disconjugacy.*

Proof. Only equations with at least one pair of nonreal eigenvalues need to be considered. Without loss of generality it may be assumed that $m(\lambda)=(\lambda^2+\omega^2)p(\lambda)$, where all eigenfrequencies of $p(\lambda)$ are $\leqq\omega$ and $\deg(p)\geqq 1$. Now, induction arguments are used. Assume first that for some $n>3$ there is an equation (1) having $N \geqq n$ zeros in $[0, \pi/\omega]$. The function $w:=\ddot{z}+\omega^2 z$ is a solution of the $(n-2)$th order equation $P[w]=0$. From Lemma 5 it can be deduced that w has at least $N-2 \geqq n-2$ zeros in $[0, \pi/\omega]$. Therefore, it is left to show that the assertion is true for $n=3, 4$.

For $n=3$ one can restrict to functions $z(t)=A\sin(\omega t+\gamma)+\exp(\delta t)$. Then, $\eta_0>\pi/\omega$ follows immediately from the positivity of $\exp(\delta t)$ for real δ. For $n=4$ a careful investigation of the various possibilities shows again that there is no solution of (1) with four zeros in $[0, \pi/\omega]$. The results of these simple but somewhat lengthy investigations are given in Fig. 1 and Fig. 2. One obtains that η_0 approaches π/ω only in case the real parts of both characteristic roots of p approach $\pm\infty$.

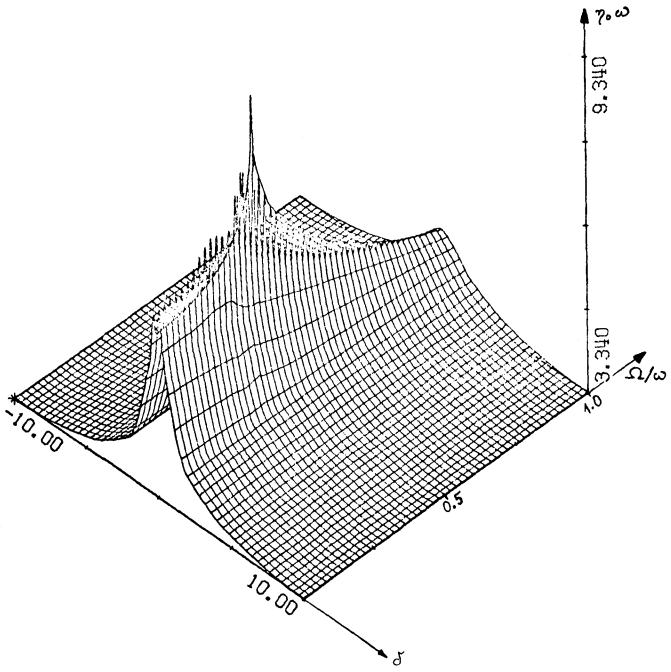


FIG. 1. η_0 *as a function of* Ω *and* δ *for equations* (1) *with characteristic polynomial* $p(\lambda) = (\lambda^2+\omega^2)(\lambda^2-2\delta\lambda+\delta^2+\Omega^2)$.

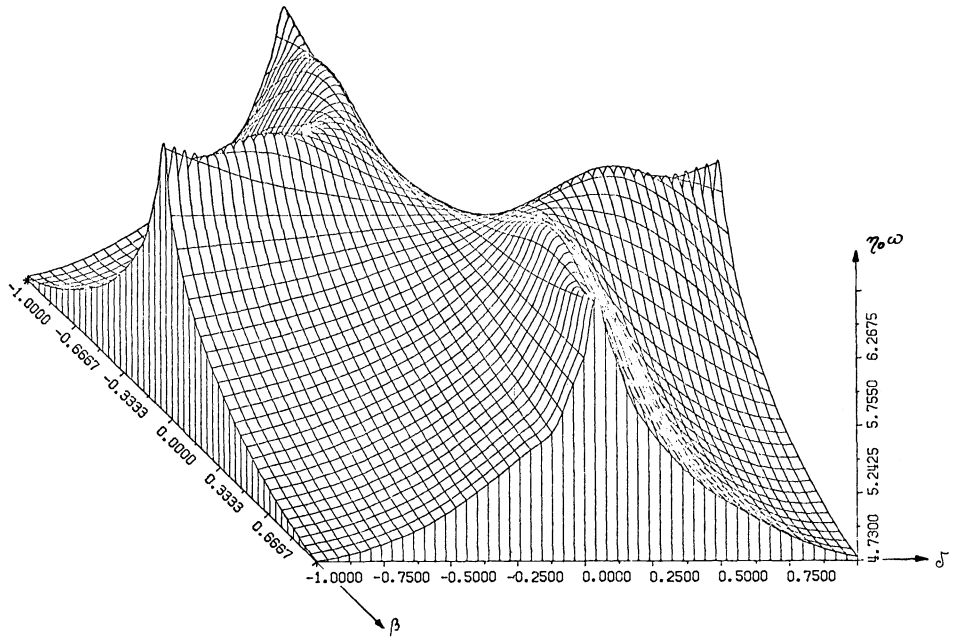


FIG. 2. η_0 *as a function of* β *and* δ *for equations* (1) *with characteristic polynomial* $p(\lambda) = (\lambda^2+\omega^2)(\lambda-\beta)(\lambda-\delta)$.

A rather trivial reformulation of this result is the following.

COROLLARY. $\eta_0 = \pi/\omega$ *iff* $n = 2$.

PROPOSITION 4. *For every* $n > 2$ *and every* $\varepsilon > 0$ *sufficiently small, there is an equation* (1) *with* $\eta_0 = \pi/\omega + \varepsilon$.

Proof. Lemma 3 together with a transformation of time $t' = \omega t$ allows reduction to systems (1) with characteristic polynomial

$$p_n(\lambda) = (\lambda^2 + 1)\hat{p}_n(\lambda).$$

Let P_n be the differential operator associated with $p_n(\cdot)$, where

$$\hat{p}_n(\lambda) = (\lambda - \alpha)(\lambda - 2\alpha)\cdots(\lambda - n\alpha), \qquad n = 1, 2, \cdots, \quad \alpha \text{ real.}$$

Then we have for the solution z_n of the SIVP

$$P_n[z] = 0, \quad z^{(k)}(0) = \delta_{k,n+1}, \quad k = 0, 1, \cdots, n+1,$$

the representation

$$z_n(t) = A_n \cos t + B_n \sin t + \sum_{i=1}^{n} a_{ni} \exp(i\alpha t), \qquad n = 1, 2, \cdots.$$

Now observe that z_n is also a solution of

$$\dot{z} - n\alpha z = z_{n-1}, \quad z(0) = 0, \quad n = 1, 2, \cdots,$$

with $z_0 = \sin t$. This allows us to derive the following recursion formulas for the coefficients of z_n:

$$\begin{aligned} A_n &= -(n\alpha A_{n-1} + B_{n-1})/(1 + n^2\alpha^2), \\ B_n &= (A_{n-1} - n\alpha B_{n-1})/(1 + n^2\alpha^2), \\ a_{ni} &= -a_{n-1,i}/((n-i)\alpha) \qquad i = 1, \cdots, n-1, \\ a_{nn} &= -A_n - \sum_{i=1}^{n-1} a_{ni}, \end{aligned} \qquad n = 2, 3, \cdots,$$

with starting values

$$A_1 = -1/(1 + \alpha^2), \quad B_1 = \alpha A_1, \quad a_{11} = -A_1.$$

Now, rewrite z_n as

$$z_n = (A_n^2 + B_n^2)^{1/2} \sin(t + \beta_n) - q_n(\exp(\alpha t)). \tag{6}$$

Here, $q_n(\cdot)$ is a polynomial of degree n, and the phase β_n is determined by

$$\tan \beta_n = A_n/B_n, \qquad |\beta_n| \leqq \pi/2.$$

Let $\alpha < 0$. Then we have

$$\beta_n < 0 \quad \text{and} \quad \lim_{\alpha \to -\infty} \beta_n(\alpha) = 0. \tag{7}$$

This is shown as follows. For $|\alpha|$ sufficiently large we have $B_n > 0$ for all $n = 1, 2, \cdots$. Then by induction arguments (O is the Landau symbol)

$$\begin{aligned} A_n &= O(|\alpha|^{-(n+1)}) \\ B_n &= O(|\alpha|^{-n}) \end{aligned} \qquad \text{for } |\alpha| \to \infty, \quad n = 1, 2, \cdots,$$

as well as

$$\operatorname{sgn}(A_n/B_n) = \operatorname{sgn}(\alpha) = -1, \qquad n = 1, 2, \cdots,$$

can be deduced easily. Consequently,

$$\tan \beta_n = O(|\alpha|^{-1}) \quad \text{for } |\alpha| \to \infty$$

holds. This finishes the proof of (7). From this follows further that the first expression in (6) is positive for $t < \pi - \beta_n$ and negative for $t > \pi - \beta_n > \pi$ and $\beta_n < 0$ arbitrarily close to 0 for α sufficiently large.

The second term q_n in (6) is a polynomial with argument $\exp(\alpha t)$. Now take $t \geqq \pi$ and $\alpha < 0$ with $|\alpha|$ sufficiently large. Then, again by induction arguments, it can be deduced that $\operatorname{sgn}(-q_n) = \operatorname{sgn}(a_{n1}) = 1$. From this $q_n < 0$, follows. Further, a_{n1} is bounded (in fact $a_{n1} = O(|\alpha|^{-(n+1)})$), and from this

$$q_n/(z_n + q_n) \to 0 \quad \text{for } t \neq \pi - \beta_n \text{ and } \alpha \to -\infty$$

follows. Therefore, $z_n + q_n$ and q_n must have a point of intersection $T = T(\alpha) > \pi - \beta_n$ with $T(\alpha) \to \pi - \beta_n(\alpha) \to \pi$ for $\alpha \to -\infty$, according to (7). This terminates the proof, the idea of which is illustrated in Fig. 3.

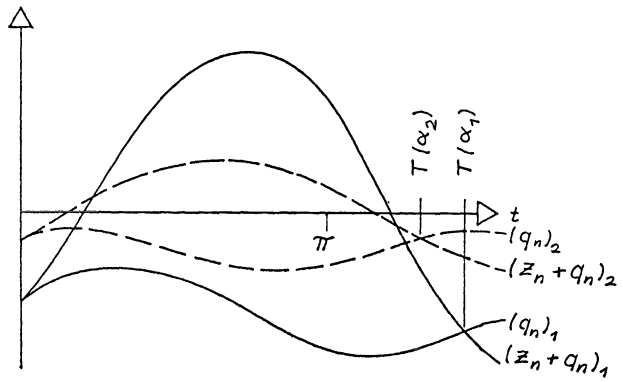


FIG. 3. *The geometric idea of the proof of Proposition* 4. *Full line*: $\alpha = \alpha_1$; *dashed line*: $\alpha = \alpha_2$; *with* $\alpha_2 < \alpha_1 < 0$.

We remark that the functions used in this proof are not the only possible ones. For example, equations with a set of characteristic roots equal to $\{k\alpha \pm \sqrt{-1}\,|k = 0, \cdots, m\} = S$ for $n = 2m$ and $S \cup \{0\}$ for $n = 2m+1$ can be used in a similar manner.

Propositions 3 and 4 prepare the results on a lower bound for η_0, whereas Proposition 5 guarantees that for every order n there is a differential equation such that η_0 equals the well-known upper bound $(n-1)\pi/\omega$. In order to relieve the proof of unessential but necessary calculations, a useful result on trigonometric functions will be given first.

LEMMA 7. (a) *Let* $z(t) = \mathcal{L}\{\sin t, \sin 3t, \cdots, \sin(2k-1)t\}$, *where* $\mathcal{L}$ *stands for "linear combination of". Then the Wronskian of order* m, W_m, *is again a linear combination of trigonometric functions. The frequencies are either all even* (m *even*) *or all odd* (m *odd*) *and bounded by* $m(2k-m)$.

(b) *Let* $z(t) = \mathcal{L}\{1, \cos 2t, \cdots, \cos 2kt\}$. *Then* W_m *is a linear combination of trigonometric functions. The frequencies are either all even* (m *odd*) *or all odd* (m *even*) *and bounded by* $m(2k+1-m)$.

Proof. The proof uses induction arguments and is based on the elementary formulas for the addition of trigonometric functions. It is given for the first assertion and

m even only. The Wronskian W_m, with $S := \{1, 3, \cdots, 2k-1\}$, is of the form

$$W_m = \mathscr{L}\{\sin \omega_1 t \cdots \sin \omega_m t | \omega_i \in S\} = \mathscr{L}\{\cos(\omega_1 \pm \omega_2 \pm \cdots \pm \omega_m)t | \omega_i \in S\}.$$

From this the first part of the assertion is evident. It is left to show that the maximum frequency $\omega_{k,m}$ of W_m has the given bound. For $k = 1$ this is trivially true. The solution of the SIVP with characteristic polynomial $p(\lambda) = (\lambda^2+1)(\lambda^2+3^2)\cdots(\lambda^2+(2k+1)^2)$ can be written as $z_{2k+2} = x + a \sin(2k+1)t =: x + u$, where x is a solution of the same differential equation of order $2k$ as z_{2k}, whereas u is a solution of $\ddot{u} + (2k+1)^2 u = 0$. The Wronskian $W_m(z_{2k+2})$ can be represented as sum of Wronskians of order m:

$$W_m(z_{2k+2}) = W_m(x+u) = W_m(x) + W(x, \dot{x}, \cdots, x^{(m-2)}, u^{(m-1)}) + \cdots + W_m(u).$$

Here, only those determinants do not vanish identically which have at most two rows containing u or derivatives of it and at most $2k$ rows built with the function x. We consider three cases.

(i) Let W have two rows built with $u^{(s)}$ and $u^{(r)}$. From Laplace's theorem it follows that W is a sum of terms

$$W(x^{(s_1)}, \cdots, x^{(s_{m-2})}) \cdot \begin{vmatrix} u^{(s)} & u^{(r)} \\ u^{(s+h)} & u^{(r+h)} \end{vmatrix}, \qquad 1 \leqq h \leqq m-1.$$

The first determinant is a linear combination of trigonometric functions with frequencies $\leqq \omega_{k,m-2}$, whereas the second factor is a constant.

(ii) In case W contains only one row built from u, W is a sum of terms with frequencies $\leqq(\omega_{k,m-1} + 2k + 1)$, where $m \leqq 2k$ holds also.

(iii) If W contains no row built from u, frequencies $\leqq \omega_{k,m}$ may appear with $m \leqq 2k$.

Putting the pieces together now results in

$$\begin{aligned} \omega_{k+1,m} &\leqq \max\{\omega_{k,m-2}, \omega_{km}, \omega_{k,m-1} + 2k + 1\} \\ &= \max\{(m-2)(2k-m+2), m(2k-m), (m-1)(2k-m+1)+2k+1\} \\ &= \max\{(m(2k-m), m(2k+2-m)\} = m(2(k+1)-m), \end{aligned}$$

and this terminates the proof.

PROPOSITION 5. *For every $n \geqq 2$ there is a differential equation with $\eta_0 = (n-1)\pi/\omega$.*

Proof. Consider (1) with characteristic polynomial

$$\begin{aligned} p_n(\lambda) &= (\lambda^2+1)(\lambda^2+3^2)\cdots(\lambda^2+(2k-1)^2) \quad \text{for } n = 2k, \\ p_n(\lambda) &= (\lambda^2+2^2)(\lambda^2+4^2)\cdots(\lambda^2+(2k)^2) \quad \text{for } n = 2k+1. \end{aligned}$$

Proposition 1, requires that none of the Wronskians $W_m(z_n)$, $1 \leqq m \leqq [n/2]$, vanish for some t with $0 < t < \pi = (n-1)\pi/\omega$. The solution z_n of the SIVP is a linear combination of trigonometric functions of the form

$$z_n(t) = \begin{cases} \sum_{i=1}^{k} a_i \sin(2i-1)t, & n = 2k, \\ \sum_{i=0}^{k} b_i \cos 2it & n = 2k+1, \end{cases}$$

where the coefficients a_i and b_i can be characterized as quotients of certain Vandermonde determinants. By definition, $z_n(t)$ has a zero of order $n-1$ at $t = 0$, and because

of $\sin 0 = \sin j\pi = 0 (j \in \mathbb{N})$ and $\cos 0 = \cos 2j\pi = 1$, it has a zero of the same order at $t = \pi$.

Now by Lemma 7 every Wronskian W_m is a linear combination of trigonometric functions with frequencies

$$m(n-m), m(n-m)-2, \cdots, (1-(-1)^m)/2, \qquad m = 1, 2, \cdots, n-1,$$

with only sine or cosine terms appearing. The derivative of order j can be written as

$$W_m^{(j)} = \mathscr{L}\{W(z_n^{(s_1)}, \cdots, z_n^{(s_m)})\} \quad \text{with } s_1 + \cdots + s_m = j + m(m-1)/2,$$

and can differ from zero only when at least one Wronskian in this sum differs from zero. At $t = 0$ or $t = \pi$, this is for the first time true for $s_m = s_{m-1} + 1 = \cdots = s_1 + m - 1 = n - 1$. Consequently, $j = m(n-m)$.

By a well-known theorem, see, e.g., [4], no linear combination of the functions $1, \cos t, \cdots, \cos Mt, \sin t, \cdots, \sin Mt$ has more than $2M$ zeros in $[0, 2\pi)$, because they build a complete Chebyshev system over this interval. In our terminology this is equivalent to "Equation (1) with characteristic polynomial $p(\lambda) = (\lambda^2+1)(\lambda^2+2^2)\cdots(\lambda^2+M^2)$ is disconjugate over $[0, 2\pi)$ for every natural number M". W_m is a function of this type with $M = m(n-m)$ which has M zeros at $t = 0$ and M further zeros at $t = \pi$. Therefore, W_m cannot vanish at any point of $(0, \pi)$. This completes the proof.

PROPOSITION 6. *Let* (1) *depend continuously on a real parameter* s. *Then* $\eta_0 = \eta_0(s)$ *is a continuous function of* s.

Let us remark, first, that this result is already known [9] and is valid even for nonautonomous differential equations. Nevertheless, we shall present here a proof which is probably more simple than the one given in [9]. It is pointed out that it can be used for the general case of [9] in a straightforward way. The proof is based again on the strong connection between disconjugacy properties and the concept of Chebyshev system.

Proof. From standard theorems about differential equations, e.g., [5], it follows that not only z, but also the zeros of z or one of the Wronskians $W_m(z)$ are continuous functions of s, as long as there is a real zero. But it might happen that for some $\hat{\eta}_0 = \eta_0(\hat{s})$ all Wronskians having a zero at $\hat{\eta}_0$ have a zero of even order there. Consequently, the possibility that all these zeros disappear simultaneously has to be taken into account. Therefore, we have to show that no pair $\hat{s}$, $\hat{\eta}_0 = \eta_0(\hat{s})$ exists such that all Wronskians which vanish at $\hat{P} := (\hat{s}, \hat{\eta}_0)$ are sign-definite in a neighbourhood U of $\hat{P}$. Assume now that W_m and W_{m+1} both have a zero of even order at $\hat{P}$, and that $W_{m+1} \neq 0$ at $\hat{P}$ (remark that in case W_m has a zero of order >1 W_{m+1} must also vanish at this point), and

(i) $\dot{W}_m$ has in $\hat{P}$ a zero of odd order, and therefore a real number α can be found such that

$$w(\hat{s}, t) := W_m(\hat{s}, t) + \alpha \dot{W}_m(\hat{s}, t)$$

has a zero for $t = \tau$ with $(\hat{s}, \tau) \in U$ (in fact τ can be chosen arbitrarily close to $\hat{\eta}_0$). Now, from $w(\hat{s}, \tau) = w(\hat{s}, \hat{\eta}_0) = 0$ and continuity of $w(\cdot, \cdot)$, it follows that numbers $\tau_1, \tau_2, \tilde{s}$ can be found such that $(\tilde{s}, \tau_i) \in U$ for $i = 1, 2$ and $w(\tilde{s}, \tau_1) = w(\tilde{s}, \tau_2) = 0$.

(ii) On the other hand, the assumption that W_{m+1} is definite in U is equivalent to the definiteness of $W_m \ddot{W}_m - \dot{W}_m^2$, since [7, VII, Problem 59] $W_m \ddot{W}_m - \dot{W}_m^2 = W_{m-1} W_{m+1}$. Consequently, W_m and the Wronskian $W_2(W_m) = W(W_m, \dot{W}_m)$ are both definite in U. This means that every function $v(s, t) := \alpha_1 W_m(s, t) + \alpha_2 \dot{W}_m(s, t)$ has at most one zero in U for any $s \neq \hat{s}$. But from this follows that $w(\tilde{s}, t)$ as defined in (i) can have at most one zero in U. This contradiction terminates the proof.

Finally, we present the proof of Theorem 3.

Proof of Theorem 3. The representation (5) of the characteristic polynomial allows us to write the solution to be investigated as [5]

$$z(t) = x(t) + y(t), \quad \text{with } P[y] = 0 \text{ and } Q[x] = 0.$$

Here, P and Q are the differential operators corresponding to $p(\cdot)$ and $q(\cdot)$. From this it follows that x can be written as

$$x(t) = \sum_{i=1}^{j} p_i(t) \exp(\lambda_i t),$$

where j is the number of distinct zeros of $q(\cdot)$, and $p_i(t)$ are polynomials of degree r_i (corresponding to the multiplicities of the real zeros λ_i) with $r_1 + \cdots + r_j + j = r$. By Lemma 3 $w_1(t) = \exp(-\lambda_1 t) z(t)$ has the same zeros (and of the same multiplicity!) as $z(t)$. Consequently, if $z(t)$ has N zeros counting multiplicities on I, then the (r_1+1)th derivative of $w_1(t)$ has at least $N - r_1 - 1$ zeros on I. Multiplying again by $\exp(-\lambda_2 t)$ and differentiating yields finally that a function of the form

$$\exp(-\lambda_1 t) \cdots \exp(-\lambda_j t)(b_0 y(t) + \cdots + b_r y^{(r)}(t))$$

has at least $(N - r)$ zeros on I. From this follows that

$$w(t) = b_0 y(t) + \cdots + b_r y^{(r)}(t),$$

which is again a solution of $P[y] = 0$, has at least $(N - r)$ zeros on I. And from Theorem 1, $w(t)$ has at most $[T/\eta_{01}]$ zeros. Therefore,

$$N - r \leqq (n - r - 1)[T/\eta_{01}]^*$$

which yields the desired estimate.

REFERENCES

1. W. A. COPPEL, *Disconjugacy*, Springer-Verlag, Berlin, 1971.
2. O. HÁJEK, *Terminal manifolds and switching locus*, Math. Systems Theory, 6 (1973), pp. 289–301.
3. O. HÁJEK, *On the number of roots of exp-trig polynomials*, Computing, 18 (1977), pp. 177–183.
4. S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York, 1966.
5. H. W. KNOBLOCH AND F. KAPPEL, *Gewöhnliche Differentialgleichungen*, Teubner-Verlag, Stuttgart, 1974.
6. J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations V, Princeton University Press, Princeton, NJ, 1960.
7. G. PÓLYA AND G. SZEGÖ, *Aufgaben und Lehrsätze der Analysis*, 2. Band, Springer-Verlag, Berlin, 1971.
8. L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *Mathematische Theorie optimaler Prozesse*, Oldenbourg, Munich, 1967.
9. T. L. SHERMAN, *Conjugate points and simple zeros for ordinary linear differential equations*, Trans. Amer. Math. Soc., 146 (1969), pp. 397–411.
10. I. TROCH, *Bemerkungen zur n-Beobachtbarkeit und n-Steuerbarkeit*, Z. Angew. Math. Mech., 51 (1971), pp. 255–264.
11. ———, *Über die Lösungen linearer autonomer Differentialgleichungssysteme als Tschebyscheff-System*, Z. Angew. Math. Mech., 52 (1972), pp. 193–194.
12. ———, *Über die Bedeutung von Beobachtbarkeits-und Steuerbarkeitsindex*, Z. Angew. Math. Mech., 51 (1971), pp. 265–269.
13. ———, *Sampling with arbitrary choice of the sampling instants*, Automatica, 9 (1973), pp. 117–124.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0009 $01.00/0

A QUADRATIC TRANSFORMATION TECHNIQUE FOR THE SOLUTION OF NONLINEAR SYSTEMS*

ROBERT A. HOWLAND, JR.†

Abstract. A new technique is developed for the formal analytical solution of systems of differential equations through an easily invertible Lie transform method. It is based on an approach due to Kolmogorov and Arnol'd and features the ability to determine the transformed equation to twice the order of explicit transformation, particularly useful in the solution of nonlinear systems. Van der Pol's equation is treated as an example in which a single transformation results in a second-order explicit periodic solution with frequencies accurate to *fourth* order.

1. Introduction. It is a fact which must be lived with that few differential equations can be solved exactly, and fewer still which mathematically model, even approximately, physically meaningful systems. On the other hand, many systems can be described through a small function perturbing an otherwise soluble system, and there are various techniques used to obtain at least formal solutions to such problems, as power series in the magnitude of the perturbation. Some deal with the original system (iterative successive approximation techniques, for example), others with a transformed (and hopefully soluble) system in new variables; in the latter case, the original variables' solution can then be found through the inverse transformation from the soluble system's variables. To be useful, however, the transformations involved must a) give a predictable form for the transformed differential equation, b) be manageable, and c) be easily inverted.

Hamiltonian systems exemplify this latter approach. Their form under canonical transformations is invariant, and the sufficiency condition for such transformations is sufficiently general to be practical, satisfying criteria a) and b); unfortunately, the traditional transformations involve expressions in mixed (transformed and untransformed) variables, making inversion difficult. In the past decade, however, a new easily invertible canonical transformation, the *Lie transform* (expressed in *homogeneous* [all old or all new] variables) has met with some success in methods by Hori [1], Deprit [2], Kamel [3], and Henrard [4], [5], and has been generalized to non-Hamiltonian systems by the latter two [4], [6] more for its pleasing inversion properties than for its canonicity.

The solutions obtained by these methods, each of which determines sequentially orders of a generating function transforming the original system and thus may be termed *linearly* transforming, are purely formal *asymptotic series*; mathematical convergence is generally not only unable to be demonstrated, but may even be violated (see, for example, Siegel's results regarding Hamiltonian systems [7, § 3(a)], although a theorem by Poincaré [8, VIII, § 3] justifies their use over short intervals of time. Yet another approach suggested by Kolmogorov [9] and generalized and proven convergent by Arnol'd [10] for the solution of Hamiltonian systems *does* give convergent series solutions under certain circumstances; the technique relies on determining a sequence of *complete* transformations, and results in *quadratic* transformation and convergence. As originally presented, their approach utilized the traditional mixed-variable generating function, making it especially impractical for application, but a recent method proposed by the author [11], and subsequently modified in the interest of greater efficiency [12] utilizing Lie transforms makes their approach a viable technique for the solution of Hamiltonian systems. (Unlike previous linear perturbation methods,

* Received by the editors October 2, 1977 and in final revised form October 16, 1978.

† Department of Mechanical Engineering, Rose-Hulman Institute of Technology, Terre Haute, Indiana, 47803.

this technique relies on a demonstrably convergent mathematical approach and might thus have some appeal. Convergence, however, requires imposition of the *irrationality condition* [10, V, § 3.2] which, in practical problems, is impossible to verify as discussed by the author [11, § 1.1]; no claim, then, can be made regarding convergence of these solutions.)

Not all systems can be represented by a classical Hamiltonian: dissipative systems, for example, involve an additional nonconservative work term in the momentum equations of motion whose mere transformation does not simplify. Kamel [13] has proposed a formal Hamiltonization procedure applicable even to dissipative systems, but it might be deemed preferable to work with the original non-Hamiltonian systems. In such cases, the non-Hamiltonian formulations [4], [6] could be utilized, but, like their Hamiltonian counterparts, these methods are only linearly transforming. (The formal Hamiltonization procedure and most of the linear techniques, though none of the quadratic schemes, cited above are described in Nayfeh's survey of perturbation methods [14, § 5.7].)

The present paper, then, generalizes the author's Hamiltonian quadratically transforming technique, resulting in an algorithm for the formal asymptotic solution of such non-Hamiltonian systems. It might be expected to be more efficient in many cases, but in any event has a special feature which might recommend its use in some applications: the ability to obtain nonlinearly perturbed frequencies of vibration to exceptionally high accuracy. Basic theory of the vector generalization is developed in the next section; application of the method, an example (van der Pol's equation), and comparisons with linear counterparts are made in subsequent ones.

2. Basic theory. Consider a vector function $f(x; \kappa, \varepsilon)$ of the vector variable x and the two scalar parameters κ and ε, analytic in x at $\kappa = 0$ and $\varepsilon = 0$. (With the exception of exponents and indices, Latin letters throughout this paper will denote n-vectors unless otherwise stated.) Then, for the [vector] C^∞ function $W(x; \varepsilon)$, define the formal *Lie operator*

$$L_W f \equiv \left(\frac{\partial f}{\partial x}\right) \cdot W \tag{2.1}$$

(where $(\partial f/\partial x)$ in this matrix product is the Jacobian of f with respect to x, and W is represented as a column vector), and its iterate

$$L_W^i f \equiv L_W(L_W^{i-1} f), \qquad L_W^0 f \equiv f.$$

Now defining the *Lie transform of f under W*,

$$\exp(\kappa L_W) f \equiv \sum_{i \geqq 0} \frac{\kappa^i}{i!} L_W^i f = f + \kappa L_W f + \frac{\kappa^2}{2!} L_W^2 f + \cdots,$$

we can make the formal transformation $x \leftrightarrow (y; \kappa, \varepsilon)$

$$x \equiv \exp(\kappa L_W) y \tag{I}^1$$

where now $W = W(y; \varepsilon)$. (Note that if $W = JS_{(y,Y)}$, where J is the symplectic matrix and $S = S(y, Y; \varepsilon)$ a *scalar* function, this transformation reduces to the canonical transformation (I) in the author's treatment of Hamiltonian systems [11]).

The following theorems can be proven about the transformation (I):

[1] Equations fundamental to the technique are identified with Roman numerals.

THEOREM (Transformation). *Under* (I),

$$f^*(y;\kappa,\varepsilon)\equiv f(\exp(\kappa L_{W(y;\varepsilon)})y);\varepsilon)=\exp(\kappa L_{W(y;\varepsilon)})f(y;\varepsilon)). \tag{II}$$

Proof. Expanding $f(\exp(\kappa L_W y);\varepsilon)$ about $\kappa=0$, we get

$$f(x(y;\kappa,\varepsilon);\varepsilon)\equiv f^*(y;\kappa,\varepsilon)=\sum_{i\geq 0}\frac{\kappa^i}{i!}\frac{d^i f^*}{d\kappa^i}\bigg|_{\kappa=0}.$$

But

$$\frac{df^*}{d\kappa}=\left(\frac{\partial f}{\partial x}\right)\cdot\frac{dx}{d\kappa}\bigg|_{x=x(y;\kappa,\varepsilon)},$$

and $L_W y$ is a function of y and ε only. Thus, by (I),

$$\frac{dx}{d\kappa}=\sum_{i\geq 1}\frac{\kappa^{i-1}}{(i-1)!}L_W^i y=L_W\sum_{j\geq 0}\frac{\kappa^j}{j!}L_W^j y=L_W x\equiv\left(\frac{\partial x}{\partial y}\right)\cdot W, \tag{2.2}$$

so direct calculation gives

$$\frac{df^*}{d\kappa}=\left(\frac{\partial f}{\partial x}\right)\cdot\left(\left(\frac{\partial x}{\partial y}\right)\cdot W\right)=\left(\left(\frac{\partial f}{\partial x}\right)\cdot\left(\frac{\partial x}{\partial y}\right)\right)\cdot W=\frac{\partial f}{\partial y}\cdot W\equiv L_W f,$$

and, inductively,

$$\frac{d^i f^*}{d\kappa^i}=L_W^i f(x(y;\kappa,\varepsilon);\varepsilon).$$

But, then, from the form of the explicit transformation (I),

$$\frac{d^i f^*}{d\kappa^i}\bigg|_{\kappa=0}=L_W^i f(x(y;0,\varepsilon);\varepsilon)=L_W^i f(y;\varepsilon),$$

from which the result follows.

THEOREM (Inverse). *The inverse of transformation* (I) *is generated by* $(-W)$; i.e.,

$$f^*(y(x;\kappa,\varepsilon);\kappa,\varepsilon)=\exp(\kappa L_{-W(x;\varepsilon)})f^*(x;\kappa,\varepsilon)). \tag{III}$$

The proof is the same as ones by Deprit [2, § 2] and the author [11, § 2] for this property in their Hamiltonian developments.

The basic theory presented here compares directly Deprit's development of his linear technique for Hamiltonian systems [2, § 2] in which the generating function is independent of the small parameter in terms of which expansions are made. In common with the other linear Lie transform techniques cited in § 1 of this article, however, he chooses to develop transformations (as well as the generating function) in powers of the original small parameter, ε, necessitating additional differentiation with respect to this parameter, and expansions are generally implemented through the use of auxiliary functions. In these methods, then, the coordinate transformation (I) assumes the form

$$x(y;\varepsilon)=\exp(\varepsilon L_{W(y;\varepsilon)})y),$$

so that x becomes the solution of the differential equation ([6, § 2])

$$\frac{dx}{d\varepsilon}=W(x;\varepsilon);$$

indeed, this equation is the starting point for Henrard's developments of both his Hamiltonian and non-Hamiltonian methods. In the present approach, expansions at

each transformation "step" are made in terms of a second parameter κ, functionally independent of ε, introduced to measure the magnitude of the perturbation at that step; the resulting transformation (I) can then be interpreted (see (3.1.5) below) as the solution to

$$\frac{dx}{d\kappa} = W(x; \varepsilon).$$

The functional independence of the right-hand side of this equation of κ effects considerable simplification.

3. Application of the transformation to differential equations.

3.1. Transformation of differential equations under (I). Consider a system of differential equations in vector form,

$$\dot{x} = f(x; \kappa, \varepsilon) \equiv \sum_{j \geqq 0} \kappa^j f_j(x; \varepsilon). \tag{3.1.1}$$

Under any transformation $x \to (y; \kappa, \varepsilon)$,

$$\dot{x}(y; \kappa, \varepsilon) = \frac{\partial x(y; \kappa, \varepsilon)}{\partial y} \cdot \dot{y} = f(x(y; \kappa, \varepsilon); \kappa, \varepsilon),$$

so that

$$\dot{y} = \left(\frac{\partial x(y; \kappa, \varepsilon)}{\partial y}\right)^{-1} \cdot f(x(y; \kappa, \varepsilon); \kappa, \varepsilon) \equiv g(y; \kappa, \varepsilon),$$

or, for $f^*(y; \kappa, \varepsilon) \equiv f(x(y; \kappa, \varepsilon); \kappa, \varepsilon)$,

$$\frac{\partial x}{\partial y} \cdot g = f^*, \tag{3.1.2}$$

in which all functions are expressed in $(y; \kappa, \varepsilon)$.

Under the particular transformation (I), write g (after Henrard [4])

$$g(y; \kappa, \varepsilon) \equiv \sum_{i \geqq 0} \kappa^i g_i(y; \varepsilon) \equiv (\mathscr{L}(W)f)(y; \kappa, \varepsilon). \tag{3.1.3}$$

Then, by (I),

$$\frac{\partial x}{\partial y} = \sum_{j \geqq 0} \frac{\kappa^j}{j!} \frac{\partial}{\partial y}(L_W^j y),$$

while, by (II),

$$f^* = \exp(\kappa L_W f) = \sum_{i \geqq 0} \frac{\kappa^i}{i!} L_W^i \sum_{j \geqq 0} \kappa^j f_j,$$

so (3.1.2) becomes

$$\sum_{k \geqq 0} \kappa^k \sum_{j=0}^{k} \frac{1}{(k-j)!} \left(\frac{\partial}{\partial y}(L_W^{k-j} y)\right) \cdot g_j = \sum_{k \geqq 0} \kappa^k \sum_{j=0}^{k} \frac{1}{(k-j)!} L_W^{k-j} f_j.$$

Thus, equating orders of κ, using (2.1), and noting that $L_W y = W$, we obtain finally for

the g_i in (3.1.3):

(IV)
$$k=0: \quad \frac{\partial}{\partial y}(L_W^0 y)\cdot g_0 = g_0 = f_0,$$
$$k\geqq 1: \quad g_k = f_k + \sum_{j=0}^{k-1} \frac{1}{(k-j)!}(L_W^{k-j} f_j - L_{g_j}(L_W^{k-j-1} W))$$

Observe that, unlike the other non-Hamiltonian Lie transform techniques [4], [6], the recursive relation for $k\geqq 1$ in (IV) does not involve the calculation of intermediate functions resulting from recursive applications of the Lie operator, subsequently to be operated upon again by Lie operators to determine the transformed differential equation (cf., e.g., Henrard's Algorithms II and IV [4], which also describe Kamel's technique). Required only are the iterates $L_W^i f_j$ and $L_W^i W$, the latter of which it is advisable to retain for later applications of L_{g_j}. It should also be kept in mind that W, f_i, and g_i are, in practice, all (vector) power series in ε.

Note that as a special case of (IV),

$$(\mathscr{L}(W)W)(y;\varepsilon) = W(y;\varepsilon); \tag{3.1.4}$$

indeed, for $f\equiv W$ in (3.1.1), $f_0 = W$ while $f_j \equiv 0$ for $j\geqq 1$, and inductively $g_k \equiv 0$ for $k\geqq 1$. Thus, using (3.1.3) to write (3.1.2) in the form

$$\frac{\partial x(y;\kappa_0,\varepsilon)}{\partial y}\cdot(\mathscr{L}(W)f)(y;\kappa_0,\varepsilon) = f(x(y;\kappa_0,\varepsilon);\varepsilon)$$

and using the result (3.1.4), we get

$$\frac{\partial x(y;\kappa,\varepsilon)}{\partial y}\cdot(\mathscr{L}(W)W)(y;\varepsilon) = \frac{\partial x(y;\kappa,\varepsilon)}{\partial y}\cdot W(y;\varepsilon) = W(x(y;\kappa_0,\varepsilon);\varepsilon).$$

But then, by (2.2), $x(y;\kappa,\varepsilon)$ under (I) satisfies

$$\frac{dx}{d\kappa} = \frac{\partial x}{\partial y}\cdot W = W(x;\varepsilon). \tag{3.1.5}$$

Such an equation is the starting point for Kamel's and Henrard's developments [3], [6] leading to their transformation algorithms; in the present case the transformation (I) itself has been. But having established (3.1.5) it is possible to obtain an initially appealing result regarding the composition of transformations (I) analogous to one of Henrard's ([4, § 4]):

THEOREM (Composition). *The mapping*

$$x = \exp(\kappa_0 L_{W_0} z),$$

giving the composition of the two successive mappings

$$x = \exp(\kappa_0 L_{W_1} y), \qquad y = \exp(\kappa_1 L_{W_2} z),$$

generated by the two C^∞ vector functions $W_1(y;\varepsilon)$ and $W_2(z;\varepsilon)$ for $\kappa_1 = \kappa_1(\kappa_0)$, is generated by

$$W_0(z;\kappa_0,\varepsilon) = W_1(z;\varepsilon) + \frac{d\kappa_1(\kappa_0)}{d\kappa_0}(\mathscr{L}^{-1}(W_1)W_2)(z;\varepsilon).$$

Proof. The proof follows exactly as Henrard's, showing through direct differentiation that $x(y(z;\kappa_1(\kappa_0),\varepsilon);\kappa_0,\varepsilon)$ satisfies (3.1.5),

$$\frac{dx}{d\kappa_0} = W_0(x;\kappa_0,\varepsilon),$$

for the W_0 above.

Since the application given in the next section relies on successive transformation of the original differentiation equation, this theorem would appear to be useful in recovering the original variables' solution from that of the transformed variables; but implementation of this result would require additional transformation of W_2 under the inverse of that generated by W_1 (transformation repeated for successive transformations) and more importantly, development of the inverse operator and generalization of the entire transformation algorithm to allow dependence of the generating function on the perturbation parameter κ. A similar occurrence in the author's Hamiltonian scheme [12, § 3] led to investigation of the advisability of such generalization (which ultimately would reduce to adapting one of the available linear schemes in which generating functions depend on the small parameter in terms of which expansions are made) merely for use in solution inversion. Consideration of the number of Lie operator applications required for a composite generating function of requisite length suggested that more effort would be required to apply the composite to the variables than to use the separate generating functions of shorter length successively—a result ignoring the additional transformations of the generating function itself ([12, § 5]).

3.2. Application of the basic method. Consider a vector system of differential equations

$$\dot{x} = f_0(x;\varepsilon) + \varepsilon^m f_p(x;\varepsilon), \tag{3.2.1}$$

in which f_0 and f_p are assumed of order unity and equation $\dot{x} = f_0$ is soluble, though the full system is not. If, furthermore

$$f_p = \bar{f}_1 + \tilde{f}_1 + \varepsilon^m F_2,$$

in which $\dot{x} = f_0(x;\varepsilon) + \varepsilon^m \bar{f}_1(x;\varepsilon) \equiv f_0^*(x;\varepsilon)$ is soluble, the $\varepsilon^m \tilde{f}_1$ can be said to produce a "perturbation of order m [in ε]" on f_0^*. One endeavors to transform $x \to (y;\kappa,\varepsilon)$ to eliminate the part $\tilde{f}_1$.

Since the perturbation is of the given magnitude, this suggests introducing a new "perturbation parameter" $\kappa = \varepsilon^m$ and writing F_2 in the form (3.1.1):

$$\varepsilon^{2m} F_2 \equiv \sum_{j\geqq 2} \varepsilon^{mj} \sum_{k=0}^{m-1} \varepsilon^k f_{jk}(x) \equiv \sum_{j\geqq 2} \kappa^j f_j(x;\varepsilon).$$

A transformation (I) results in the equation for $k = 1$ in (IV) (using the fact that $f_0 = g_0$) of the form

$$g_1 = \bar{f}_1 + \tilde{f}_1 + L_W f_0 - L_{f_0} W.$$

If now W, assumed of the form

$$W(y;\varepsilon) \equiv \sum_{i=0}^{m-1} \varepsilon^i W_i(y), \tag{3.2.2}$$

satisfies the differential equation

$$L_{f_0}W - L_W f_0 \equiv \frac{\partial W}{\partial y}\cdot f_0 - \frac{\partial f_0}{\partial y}\cdot W = \tilde{f}_1 + \varepsilon^m R, \tag{V}$$

there results

$$g_1 = \bar{f}_1 - \varepsilon^m R, \tag{3.2.3}$$

and the new differential equation for y becomes

$$\dot{y} = f_0(y;\varepsilon) + \kappa\bar{f}_1(y;\varepsilon) - \kappa\varepsilon^m R(y;\varepsilon) + \sum_{k\geqq 2}\kappa^k g_k(y;\varepsilon), \tag{3.2.4}$$

in which the g_k, $k \geqq 2$, are calculated using (IV), including $\varepsilon^m R$ in (3.2.3) in L_{g_1} in particular. Now it can be recognized that $\kappa = \varepsilon^m$ "in value", so (3.2.4) can be written, for $f_2^* = g_2 - R$, $f_k^* = g_k$, $k \geqq 3$

$$\dot{y} = (f_0 + \varepsilon^m\bar{f}_1) + \sum_{k\geqq 2}\varepsilon^{mk} f_k^*,$$

and the system is soluble to order $2m$, rather than just $m+1$ as with a linear method, through an equation of the original form (3.2.1). The process can be repeated with a new $\kappa^* = \varepsilon^{2m}$ to determine a new transformation generated by W^*, say, resulting in an equation soluble to order $4m$; iteration of the process effects *quadratic transformation* and can be carried through the desired order. The solution for the original variables can be recovered by applying directly the sequence of separate transformations generated by W, W^*, etc. through (I) and (II) to express the original in terms of the transformed variables' solution; initial conditions for the latter may be found by transforming those of the original variables through the inverse transformation (III) (in reverse order).

It is perhaps advisable to collect the results of the development (equations (I)–(III) from § 2 and (IV) and (V) above) in one place:

(I) *Coordinate Transformation*: $x(y;\kappa,\varepsilon) = \exp(\kappa L_{W(y;\varepsilon)}y)$.

(II) *Functional transformation*:

$$f(\exp(\kappa L_{W(y;\varepsilon)}y);\varepsilon) = \exp(\kappa L_{W(y;\varepsilon)}f(y;\varepsilon)).$$

(III) *Transformation Inverse*: $[\exp(\kappa L_W f^*)]^{-1} = \exp(\kappa L_{-W}f^*)$.

(IV) *Differential Transformation*: $\dot{x} = \sum\kappa^i f_i \to \dot{y} = \sum\kappa^i g_i$, *where*

$$g_0 = f_0,$$

$$g_k = f_k + \sum_{j=0}^{k-1}\frac{1}{(k-j)!}(L_W^{k-j}f_j - L_{g_j}(L_W^{k-j-1}W)).$$

(V) *Determining Equation*: $L_{f_0}W - L_W f_0 = \tilde{f}_1 + \varepsilon^m R$.

At this point, two questions might be raised: 1) why the remainder term, R in (V), and 2) why not determine higher orders of W than required to eliminate only f_1? Although the two matters are related, the answer to the first lies in the fact that a W (3.2.2) satisfying (V) with $R \equiv 0$ would generally require a power series extending to the ultimate order of transformation reduced only by the order of perturbation at that stage, a power series which must be applied to the f_j and W through (IV). But assuming that no higher orders of generating function are to be found—and this has still to be justified—perturbation is eliminated only to order $2m$, and only m orders of generating

function effect this elimination; more would result merely in unproductive effort. It is just this observation which led to modification [12] of the author's original technique.

This can be demonstrated, and the calculation of the explicit R assisted, by noting that for

$$\tilde{f}_1 = \sum_{i=0}^{m-1} \varepsilon^i \tilde{f}_{1i}, \quad W = \sum_{j=0}^{m-1} \varepsilon^j W_j, \quad f_0 = \sum_{k=0}^{m-1} \varepsilon^k f_{0k},$$

it is possible to express (V)

$$\sum_{i=0}^{m-1} \varepsilon^i \sum_{j=0}^{i} \left(\frac{\partial W_j}{\partial y} \cdot f_{0(i-j)} - \frac{\partial f_{0(i-j)}}{\partial y} \cdot W_j \right)$$
$$+ \sum_{i=m}^{2m-2} \varepsilon^i \sum_{j=(1-m)+i}^{m-1} \left(\frac{\partial W_j}{\partial y} \cdot f_{0(i-j)} - \frac{\partial f_{0(i-j)}}{\partial y} \cdot W_j \right) = \sum_{i=0}^{m-1} \varepsilon^i \tilde{f}_{1i};$$

the first (double) summation can be recognized as that determining the W_j, $0 \leqq j \leqq m-1$; the second is just the explicit expression for $\varepsilon^m R$.

Again, the fact that W, f_i and g_i are power series in ε bears mention. In particular, using the transformation equation (IV), it is a simple matter to prove inductively that the g_k in (3.2.4) are of the form

$$g_k = \sum_{i=0}^{(k+1)(m-1)} \varepsilon^i g_{ki}(y);$$

thus, for $\kappa = \varepsilon^m$, $\kappa^i g_i$ has terms from order im to $(im + (i+1)(m-1))$ inclusive (in ε), and there is an "overlapping" in magnitude with later such terms.

The second question will be answered in the next section.

3.3. A special feature of the method for nonlinear vibrations. In any transformation method for the solution of differential equations, the solution for the original variables is found by transformation of the final solution (known to the order of the transformed differential equation) through the explicit transformation (to its order). The method developed above has the ability to eliminate perturbation to twice the order of that present at a given step; in particular, at the last step of this procedure, perturbation would comprise half the terms in the differential equation, and the subsequent determination of an appropriate generating function would eliminate these. In this assurance, one can merely *drop* the perturbing terms at the last step without explicitly determining the generating function; but then the transformed differential equation (and its solution) is known to *twice* the order of the explicit transformation (and the *original* variables' solution).

This apparently self-defeating feature has important applications in the solution of nonlinearly vibrating systems, in which transformation generally is utilized to force the system, through averaging, to assume the form of a linear oscillator whose frequencies can then be determined. These frequencies should be determined as accurately as possible, for an error in the frequencies generates a time-linear (secular) "drift" away from the "actual" solution vis a vis the purely periodic error generated by truncation of the generating function (generally periodic in form). In a given problem, furthermore, the frequencies of vibration under perturbation can be more important than the actual solution. Thus it is typical in a linear method to average the next order in the transformed equation above the last explicit generating function's in the assurance that any periodic perturbation could be removed to that order, and resulting in a frequency known to one order higher than the transformation, and thus solution. But for times on

the order of the inverse square of the [original] small parameter, the secular error in the frequencies will already have encroached on the highest order of explicit solution, independent of the order to which transformation has been made. In the present technique, terms in the transformed equation can be averaged to *twice* the order of perturbation at the last step, and for times on the order of the inverse of the perturbation parameter at this *last* step (as opposed to the *original* perturbation parameter), secular error will still be of the same order as the periodic error from the generating function termination. In any event, the frequencies of the perturbed system have been determined to very high accuracy relative to the solution—an accuracy which increases with the order of transformation, in contrast with a linear method.

Though in principle higher orders of the generating function could be determined in a given transformation, it is the use of *separate* complete transformations which effects the continual redoubling of transformation elimination in the above technique (and the approach of Kolmogorov and Arnol'd on which the method is based), advising against finding such higher orders at each transformation step.

The efficiency of the present method in determining frequencies of nonlinear oscillation is exemplified in the next section.

4. Example: van der Pol's equation. In illustration of his generalized Lie transform technique, Kamel [6, § 4] considers the one-dimensional second-order van der Pol equation

$$\ddot{q}+q=\varepsilon(1-q^2)\dot{q}.$$

(Note that in this equation q is a *scalar*; so also in the subsequent discussion will be the quantities A, B, and C, which, along with ϕ, ψ, and σ, will be components of the 2 *vectors* x, y, and z respectively.)

Through a variation of parameters, letting

$$q=A\sin\phi, \qquad \dot{q}=A\cos\phi,$$

he puts the original equation in the form

$$\dot{x}=\frac{d}{dt}\begin{pmatrix}A\\ \phi\end{pmatrix}=\begin{pmatrix}0\\ 1\end{pmatrix}+\varepsilon\left(\begin{pmatrix}\frac{A}{2}\left(1-\frac{A^2}{4}\right)\\ 0\end{pmatrix}+\begin{pmatrix}\frac{A}{2}C_2+\frac{A^3}{8}C_4\\ \left(\frac{A^2}{4}-\frac{1}{2}\right)S_2-\frac{A^2}{8}S_4\end{pmatrix}\right)$$

$$\equiv f_0+\varepsilon(\bar{f}_1(A)+\tilde{f}_1(A,\phi)),$$

in which $C_i\equiv\cos i\phi$, $S_i\equiv\sin i\phi$—a two-dimensional, first-order system. By applying the present method to this problem, some comparison of the linear quadratic techniques can be made. As with Kamel, we shall transform this to new variables $y=(B,\psi)^T$ in terms of which the new differential equation is free of ψ. The procedure outlined above will be used to find the explicit transformation and solution to second order while determining the transformed equation to fourth.

Elimination of ψ to second order. We start, of course, with the fact from (IV) that

$$g_0=f_0=\begin{pmatrix}0\\ 1\end{pmatrix}.$$

That ϕ-dependent terms—for us the "perturbation"—are present to order ε recommends the perturbation parameter for the first step to be $\kappa=\varepsilon$. The determining equation (V) to be satisfied by the generating function $W=\begin{pmatrix}W_1\\ W_2\end{pmatrix}$ at this step by (3.2.2),

since $L_W f_0 \equiv 0$, becomes

$$L_{f_0} W \equiv \frac{\partial W}{\partial y} \cdot f_0 = \begin{pmatrix} \frac{\partial W_1}{\partial \psi} \\ \frac{\partial W_2}{\partial \psi} \end{pmatrix} = \tilde{f}_1 = \begin{pmatrix} \frac{B}{2} C_2 + \frac{B^3}{8} C_4 \\ \left(\frac{B^2}{4} - \frac{1}{2}\right) S_2 - \frac{B^2}{8} S_4 \end{pmatrix},$$

where now $C_i \equiv \cos i\psi$, $S_i \equiv \sin i\psi$. Thus,

$$W = \begin{pmatrix} \frac{B}{4} S_2 + \frac{B^3}{32} S_4 \\ -\left(\frac{B^2}{8} - \frac{1}{4}\right) C_2 + \frac{B^2}{32} C_4 \end{pmatrix} \tag{4.1}$$

and

$$g_1 = \begin{pmatrix} \frac{B}{2}\left(1 - \frac{B^2}{4}\right) \\ 0 \end{pmatrix}.$$

In this case, since the ε-power series for both f_0 and f_1 have only the single zeroth order term, the determining equation is satisfied exactly, so that $R \equiv 0$ and g_1 also has only a single term.

Calculation of the frequencies to fourth order. With only the explicit generating function W, we shall calculate the transformed (angle-independent) equation to fourth order, that to which a subsequent explicit transformation would be expected to eliminate the angles. Since we already have g_1 above, this will involve determining the angle-independent parts of both g_2 and g_3. By (IV), the latter in particular requires operation on g_2 by L_W, periodic with arguments 2ψ and 4ψ; thus, both secular and periodic terms with these arguments will be found for g_2.

From (IV), since $f_2 \equiv 0$,

$$g_2 = 0 + \frac{1}{2!}(L_W^2 f_0 - L_{g_0}(L_W W)) + \frac{1}{1!}(L_W f_1 - L_{g_1} W).$$

But $L_W^2 f_0 \equiv 0$, so the first term

$$-\frac{1}{2!} L_{g_0}(L_W W) \equiv -\frac{1}{2!}\left(\frac{\partial}{\partial y}\left(\frac{\partial W}{\partial y} \cdot W\right)\right) \cdot g_0$$

$$= -\begin{pmatrix} \left(-\frac{5B^3}{128} + \frac{B^5}{128}\right) S_2 + \left(-\frac{B}{16} + \frac{B^3}{16}\right) S_4 + \cdots \\ -\frac{B^2}{64} C_2 + \left(-\frac{1}{8} + \frac{B^2}{16} - \frac{B^4}{16}\right) C_4 + \cdots \end{pmatrix}.$$

Similarly

$$L_W f_1 \equiv \frac{\partial f_1}{\partial y} \cdot W$$

$$= \begin{pmatrix} \left(\frac{B}{8} - \frac{23B^2}{128} + \frac{B^5}{128}\right) S_2 + \left(-\frac{B}{16} + \frac{5B^3}{64} - \frac{3B^5}{256}\right) S_4 + \cdots \\ \left(-\frac{1}{8} + \frac{3B^2}{16} - \frac{11B^4}{256}\right) + \left(-\frac{7B^2}{64} + \frac{3B^4}{64}\right) C_2 + \left(-\frac{1}{8} + \frac{B^2}{16} - \frac{B^4}{32}\right) C_4 + \cdots \end{pmatrix},$$

and finally

$$-L_{g_1}W \equiv -\frac{\partial W}{\partial y}\cdot g_1 = -\begin{pmatrix} \left(\frac{B}{8}-\frac{B^3}{32}\right)S_2+\left(\frac{3B^3}{64}-\frac{3B^5}{256}\right)S_4 \\ \left(-\frac{B^2}{8}+\frac{B^4}{64}\right)C_2+\left(\frac{B^2}{32}-\frac{B^4}{128}\right)C_4 \end{pmatrix}.$$

Thus

$$g_2 = \begin{pmatrix} \left(-\frac{7B^3}{64}+\frac{3B^5}{128}\right)S_2+\left(-\frac{B^3}{32}\right)S_4+\cdots \\ \left(-\frac{1}{8}+\frac{3B^2}{16}-\frac{11B^4}{256}\right)+\left(\frac{B^2}{32}+\frac{B^4}{64}\right)C_2+\left(-\frac{B^2}{32}+\frac{B^4}{128}\right)C_4+\cdots \end{pmatrix}.$$

We are now prepared to calculate g_3. Again by (IV) (and $f_3 \equiv 0$),

$$g_3 = 0+\frac{1}{3!}(0-L_{g_0}(L_W^2W))+\frac{1}{2!}(L_W^2f_1-L_{g_1}(L_WW))+\frac{1}{1!}(L_W0-L_{g_2}W).$$

Here only purely secular terms need be found. Routine calculations yield that

$$-\frac{1}{3!}L_{g_0}(L_W^2W) = \text{(purely periodic)};$$

$$\frac{1}{2!}L_W^2f_1 = \begin{pmatrix} \left(\frac{3B}{128}-\frac{17B^3}{256}+\frac{239B^5}{8192}-\frac{91B^7}{32768}\right)+\text{(periodic)} \\ \text{(periodic)} \end{pmatrix},$$

$$-\frac{1}{2!}L_{g_1}(L_WW) = -\begin{pmatrix} \left(\frac{3B}{128}-\frac{15B^3}{512}+\frac{83B^5}{8192}-\frac{35B^7}{32768}\right)+\text{(periodic)} \\ \text{(periodic)} \end{pmatrix},$$

$$-L_{g_2}W = -\begin{pmatrix} \left(-\frac{3B^3}{512}+\frac{7B^5}{2048}+\frac{B^7}{2048}\right)+\text{(periodic)} \\ \text{(periodic)} \end{pmatrix}.$$

Thus, finally

$$g_3 = \begin{pmatrix} \left(-\frac{B^3}{32}+\frac{B^5}{64}-\frac{9B^7}{4096}\right)+\text{(periodic)} \\ \text{(periodic)} \end{pmatrix}.$$

Through third order, then, the transformation generated by W, (4.1), yields the new differential equation

$$\frac{d}{dt}\begin{pmatrix} B \\ \psi \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}+\varepsilon\begin{pmatrix} \frac{B}{2}-\frac{B^3}{8} \\ 0 \end{pmatrix}+\varepsilon^2\begin{pmatrix} 0 \\ -\frac{1}{8}+\frac{3B^2}{16}-\frac{11B^4}{256} \end{pmatrix}+\varepsilon^3\begin{pmatrix} -\frac{B^3}{32}+\frac{B^5}{64}-\frac{9B^7}{4096} \\ 0 \end{pmatrix} + \varepsilon^2(\text{periodic}). \tag{4.2}$$

But since the "perturbation" is of order ε^2, the next transformation, expected to be of

the form

$$y = \exp(\kappa^* L_{W^*} z),$$
$$\kappa^* = \varepsilon^2, \qquad W^*(z;\varepsilon) = W_0^*(z) + \varepsilon W_1^*(z) \tag{4.3}$$

(by (3.2.2)) for $z = (C, \sigma)^T$, would eliminate such terms to ε^4; thus the secular terms in the above expression (4.2) will represent the transformed equation after this second transformation to ε^4. The ordinary differential equation for the new variable C would be

$$\dot{C} = \varepsilon \frac{C}{2}\left(1 - \left(\frac{1}{4} + \frac{\varepsilon^2}{16}\right)C^2 + \frac{\varepsilon^2}{32}C^4 - \frac{9\varepsilon^2}{2048}C^6\right), \tag{4.4}$$

which could be solved by quadrature for $C(t)$, also reducing σ to solution by quadrature.

The above equation, due to the linear factor of C, will generally yield a solution exponential in the time. If, however, $\dot{C} = 0$, (4.2) shows that the solution for ψ will be linear in t, and thus, through purely periodic transformations, $x(z;\kappa,\varepsilon)$ will be periodic. Assuming C to be of the form

$$C = 2 + a\varepsilon + b\varepsilon^2 + c\varepsilon^3$$

and substituting this into the differential equation for C, (4.4), gives the condition that C is constant if $C = (2 - \varepsilon^2/32)$; this value of C then determines $\dot{\sigma} = (1 - \varepsilon^2/16)$—itself a constant. Note that both of these are accurate to fourth order and check with classical results ([6, p. 103]).

The transformation generated by the explicit W, (4.1), gives $q(x(y;\kappa,\varepsilon))$. The subsequent transformation giving y in terms of z, generated by W^*, has not been calculated explicitly, however; thus q is known only to second order by (4.3). (Recall that only a single explicit transformation step has been made.) To this order, then, q can be found by the mere substitution of z for y in W:

$$q^{**}(z;\kappa,\varepsilon) = q(\exp(\kappa L_{W(z)} z)) = \exp(\kappa L_{W(z)} q(z)) = q(z) + \kappa L_{W(z)} q(z),$$

where

$$q(x) = q((A, \phi)^T) = A \sin \phi.$$

But $\kappa = \varepsilon$ here, and

$$L_{W(z)} q(z) = \frac{C}{4}\left(1 - \frac{C^2}{4}\right)\cos\sigma - \frac{C^3}{32}\cos 3\sigma.$$

Thus

$$q^{**} = C \sin\sigma + \varepsilon\left(\frac{C}{4}\left(1 - \frac{C^2}{4}\right)\cos\sigma - \frac{C^3}{32}\cos 3\sigma\right),$$

where

$$\sigma = \left(1 - \frac{\varepsilon^2}{16}\right)t + \sigma_0.$$

The use of hand computation in the above example advised against explicit calculation of the second transformation (involving terms to *eighth* order); as presented, it does illustrate the ability of the proposed method to obtain transformed

differential equations (and thus their solutions) to twice the order of the transformations themselves and at least indicates the accelerated transformation property of the method. Comparison with Kamel [6, § 4] indicates the importance of the former property.

In his determination of the value $\bar{A}$ (corresponding to C above) giving periodic solution, Kamel neglects a third-order averaging despite knowledge of the explicit first- and second-order generating functions; since its time derivative is factored by ε (cf. (4.4)), this third order is required to obtain the value of $\bar{A}$ to order 2, and not doing so produces an acknowledged periodic error of second order in a solution ostensibly accurate to that order. Fortuitously, the form for both C and $\dot{\sigma}$ (corresponding to Kamel's ψ) in even powers of ε avoids a corresponding third-order error in the frequency in this case.

Contrast this with the present example, in which although an *explicit* solution is known only to second order (the explicit W^* above required to bring it to fourth) both C and $\dot{\sigma}$ are known to order 4; the same device employed after a second explicit transformation would give the explicit solution to order 4 and frequency to order 8. In addition, this solution erring periodically at order 2 avoids *secular* error for times comparable with ε^{-2}, and a second transformation would do so for $t \sim \varepsilon^{-4}$; Kamel's—or *any* linear—method can only hold for times on the order of ε^{-1} before suffering such secular error.

5. Comparisons with available techniques. Attention will be limited to methods using Lie transforms, in order to maintain the inversion properties used to obtain the original solution in terms of the ultimately transformed one.

In this regard, the linear methods of both Kamel [6, § 3] and Henrard [4, Algorithm IV] are highly efficient due to the fact that only nonvanishing terms in the original perturbed differential equation generate the intermediate functions mentioned in § 3.1 above; previous algorithms, like their earlier Hamiltonian counterparts, generated the same number of these functions independent of the form of the original equation (cf., e.g., Kamel [6, § 2] or Henrard [4, Algorithm II]). Henrard [4] shows the vector formulations of both these techniques to be straightforward generalizations of the respective scalar developments (his Algorithms I and III), justifying comparison of the scalar methods as an indication of the efficiency of the vector ones.

The author [12, § 5] has made a detailed study comparing his Hamiltonian quadratic formulation with the "inverse algorithms" of Kamel [3] and Henrard [5] (which are generalized to their efficient vector methods) and more traditional linear methods, using the number of Lie operator applications as a criterion. Briefly stated, the ("nondegenerate") results indicate that to obtain the original solution to the same order as the transformed equation for starting Hamiltonians with very few terms, the inverse algorithms require fewer Lie operator applications both to transform the Hamiltonian and to invert the transformed solution for that of the original variables; but for starting Hamiltonians with second- or third-order terms, the quadratic technique becomes significantly more efficient in transformation. Both are more efficient in both transformation and inversion than previous linear methods.

But the present technique also allows the option—not available in *any* linear method—of determining the transformed differential equation to twice the order of solution (cf. § 3.3 above). Thus to find the perturbed frequencies in a nonlinear system to a given order requires some fewer Lie operator applications than needed to find both the frequencies *and* solution to the same order ([12, § 5]) and might be desired independent of the effort in any event.

REFERENCES

[1] G. HORI, *Theory of general perturbations with unspecified canonical variables*, Pub. Ast. Soc. Jap., 18 (1966), pp. 287–296.

[2] A. DEPRIT, *Canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 12–30.

[3] A. A. KAMEL, *Expansion formulae in canonical transformations depending on small parameters*, Celestial Mech., 1 (1969), pp. 190–199.

[4] J. HENRARD, *On a perturbation theory using Lie transforms*, Celestial Mech., 3 (1970), pp. 107–120.

[5] ———, *The Algorithm of the Inverse for Lie Transform*, Recent Advances in Dynamical Astronomy, B. D. Tapley and V. Szebehely, eds., D. Reidel Publishing Co., Dordrecht, Holland, 1973, pp. 250–259.

[6] A. A. KAMEL, *Perturbation theory in the theory of nonlinear oscillations*, Celestial Mech., 3 (1970), pp. 90–106.

[7] J. MOSER, *Lectures on Hamiltonian systems*, Mem. Amer. Math. Soc., 81 (1968), pp. 1–60.

[8] H. POINCARÉ, *Les methodes nouvelles de la mécanique céleste*, Gauthier-Villars, Paris, 1892. NASA TT F-451, National Aeronautics and Space Administration, Washington, DC 1967.

[9] A. N. KOLMOGOROV, *The conservation of conditionally periodic motions with a small change in the Hamiltonian*, Dokl. Akad. Nauk SSSR, 98 (1954), pp. 527–530.

[10] V. I. ARNOL'D, *Small denominators and problems of stability of motion in classical and celestial mechanics*. Russian Math. Surveys, 18 (1963), pp. 85–191.

[11] R. A. HOWLAND, *An accelerated elimination technique for the solution of perturbed Hamiltonian systems*, Celestial Mech., 15 (1977), pp. 327–352.

[12] ———, *An improved transformation-elimination technique for the solution of perturbed Hamiltonian systems*, Celestial Mech., 19 (1979), pp. 95–110.

[13] A. A. KAMEL, *Lie transforms and the Hamiltonization of non-Hamiltonian systems*, Celestial Mech., 4 (1971), pp. 397–405.

[14] A. H. NAYFEH, *Perturbation Methods*, John Wiley, New York, 1973.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0010 $01.00/0

REGULARITY OF SINGULAR TWO-POINT BOUNDARY VALUE PROBLEMS*

ROBERT SCHREIBER†

Abstract. Two-point boundary value problems like $(p(x)x^{\sigma}u')' = f(x)$ on (0, 1) have a regular singular point at the boundary, and solutions that are not smooth but behave like $x^{1-\sigma}$ near 0. Bounds on weighted Sobolev norms of their solutions, essential to the theory of finite element approximations for such problems, are obtained here.

1. Introduction and main results. In this paper, we prove a regularity theorem for the singular two-point boundary value problem

$$-D(p(x)Du)+q(x)u=f(x), \qquad 0<x<1, \tag{1.1a}$$

$$u(0)=u(1)=0, \tag{1.1b}$$

where

$$D\equiv\frac{d}{dx},$$

and

$$p(x)=x^{\sigma}\rho(x), \qquad q(x)=x^{\sigma}\gamma(x), \tag{1.2a}$$

for some $0\leqq\sigma<1$,

$$\rho(x)\geqq\rho_{\min}>0, \tag{1.2b}$$

and

$$f\in H^m, \quad \rho\in W^{m+1,\infty}, \quad \gamma\in W^{m,\infty}, \tag{1.2c}$$

for some integer $m\geqq 0$. Here, $W^{m,p}$ denotes the usual Sobolev space of functions with weak derivatives of order up to m in $L^p(0, 1)$, with norm

$$\|u\|_{m,p}^p\equiv\sum_{j=0}^{m}\|D^j u\|_{L^p(0,1)}^p.$$

We let $H^m = W^{m,2}$ and $\|\cdot\|_m = \|\cdot\|_{m,2}$. $H_0^1\subset H^1$ consists of the H^1 functions which vanish at 0 and 1. We denote by $(f, g)_m$ the inner product of H^m.

Problems of this type arise in the study of generalized axisymmetric potentials [7], [4]; their numerical solution by finite element methods is discussed by the author [6]. That paper assumes the bounds on the solution to be proved here.

Let S be the weighted Sobolev space of absolutely continuous functions u such that $u(0) = u(1) = 0$ and $x^{\sigma/2}Du\in L^2(0, 1)$. The solution u of (1.1) is in S but is not always in H_0^1 [1].

We note that the form

$$a(u, v)\equiv\int_0^1 (pDuDv+quv)\,dx,$$

* Received by the editors January 23, 1980. This work constitutes part of the author's doctoral dissertation at Yale University, New Haven, Connecticut.

† Computer Science Department, Stanford University, Stanford, California 94305.

is positive definite over S, and therefore, there exists a unique $u \in S$ satisfying

$$a(u, v) = (f, v)_0,$$

for all $v \in S$, which we call the generalized solution of (1.1). The solution is bounded by the right-hand side,

$$\|u\|_0 \leqq C_0 \|f\|_0, \tag{1.3}$$

where C_0 is a constant independent of u and f.

MAIN THEOREM. *There exists a positive constant* Γ *independent of* u *and* f *such that, if* u *is the generalized solution of* (1.1), *then for all* $0 \leqq l \leqq m$,

$$\|D(x^\sigma Du)\|_l \leqq \Gamma \|f\|_l.$$

2. Proof of main theorem. The proof requires three lemmas.

LEMMA 1. *If* $g \in C^{m+1}[0, 1]$, $g(0) = 0$ *and* $f(x) \equiv g(x)/x$, $0 < x \leqq 1$, *then* $f \in C^m[0, 1]$ *and*

$$\lim_{x \to 0} D^l f(x) = \frac{D^{l+1} g(0)}{l+1},$$

for all $0 \leqq l \leqq m$.

Proof. It is clear that $f \in C^m(0, 1]$. A simple computation shows that

$$D^l f(x) = x^{-(l+1)} \sum_{j=0}^{l} \frac{l!}{j!} (-1)^{l-j} x^j D^j g,$$

for all $x > 0$. Taking limits with the aid of L'Hôpital's rule, we obtain

$$\lim_{x \to 0} D^l f(x) = \lim_{x \to 0} \frac{\sum_{j=0}^{l} (l!/j!)(-1)^{l-j} (j x^{j-1} D^j g + x^j D^{j+1} g)}{(l+1) x^l} = \lim_{x \to 0} \frac{D^{l+1} g}{l+1}. \qquad \square$$

Hardy's inequality [3] states that when $g \in H^1$ and $g(0) = 0$, then $\|g/x\|_0 \leqq 2\|Dg\|_0$. We now generalize to the case $g \in H^{m+1}$.

LEMMA 2. *If* $g \in H^{m+1}$, $g(0) = 0$ *and* $f(x) \equiv g(x)/x$, *then* $f \in H^m$ *and*

$$\|D^l f\|_0 \leqq \frac{2}{2l+1} \|D^{l+1} g\|_0. \tag{2.1}$$

Proof. We shall show (2.1) for $g \in C^{m+1}$. This will suffice, since $\{g \in C^{m+1} | q(0) = 0\}$ is dense in $\{g \in H^{m+1} | q(0) = 0\}$ by Sobolev's inequality [2].

A routine computation shows that, for $1 \leqq l \leqq m$,

$$x D^l f = D^l g - l D^{l-1} f. \tag{2.2}$$

Since for $l = 0$ the result is just Hardy's inequality, we assume $l \geqq 1$. Integrating by parts,

$$\begin{aligned} \|D^l f\|_0^2 &= \int_0^1 x^{-2} [x D^l f]^2 \, dx \\ &= -x^{-1} [x D^l f]^2 \Big|_0^1 + 2 \int_0^1 D^l f D(x D^l f) \, dx. \end{aligned}$$

The integrated term is nonpositive, since at $x = 0$, $D^l f$ is finite (by Lemma 1). Thus, by (2.2)

$$\|D^l f\|_0^2 \leqq 2 \int_0^1 D^l f (D^{l+1} g - l D^l f),$$

whence, by the Cauchy-Schwarz inequality,

$$(2l+1)\|D^l f\|_0^2 \leqq 2\|D^l f\|_0 \|D^{l+1} g\|_0. \qquad \square$$

The next lemma is of interest for its own sake, as it shows that if $D(x^\sigma Du)$ is smooth, then u is of the form $x^{-\sigma}v$, with v smooth.

LEMMA 3. *If $u \in S$, and $D(x^\sigma Du) \in H^m$, then $x^\sigma u \in H^{m+2}$, and there exists a positive constant Γ_1 independent of u such that, for $1 \leqq l \leqq m+2$,*

$$\|D^l(x^\sigma u)\|_0 \leqq \Gamma_1 \|D^{l-1}(x^\sigma Du)\|_0. \tag{2.3}$$

Proof. Define

$$v_0(t) = t^\sigma Du(t), \qquad 0 \leqq t \leqq 1,$$

and, for $1 \leqq i \leqq m+1$,

$$v_i(t) = \frac{(v_{i-1}(t) - v_{i-1}(0))}{t}, \qquad 0 < t \leqq 1.$$

By hypothesis, $v_0 \in H^{m+1}$; furthemore, it follows from Lemma 2 that $v_i \in H^{m+1-i}$ and, for $i < l \leqq m+2$,

$$\|D^{l-i-1} v_i\|_0 \leqq \left(\prod_{j=l-i-1}^{l-2} \left(\frac{2}{2j+1} \right) \right) \|D^{l-1} v_0\|_0. \tag{2.4}$$

We now show by induction on l that

$$D^l(x^\sigma u) = D^{l-1}v_0 + c_1 D^{l-2} v_1 + \cdots + c_{l-1} D^0 v_{l-1} + c_l x^{\sigma-l} \int_0^x t^{l-1-\sigma} v_{l-1}(t)\, dt, \tag{2.5}$$

where $c_k = \prod_{i=0}^{k-1} (\sigma - i)$. Since $u \in S$,

$$x^\sigma u(x) = x^\sigma \int_0^x t^{-\sigma} (t^\sigma Du(t))\, dt,$$

and

$$D(x^\sigma u) = v_0 + \sigma x^{\sigma-1} \int_0^x t^{-\sigma} v_0(t)\, dt,$$

which is (2.5) for $l = 1$. For the induction step, first note that the last term on the right-hand side of (2.5) satisfies

$$x^{\sigma-l} \int_0^x t^{l-1-\sigma} v_{l-1}(t)\, dt = x^{\sigma-l} \int_0^x t^{l-1-\sigma} [v_{l-1}(0) + t v_l(t)]\, dt$$

$$= \frac{1}{l-\sigma} v_{l-1}(0) + x^{\sigma-l} \int_0^x t^{l-\sigma} v_l(t)\, dt.$$

Now assume (2.5) holds for l. Differentiating, we obtain

$$D^{l+1}(x^\sigma u) = D^l v_0 + c_1 D^{l-1} v_1 + \cdots + c_{l-1} D v_{l-1}$$

$$+ c_l v_l + (\sigma - l) c_l x^{\sigma-l-1} \int_0^x t^{l-\sigma} v_l(t)\, dt,$$

which is (2.5) for $l+1$.

Inequality (2.3) now follows from (2.4) and (2.5) provided we can show that

$$\left\| x^{\sigma-l} \int_0^x t^{l-1-\sigma} v_{l-1}(t)\, dt \right\|_0 \leqq C \|v_{l-1}\|_0,$$

for all $1 \leqq l \leqq m+2$, with C a constant independent of v. Let

$$I(x) \equiv \int_0^x t^{l-1-\sigma} v_{l-1}(t)\, dt.$$

Integrating by parts, we obtain

$$\begin{aligned} \|x^{\sigma-l} I(x)\|_0^2 &= \int_0^1 x^{2\sigma-2l} [I(x)]^2\, dx \\ &= \frac{x^{2\sigma-2l+1}}{2\sigma-2l+1} \int_0^1 x^{2\sigma-2l+1} I(x) x^{l-1-\sigma} v_{l-1}(x)\, dx. \end{aligned}$$

If the integrated term is nonpositive, then, by the Cauchy-Schwarz inequality,

$$\|x^{\sigma-l} I(x)\|_0^2 \leqq \left| \frac{2}{2l-2\sigma-1} \right| \|x^{\sigma-1} I(x)\|_0 \|v_{l-1}\|_0.$$

As for the integrated term, at $x = 1$ it is negative, except (possibly) when $l = 1$. But in that case,

$$I(1) = \int_0^1 Du(t)\, dt = u(1) - u(0) = 0,$$

so the integrated term vanishes.

At $x = 0$ it vanishes, too. For, by the Cauchy-Schwarz inequality,

$$\begin{aligned} (I(x))^2 &= \left(\int_0^x t^{l-1-\sigma} v_{l-1}(t)\, dt \right)^2 \\ &\leqq \left(\int_0^x t^{2(l-1-\sigma)}\, dt \right) \left(\int_0^x v_{l-1}^2\, dt \right) \\ &\leqq \frac{x^{2l-2\sigma-1}}{2l-2\sigma-1} \int_0^x v_{l-1}^2(t)\, dt, \end{aligned}$$

whence, since $v_{l-1} \in L^2(I)$,

$$\lim_{x \to 0+} x^{2\sigma-2l+1} (I(x))^2 \leqq \frac{1}{2l-2\sigma-1} \lim_{x \to 0+} \int_0^x v_{l-1}^2(t)\, dt = 0. \qquad \square$$

Proof of main theorem. We start by showing that $D(x^\sigma Du) \in H^m$ and

$$\|D(x^\sigma Du)\|_l \leqq \bar{C} \|f - qu\|_l, \qquad 0 \leqq l \leqq m, \tag{2.6}$$

where $\bar{C} > 0$ is independent of f.

Following Reddien [5], we explicitly construct the generalized solution. First, let

$$g(x) = \int_0^x (qu - f)(t)\, dt.$$

Clearly, $g \in L^\infty$ and

$$\|g\|_{L^\infty} \leqq \|f - qu\|_0.$$

Next, let

$$h(x)=\frac{g(x)}{p(x)}=x^{-\sigma}\frac{g(x)}{\rho(x)},$$

and

$$\hat{w}(x)=\int_0^x h(t)\,dt.$$

Finally, let

$$w(x)=\hat{w}(x)-\hat{w}(1)Y(x),$$

where Y is the solution of the problem

$$-D(p\,DY)=0,$$
$$Y(0)=0, \qquad Y(1)=1.$$

We claim that $w\in S$ and, moreover, $w=u$, the generalized solution. Clearly, $w(0)=w(1)=0$. Moreover, the construction shows that

$$x^{\sigma/2}Dw=x^{-(\sigma/2)}\frac{g}{\rho}-\hat{w}(1)x^{\sigma/2}DY,$$

which has finite square integral; hence $w\in S$.

The construction also shows that

$$(p\,Dw, Dv)=(f-qu, v)$$

for all $v\in S$, whence $(p\,D(u-w), D(u-w))=0$, i.e., $u=w$.

To obtain the bound (2.6), note that

$$D^{l+1}(x^{\sigma}D\hat{w})=D^{l+1}\left(\frac{g}{\rho}\right), \qquad 0\leqq l\leqq m,$$

whence,

$$\|D(x^{\sigma}D\hat{w})\|_l\leqq c_0\|g\|_{l+1}$$
$$\leqq c_1\|f-qu\|_l, \qquad 0\leqq l\leqq m,$$

where c_1 depends only on l and $\|\rho\|_{W^{m+1,\infty}}$. Furthermore,

$$|\hat{w}(1)|\leqq\frac{1}{\rho_{\min}}\|g\|_{L^\infty}\int_0^1 x^{-\sigma}\,dt\leqq C\|f-qu\|_0.$$

Thus,

$$\|D(x^{\sigma}Du)\|_l\leqq\|D(x^{\sigma}D\hat{w})\|_l+|\hat{w}(1)|\,\|D(x^{\sigma}DY)\|_l$$
$$\leqq\bar{C}\|f-qu\|_l.$$

Now we show that there exist constants M_l such that

$$\|f-qu\|_l\leqq M_l\|f\|_l. \tag{2.7}$$

We proceed by induction on l. For $l=0$, the a priori estimate (1.3) yields

$$\|u\|_0\leqq C_0\|f\|_0,$$

whence,

$$\|f-qu\|_0 \leqq (1+C_0\|q\|_{L^\infty})\|f\|_0$$
$$\equiv M_0\|f\|_0.$$

Since $D(x^\sigma Du)\in H^0$, we can apply Lemma 3, which yields

$$u = x^{-\sigma}v, \qquad v\in H^2,$$

and

$$\|v\|_2 \leqq \Gamma_1\|D(x^\sigma Du)\|_0 \leqq \Gamma_1\bar{C}M_0\|f\|_0.$$

Thus, $qu = (x^\sigma\gamma)(x^{-\sigma}v) = \gamma v$ and

$$\|f-qu\|_2 \leqq \|f\|_2 + \|\gamma v\|_2$$
$$\leqq \|f\|_2 + c_2\|v\|_2$$
$$\leqq M_2\|f\|_2,$$

where c_2 depends on $\|\gamma\|_{W^{2,\infty}}$ and $M_2 = 1+c_2\Gamma_1\bar{C}M_0$).

We reiterate this argument for $l = 2, 4, \cdots, m$ (if m is even) or $l = 2, 4, \cdots, m-1$ (if m is odd). To obtain (2.7) for odd values of l, possibly including m, we apply Lemma 3. Since $D(x^\sigma Du)\in H^{l-1}$, $u = x^{-\sigma}v$, where $v\in H^{l+1}$ and

$$\|v\|_{l+1} \leqq \Gamma_1\|D(x^\sigma Du)\|_{l-1}$$
$$\leqq \Gamma_1\bar{C}\|f-qu\|_{l-1}$$
$$\leqq \Gamma_1\bar{C}M_{l-1}\|f\|_{l-1}.$$

Thus,

$$\|f-qu\|_l \leqq \|f\|_l + \|\gamma v\|_l$$
$$\leqq \|f\|_l + c_l\|v\|_l$$
$$\leqq \|f\|_l + c_l\|v\|_{l+1}$$
$$\leqq M_l\|f\|_l,$$

where c_l depends on $\|\gamma\|_{W^{l,\infty}}$ and $M_l = 1+c_l\Gamma_1\bar{C}M_{l-1}$. □

Acknowledgment. The author wishes to thank Professors Martin H. Schultz and Stanley C. Eisenstat for their invaluable comments and encouragement.

REFERENCES

[1] P. G. CIARLET, F. NATTERER AND R. S. VARGA, *Numerical methods of high-order accuracy for singular nonlinear boundary value problems*, Numer. Math., 15 (1970), pp. 87–99.

[2] AVNER FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice Hall, Englewood Cliffs, NJ, 1964.

[3] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, Cambridge University Press, London, 2nd ed., 1952.

[4] SEYMOUR V. PARTER, *Numerical methods for generalized axially symmetric potentials*, SIAM J. Numer. Anal., 2 (1965), pp. 500–516.

[5] G. W. REDDIEN, *Projection methods and singular two-point boundary value problems*, Numer. Math., 21 (1973), pp. 193–205.

[6] ROBERT SCHREIBER, *Finite element methods of high-order accuracy for singular two-point boundary value problems with nonsmooth solutions*, SIAM J. Numer. Anal., 17 (1980), pp. 547–566.

[7] ALEXANDER WEINSTEIN, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–38.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0011 $01.00/0

PIECEWISE MONOTONE INTERPOLATION IN POLYNOMIAL TIME*

L. RAYMON†

Abstract. There are presently in the mathematical literature constructive proofs of the existence of algebraic polynomials P that interpolate given data and remain monotone between the data points, together with upper bound estimates on the degree of P as a function of the data. These constructions, however, are impractical for large numbers k of data points, since P is obtained as a linear combination of 2^k polynomial approximations. A constructive proof of the existence of such a piecewise monotone interpolating polynomial is given here, with P obtained as a linear combination of k polynomials, together with an upper bound on the degree of P.

1. Introduction. Let $X=\{x_i\}_0^k$ where $0=x_0<x_1<\cdots<x_k=1$, and let $Y=\{y_i\}_0^k$ be real numbers such that $y_{i-1}\neq y_i$, $i=1,\cdots,k$. An algebraic polynomial $P(x)$ with the properties

(i) $P(x_i)=y_i$, $1\leqq i\leqq k$, and

(ii) $P(x)$ is monotone on the interval (x_{i-1}, x_i), $1\leqq i\leqq k$, is said to *interpolate Y at X piecewise monotonely*. It is a result of W. Wolibner [4] and S. W. Young [5] that for each X, Y there exists a polynomial that interpolates Y at X piecewise monotonely. The smallest degree of a polynomial that interpolates Y and X piecewise monotonely is called the *degree of piecewise monotone interpolation of Y with respect to X*, and is denoted by $N=N(X, Y)$. Estimates on the degree of piecewise monotone interpolation are seen to depend on the "degree of comonotone approximation", which we proceed to define.

A function $f(x)$ on $[a, b]$ is said to be *piecewise monotone* if $[a, b]$ may be partitioned into a finite number of subintervals on which f is alternately nondecreasing and nonincreasing. $f(x)$ and $g(x)$ are said to be comonotone on $[a, b]$ if they are piecewise monotone and are alternately nondecreasing and nonincreasing on the same subintervals. If f is piecewise monotone on $[a, b]$ we denote by $\mathscr{P}_n^*(f)$ the set of all polynomials of degree $\leqq n$ comonotone with f on $[a, b]$. The *degree of comonotone approximation* of f, $E_n^*(f)$, is defined by

$$E_n^*(f)=\inf_{P\in\mathscr{P}}\|f-P\|,$$

where $\|\cdot\|$ denotes the sup norm. If S is a set of comonotone functions, the *degree of comonotone approximation to the set S* is given by

$$E_n^*(S)=\sup_{f\in S}E_n^*(f).$$

Let

$$\Delta=\Delta(Y)=\min_{1\leqq i\leqq k}|y_i-y_{i-1}|,$$

and

$$M=M(X, Y)=\max_{1\leqq i\leqq k}\left|\frac{y_i-y_{i-1}}{x_i-x_{i-1}}\right|.$$

As a consequence of a result of Passow and Raymon [3, Thm. 1], connecting the degree of piecewise monotone approximation to the degree of comonotone

* Received by the editors March 17, 1980.

† Department of Mathematics, Temple University, Philadelphia, Pennsylvania 19122.

approximation, one obtains the following estimate on $N(X, Y)$: There is an absolute constant C such that

$$N(X, Y) \leqq \frac{CM}{\Delta}. \tag{1}$$

(Although this estimate on $N(X, Y)$ is somewhat better than that claimed in [3], it does follow immediately from Theorem 1 in that article together with the now improved estimates of Newman [2] and Iliev [1] on the degree of comonotone approximation.) The estimate (1) is satisfying in the sense that it is "neat" and reasonably small. Furthermore, the proof is constructive. The rub is that the desired interpolating polynomial is a linear combination of 2^k polynomials, each of which is a comonotone approximation of a designated function. This exponential growth of the number of polynomials required as a function of the number $k+1$ of interpolation points obviously limits the practical value of the result for anything but relatively small values of k. In the spirit generated (or, at least, accelerated) by the celebrated result of Khachian in linear programming, the purpose of this article is to reduce the "exponential time" required for the production of the interpolating polynomial to "polynomial time." We are in fact able to reduce the number of polynomial approximations, of which the desired interpolating polynomial P is a linear combination, to exactly k. In order to gain this advantage, the construction is more complex, and the estimate on the degree of P is not quite as small. Here, however, the net result is, nevertheless, far more efficient "time-wise" than the previous construction, except for special cases. We summarize the result and proceed with the construction.

2. The result. *Let* $X = \{x_i\}_0^k$ *(where* $0 = x_0 < x_1 < \cdots < x_k = 1$*) and* $Y = \{y_i\}_0^k$ *be real numbers such that* $y_{i-1} \neq y_i$, $i = 1, \cdots, k$. *Then* Y *can be interpolated at* X *piecewise monotonely with an algebraic polynomial* $P(x)$. $P(x)$ *may be constructed as a linear combination of* k *polynomials, each of which is a comonotone approximation to a piecewise linear function determined by* X *and* Y. *The degree* N *of* P *satisfies*

$$N \leqq \frac{Av}{\Delta\delta}, \tag{2}$$

where A *is an absolute constant,*

$$\Delta = \Delta(Y) = \min_{1 \leqq j \leqq k} |y_j - y_{j-}| \qquad \textit{(as above)},$$

$$v = \textit{the total variation}, \qquad \sum_1^k |y_j - y_{j-1}|,$$

$$\delta = \quad (X) = \min_{1 \leqq j \leqq k} (x_j - x_{j-1}).$$

3. The construction and the derivation of the degree bound. Let S_k be the set of all continuous piecewise monotone functions f increasing on each of the intervals $[x_{i-1}, x_i]$ for which $y_i - y_{i-1} > 0$, decreasing on each of the intervals for which $y_i - y_{i-1} < 0$, $1 \leqq i \leqq k$, such that

$$\sup_{0 \leqq a, b \leqq 1} \left| \frac{f(b) - f(a)}{b - a} \right| \leqq 1.$$

By Newman's theorem [2] $\lim_{n \to \infty} E_n^*(S_k) \to 0$. For each $\eta > 0$ the smallest degree n such that $E_n^*(S_k) < \eta$ will be denoted by $n(k, \eta)$. We will construct a polynomial P that

interpolates Y at X piecewise monotonely and of degree

$$N \leqq n_k\left(\frac{\Delta\delta}{3v}\right). \tag{3}$$

$P(x)$ will be the desired polynomial, a linear combination of k polynomials. The desired estimate (2) will then follow from (3) and estimates on the degree of comonotone approximation.

There is no loss of generality if it is assumed that $y_0 = 0$. Each vector $Y = (0, y_1, \cdots, y_k)$ may be corresponded to a point in E^k as follows:

$$T\colon\ Y = (0, y_1, \cdots, y_k) \to (y_1, y_2 - y_1, \cdots, y_k - y_{k-1}).$$

Let $\bar{X} = (\bar{x}_0, \cdots, \bar{x}_k)$ and $\bar{Y} = (0, \bar{y}_1, \cdots, \bar{y}_k)$ be given. Let S be the set of all vectors with the same monotonicity as $\bar{Y}$ (i.e., all $Y = (0, y_1, \cdots, y_k)$ such that $(y_j - y_{j-1})(\bar{y}_j - \bar{y}_{j-1}) \geqq 0$, $j = 1, 2, \cdots, k$). Then T represents a one-to-one correspondence between S and the open orthant O in E^k determined by the signature of $(\bar{y}_1, \bar{y}_2 - \bar{y}_1, \cdots, \bar{y}_k - \bar{y}_{k-1})$. Let $M(N)$ be the subset of O which corresponds under T to vectors in S with degree of piecewise monotone interpolation $\leqq N$ with respect to $\bar{X}$. It is then sufficient to show that $N = [n_k(\Delta\delta/3v)] \Rightarrow T(\bar{Y}) \in M(N)$.

If $Q \in M(N)$, then $aQ \in M(N)$ for every $a > 0$. (Indeed, if $p(x)$ interpolates Y piecewise monotonely, $ap(x)$ interpolates aY piecewise monotonely; also, $T(Y) = Q \Rightarrow T(aY) = aQ$). If $Q, Q' \in M(N)$, then $Q + Q' \in M(N)$. (If $Y, Y' \in S$, then $Y + Y' \in S$, and $p(x) + p'(x)$ is a piecewise monotone interpolation of $Y + Y'$, where $p(x)$, $p'(x)$, respectively, interpolate Y, Y' piecewise monotonely; then $\left.\begin{matrix} T(Y) = Q \\ T(Y') = Q' \end{matrix}\right\} \Rightarrow$ $T(Y + Y') = Q + Q'$.) Hence, $M(N)$ is a convex cone in O. Let $\lambda_j = e_j \cdot \operatorname{sgn}(\bar{y}_i - \bar{y}_{i-1})$, $j = 1, 2, \cdots, k$, where the e_j are the ordinary e-vectors spanning E^k. The points λ_j are not elements of O, but are boundary points of O. Let H be the standard $(k-1)$-simplex determined by λ_j, $j = 1, 2, \cdots, k$ (i.e., the convex hull of the λ_j.) H is also the set of all points $(p_1, p_2, \cdots, p_k) = P \in O$ such that $|p_1| + |p_2| + \cdots + |p_k| = 1$; that is, all points $Y = (y_1, y_2 - y_1, \cdots, y_k - y_{k-1})$ such that $v = |y_1| + |y_2 - y_1| + \cdots + |y_k - y_{k-1}| = 1$. Thus, $v^{-1}Y \in H$, and

$$Y \in M(N) \Leftrightarrow v^{-1}Y \in M(N).$$

We shall complete the construction of the polynomial for (3) by finding approximations λ_j^* to λ_j in $H \cap M(N)$ sufficiently close to λ_j to ensure that $v^{-1}Y$ is in the convex hull of λ_j^*, $j = 1, \cdots, k$.

For each j, we may define a piecewise linear function $L_j(x)$ with the following properties:

(i) $L_j(x)$ is continuous;

(ii) L_j is increasing and linear on each subinterval $[\bar{x}_{i-1}, \bar{x}_i]$, $i = 1, 2, \cdots, k$ on which $\bar{y}_i - \bar{y}_{i-1} > 0$; $L_j(x)$ is decreasing and linear on each subinterval on which $\bar{y}_i - \bar{y}_{i-1} < 0$;

(iii) $L_j(\bar{x}_j) - L_j(\bar{x}_{j-1}) = 1$; $\quad i \neq j \Rightarrow |L_j(\bar{x}_i) - L_j(\bar{x}_{i-1})| \leqq \delta\Delta/3v$. ($L_j(x)$ may be obtained by defining it to have appropriate values at $x = \bar{x}_0, \bar{x}_1, \cdots, \bar{x}_k$, and defining it linearly in between.) Observe that $\delta L_j \in S_k$. Hence, there is a polynomial $q(x)$ of degree $\leqq n = n_k(\Delta\delta/3v)$ comonotone with $\delta L_j(x)$ such that

$$\|p - \delta L_j\| \leqq E_n^*(S_k) \leqq \frac{\Delta\delta}{3v};$$

then $\delta^{-1}q(x)$ is a polynomial of degree $\leqq n$ comonotone with $L_j(x)$ such that

$$\|\delta^{-1}q - L_j\| \leqq \frac{\Delta}{3v}.$$

Let $p(x) = \delta^{-1}[q(x) - q(0)]$. Then,

$$i \neq j \Rightarrow |p(\bar{x}_i) - p(\bar{x}_{i-1})| < \frac{\Delta}{v}, \qquad 1 - \frac{\Delta}{v} \leqq |p(\bar{x}_j) - p(\bar{x}_{j-1})| \leqq 1 + \frac{\Delta}{v}. \tag{4}$$

Now

$$(p(\bar{x}_0) = 0, p(\bar{x}_1), p(\bar{x}_2), \cdots, p(\bar{x}_k)) \in S$$

and

$$\bar{\lambda}_j = (p(\bar{x}_1), p(\bar{x}_2) - p(\bar{x}_1), \cdots, p(\bar{x}_k) - p(\bar{x}_{k-1}) \in M(N).$$

Also, by (4),

$$\|\bar{\lambda}_j - \lambda_j\| \leqq \frac{\Delta}{v} \qquad (\text{where } \|(z_1, z_2, \cdots, z_k)\| = \max_{1 \leqq i \leqq k} |z_i|).$$

Thus, $\bar{\lambda}_j$ lies in O and in the k-dimensional cube centered at λ_j with side $2\Delta v^{-1}$ (and diameter $2k^{1/2}\Delta v^{-1}$); i.e., $\bar{\lambda}_j$ lies in the k-dimensional rectangular parallelopiped $\bar{H}_j$ determined by the intersection of O with this cube. Multiplying $\bar{\lambda}_j$ by the appropriate positive constant (the total variation of $\bar{\lambda}_j$), we obtain a point λ_j^* of $M(N)$ in $H_j^* = \bar{H}_j \cap H$. H_j^* is then the regular $(k-1)$-simplex whose k vertices are λ_j and the $k-1$ points on the edges of H at a distance $k^{1/2}\Delta v^{-1}$ from λ_j (i.e., it is a simplex which is a "truncated tip" of the simplex H at λ_j with diameter $k^{1/2}\Delta v^{-1}$). Since $\lambda_j^* \in M(N)$ may be selected in H_j^* for each j, $j = 1, 2, \cdots, k$, we conclude that $M(N)$ includes all positive linear combinations of λ_j^*. If $\{\mu_j\}_{j=1}^k$ is a sequence of points such that $\mu_j \in H_j^*$ and is otherwise arbitrary, we let $H\{\mu_j\}$ denote the convex hull of $\mu_1, \mu_2, \cdots, \mu_j$. Then $H^* \subset M(N)$, where

$$H^* = \bigcap_{\{\mu_j\}} H\{\mu_j\}.$$

This intersection H^* is seen to be the regular $(k-1)$-simplex embedded in H of all points of H whose distance from the boundary of O is $\geqq \Delta v^{-1}$; that is, a point $P = (p_1, \cdots, p_{\ })$ of O is in H^* if and only if $\min_{1 \leqq i \leqq k} |p_i| \geqq \Delta v^{-1}$. In particular, the minimum modulus of the coordinates of $v^{-1}\bar{Y}$ is Δv^{-1}. Hence, $v^{-1}\bar{Y} \in H^*$, $v^{-1}\bar{Y} \in M(N)$, $\bar{Y} \in M(N)$, and P is the desired polynomial satisfying (3).

We now note that as an immediate corollary of Newman's construction [2] there is a constant C such that for $n > k$

$$E_n^*(S_k) \leqq \frac{c}{n},$$

i.e., that

$$n_k(\eta) \leqq \frac{c}{n}. \tag{5}$$

Combining (3) and (5), we get (2). *Q.E.D.*

4. Remark. If in the construction of P satisfying (3) the normalization to the standard $(k-1)$-simplex (by multiplying Y and v^{-1}) is replaced by a normalization to

the k-dimensional unit cube (by multiplying Y by v_0^{-1} where $v_0 = \max_j |y_j - y_{j-1}|$), then the dependence on the total variation v is ostensibly replaced by a similar dependence on the smaller v_0. However, if this is done, a dependence on k is necessarily introduced, and the resulting estimate on the degree of P is of the order $kv_0/\Delta\delta$, which is clearly weaker than (3). Nevertheless, this estimate is a bit easier to compare to (1) and helps demonstrate the time advantages of the method introduced in this article.

REFERENCES

[1] G. ILIEV, *Exact estimation under the partially monotone approximation and interpolation*, C.R. Acad. Bulgare Sci., 30 (1977), pp. 491–494.

[2] D. J. NEWMAN, *Efficient co-monotone approximation*, J. Approx. Theory, 25 (1979), pp. 189–192.

[3] E. PASSOW AND L. RAYMON, *The degree of piecewise monotone interpolation*, Proc. Amer. Math. Soc., 48 (1975), pp. 409–412.

[4] W. WOLIBNER, *Sur un polynôme d'interpolation*, Colloq. Math., 2 (1951), pp. 136–137.

[5] S. W. YOUNG, *Piecewise monotone polynomial interpolation*, Bull. Amer. Math. Soc., 73 (1967), pp. 642–643.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

0036-1410/81/1201-0012 $01.00/0

THE CHARACTERISTIC-RESISTANCE METHOD FOR GROUNDED SEMI-INFINITE GRIDS*

A. H. ZEMANIAN†

Abstract. A variety of existence and uniqueness theorems for the current flows in infinite electrical networks have been previously established, but there is virtually no information on how to compute those current flows under the practical requirement that the power dissipated and energy stored in the network be finite. This problem is addressed herein. It is shown that the characteristic-impedance method of analyzing lumped transmission lines can be adapted to semi-infinite half-plane resistive grids and to three-dimensional half-space resistive grids. The method yields the one and only current flow within the grid for which the total power dissipation is finite. It also yields a practical procedure for computing the currents and voltages in the grid. That procedure is remarkably efficient and uses very little computer time. Moreover, semi-infinite grids whose branch impedances are positive real functions can also be analyzed in a similar way. This allows the computation of transient behavior in the presence of energy-storage elements, but the required computer time is now considerably longer. Nonlinear grids can also be encompassed by an approximation technique. These results appear to provide a basis for the numerical analysis of certain boundary value problems of practical importance in engineering and the physical science.

1. Introduction. Existence and uniqueness theorems for the current flows in infinite electrical networks have been the subject of rigorous analysis for almost a decade, starting with the seminal work of Flanders [4] and continuing with the efforts of Dolezal [2] and the author [21], [23]. However, except for some highly particular cases such as the lumped transmission line, there have been very few results concerning the actual computation of the current flows in an infinite electrical network under the practical requirement that the total power dissipated and the energy stored in the network be finite. A notable accomplishment in this regard is Flanders' justification [5] of the various analyses of the constant-resistance square grid. If the assumption of finite power or finite energy is dropped, a method [22], [23] becomes available, but it requires the a priori assignment of the currents in certain branches called "joints". How to compute the current flows to ensure finite power or finite energy remains an open question for general infinite electrical networks.

The principal result of this paper is that the characteristic-impedance method of transmission line theory can be adapted to half-plane resistive grids (Fig. 1) and to three-dimensional half-space resistive grids (Fig. 6). The method yields the one and only current flow within the grid for which the total power dissipation is finite. The proof of this assertion is rather long and appears in §§ 2 through 5. It also yields a practical procedure for computing the currents and voltages in the grid (§§ 6 and 7). That procedure is remarkably efficient and uses very little computer time. Moreover, our characteristic-resistance method extends to RLC grounded grids wherein each branch is a positive real impedance (§ 8). This allows the computation of transient currents and voltages in the presence of energy-storage elements, but the required computer time is considerably longer (§ 9). Nonlinear resistive grids can also be encompassed if a linearization technique due to Dolezal [3] is employed.

In addition, our results appear to be useful to the numerical analysis of certain boundary value problems of practical importance in engineering and the physical sciences. For example, a currently active area of research in semiconductor devices is

* Received by the editors January 8, 1980 and in revised form August 2, 1980. This work was supported by the U.S. Air Force Office of Scientific Research under Grant F49620-79-C-0172.

† Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, New York 11794.

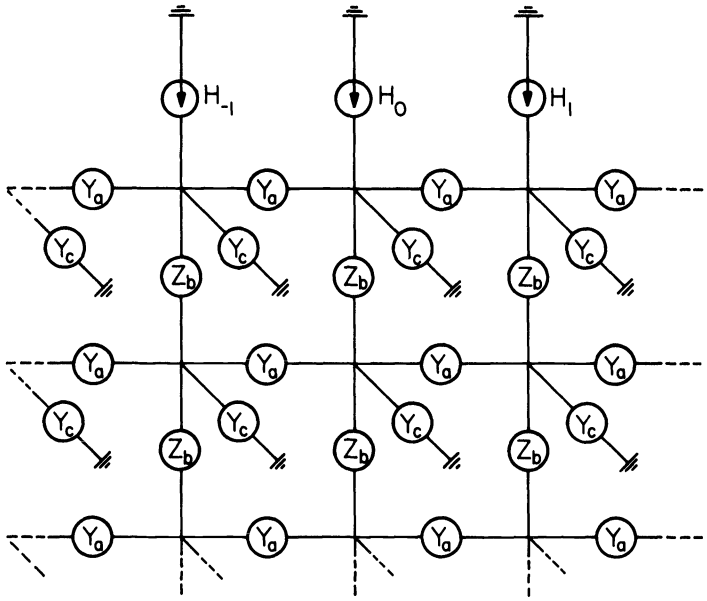


FIG. 1.

the numerical computation of the currents, potentials, and minority carrier densities in various doping configurations. More particularly, one configuration for an n-p-n lateral transistor is shown in Fig. 2 where we assume that there is no variation in the direction

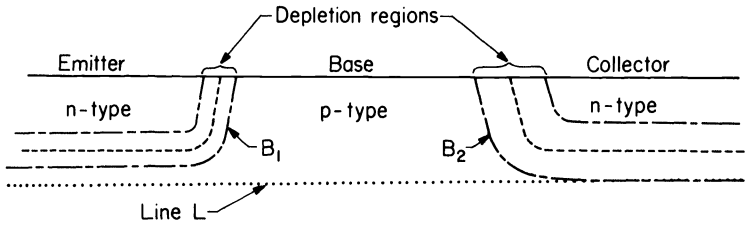


FIG. 2.

perpendicular to the plane of the page; that is, we assume a two-dimensional variation only. The two edges of the depletion regions bordering the p-type region are denoted by B_1 and B_2. Let $n = n_p - n_{p0}$, where n_p is the free-electron density in the p-type region and n_{p0} is the equilibrium value of n_p for the unbiased transistor. n varies in general with the spatial coordinates. If known biases are applied to the emitter and collector, then the values of n along B_1 and B_2 are easily determined [15]. If in addition, the locations of B_1 and B_2 can be estimated, then the following question arises: What are the values of n throughout the p-type region? n is governed by the partial differential equation [15, p. 99]:

$$\nabla^2 n = \frac{n}{D_n \tau},$$

where ∇^2 is the two-dimensional Laplacian operator, D_n is the electron diffusion constant, and τ is the electron (minority carrier) lifetime. A difference equation approximation of this equation yields the resistive grounded-grid representation,

shown in Fig. 1, for the p-type region with the node potentials being the discrete values of n and with given voltage sources at the nodes along B_1 and B_2 being the discrete boundary values for n.

But now we are faced with a dilemma. Ordinary circuit analysis does not permit us to compute the node voltages in the case where the p-type semiconductor of Fig. 2 extends infinitely downward. On the other hand, truncating that material at some vertical distance downward leads to analyses requiring impractically long computer times if the truncation occurs at realistic distances from the top surface. Truncations that are sufficiently close to the top surface to allow practicable computer times are too close to be representative of the actual configuration. This proximity of the truncation distorts the variations in n from their actual values. A possible alternative, one that would not introduce unacceptable distortions, would be to allow an infinitely deep configuration in Fig. 2.

Here is where our method for analyzing the semi-infinite grid becomes useful. We can represent all the material below a horizontal line L passing along the lowest horizontal portions of B_1 and B_2 by an operator-valued characteristic resistance and then analyze the p-type region above that line L. This determines n in that region and along L. Then our characteristic-resistance technique can be used again to determine n below the line L with comparatively little additional computer time. This numerical technique will be the subject of a subsequent paper.

Heat flow in thin films is another physical situation for which our characteristic-resistance method provides a numerical analysis [24]. Still other physical phenomena whose difference equation approximations lead to uniform grounded grids are discussed in the many papers of Kron; see, for example, [8] and [9]. These latter cases lead to RLC grounded grids, rather than purely resistive ones. Nevertheless, our method extends to this case too, as was mentioned before.

Our method also works for networks with more complicated graphs than the one shown in Fig. 1. We can allow each node of that figure to be adjacent to more than two nodes in the same horizontal row. We need only require that each node have a finite degree and that the grid remain uniform. (See [18, Fig. 5].) Moreover, the element values and the graph can vary from horizontal row to horizontal row and even in the values of Z_b so long as some periodicity in these variables occurs along the vertical direction and there is no variation in the horizontal direction. In short, the grounded-grid computations established in this work may become of considerable value in the numerical analysis of certain boundary value problems over half-planes or three-dimensional half-volumes, a matter that is currently under investigation.

2. An existence and uniqueness theorem. We will need an extension of the existence and uniqueness theorem of [19] to the case where the currents and voltages are Hilbert-space-valued and the branch conductances are positive invertible operators. In this section, we will state what alterations are needed and then will prove a simplified version of that theorem that is appropriate to the grid of Fig. 1. We continue to use the terminology of algebraic topology as introduced into the subject of infinite electrical networks by Flanders [4]. In the following, H_r denotes a real Hilbert space with the inner product $(\cdot, \cdot)$.

Conditions A. Let N be a connected countably infinite electrical network having no self loops. The currents and voltages of N are members of H_r. Each branch $\mathbf{B}_j$ of N is a parallel connection of a (possibly zero) current source $h_j \in H_r$ and a (nonzero) conductance g_j which is a positive invertible operator mapping H_r into H_r. There are no other current sources and no voltage sources.

$\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \cdots$ denote the branches in N. The typical branch $\mathbf{B}_j$, which we take to be oriented, is illustrated in Fig. 3. The branch voltage (drop) $v_j \in H_r$ and the branch

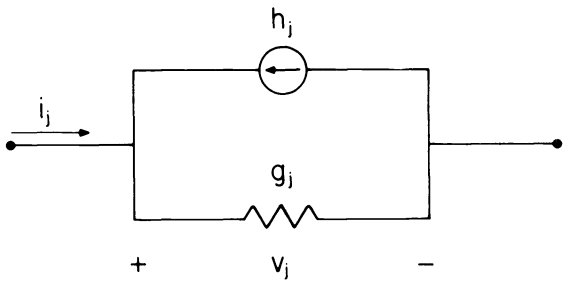


FIG. 3.

current $i_j \in H_r$ are measured with respect to the orientation of $\mathbf{B}_j$. Thus, $i_j = g_j v_j - h_j$, $\mathbf{i} = \sum i_j \mathbf{B}_j$ is the 1-chain of branch currents, $\mathbf{v} = \sum v_j \mathbf{B}_j'$ is the 1-cochain of branch voltages, and $\mathbf{h} = \sum h_j \mathbf{B}_j$ is the 1-chain of current sources. Kirchhoff's node law states that $\mathbf{i}$ is a cycle, and his loop law states that $\mathbf{v}$ is a coboundary. At this point, nothing is gained by requiring that N be locally finite. This will be required later on, the ground node excepted.

A 1-cochain $\mathbf{w}' = \sum w_j \mathbf{B}_j'$, where $w_j \in H_r$, is defined as a functional on a 1-chain $\mathbf{x} = \sum x_j \mathbf{B}_j$, where $x_j \in H_r$, by $\langle \mathbf{w}', \mathbf{x} \rangle = \sum (w_j, x_j)$ whenever $\sum (w_j, x_j)$ exists. The latter will certainly be the case when $\mathbf{x}$ is a finite 1-chain (i.e., when all but a finite number of the x_j are zero). We now let $\mathcal{V}$ be the Hilbert space of all coboundaries $\mathbf{v}' = \sum v_j \mathbf{B}_j'$ such that

$$\sum (v_j, g_j v_j) < \infty. \tag{2.1}$$

The inner product of two coboundaries $\mathbf{v}'$ and $\mathbf{w}'$ in $\mathcal{V}$ is defined to be $\sum (v_j, g_j w_j)$. Thus, the norm of $\mathbf{v}'$ is

$$\|\mathbf{v}'\| = [\sum (v_j, g_j v_j)]^{1/2}.$$

Finally, we define the conductance operator G of N to be the mapping of any 1-cochain $\mathbf{v}' = \sum v_j \mathbf{B}_j'$ into the 1-chain $G\mathbf{v}' = \sum g_j v_j \mathbf{B}_j$.

All the arguments of [19] carry over to this case of an operator network. They establish the following existence and uniqueness theorem.

THEOREM 2.1. *Let N satisfy Conditions* A, *and let its branch parameters satisfy*

$$\sum (g_j^{-1} h_j, h_j) < \infty. \tag{2.2}$$

Then there exists a unique $\mathbf{v}' \in \mathcal{V}$ such that

$$\langle \mathbf{w}', \mathbf{h} - G\mathbf{v}' \rangle = \mathbf{0}, \tag{2.3}$$

for all $\mathbf{w}' \in \mathcal{V}$.

This theorem states in effect that four conditions determine a unique set of branch voltages: Kirchhoff's loop law ($\mathbf{v}'$ is a coboundary), the finite-power dissipation condition (2.1), the finite-power-available condition (2.2), and a generalized form of Tellegen's theorem (2.3), which encompasses Kirchhoff's node law and Ohm's law as consequences.

Our infinite grids possess two more properties that will be worth exploiting. They lead to a special case of Theorem 2.1 (namely, Theorem 2.2 below) that is more convenient for our purposes than the global condition (2.3).

Conditions B. (i) N is locally finite except possibly for a single ground node, which may be of infinite degree.

(ii) There are two positive numbers P and Q such that $\|g_j\| \leqq P$ and $\|g_j^{-1}\| \leqq Q$ for all j.

It follows from B(ii) and the spectral mapping theorem for positive operators that $\|g_j^{1/2}\| \leqq P^{1/2}$ and $\|g_j^{-1/2}\| \leqq Q^{1/2}$.

We can now identify $\mathscr{V}$ with a certain subspace of all coboundaries as follows.

LEMMA 2.1. *Conditions* B(ii) *and the positivity and invertibility of each* g_j *imply that* $\mathbf{v}' = \sum v_j \mathbf{B}_j' \in \mathscr{V}$ *if and only if the* v_j *satisfy Kirchhoff's loop law and* $\sum \|v_j\|^2 < \infty$.

Proof. Kirchhoff's loop law is equivalent to the assertion that $\mathbf{v}'$ is a coboundary. Also,

$$\sum (v_j, g_j v_j) \leqq \sum \|g_j\| \, \|v_j\|^2 \leqq P \sum \|v_j\|^2.$$

Conversely, since g_j is positive and invertible, it has a square root $g_j^{1/2}$ with the same properties. Then

$$\sum \|v_j\|^2 = \sum \|g_j^{-1/2} g_j^{1/2} v_j\|^2 \leqq Q \sum \|g_j^{1/2} v_j\|^2 = Q \sum (v_j, g_j v_j).$$

This proves the lemma.

In the following, $l_2(H_r)$ will denote the space of all sequences $a = \{a_j : j = 1, 2, 3 \cdots\}$, where $a_j \in H_r$ and $\sum \|a_j\|^2 < \infty$, with the inner product $(a, b) = \sum (a_j, b_j)$.

LEMMA 2.2. *Under Conditions* A *and* B, (2.3) *is a consequence of Kirchhoff's node and loop laws, Ohm's law, and the finite-power conditions* (2.1) *and* (2.2).

Proof. Kirchhoff's loop law and (2.1) assert that $v' = \sum v_j \mathbf{B}_j'$ is a member of $\mathscr{V}$. We now show that, for every $\mathbf{w}' \in \mathscr{V}$, the series

$$\langle \mathbf{w}', \mathbf{h} - G\mathbf{v}' \rangle = \sum (w_j, h_j - g_j v_j), \tag{2.4}$$

converges absolutely. By Schwarz's inequality,

$$\begin{aligned}
\sum |(w_j, h_j - g_j v_j)| &\leqq \sum |(w_j, h_j)| + \sum |(w_j, g_j v_j)| \\
&= \sum |(g_j^{1/2} w_j, g_j^{-1/2} h_j)| + \sum |(g_j^{1/2} w_j, g_j^{1/2} v_j)| \\
&\leqq \sum \|g_j^{1/2} w_j\| \, \|g_j^{-1/2} h_j\| + \sum \|g_j^{1/2} w_j\| \, \|g_j^{1/2} v_j\| \\
&\leqq [\sum (w_j, g_j w_j) \sum (g_j^{-1} h_j, h_j)]^{1/2} + [\sum (w_j, g_j w_j) \sum (v_j, g_j v_j)]^{1/2}.
\end{aligned}$$

All four series in the last expression converge by virtue of (2.1) and (2.2). This verifies our assertion concerning (2:4). Consequently, we may rearrange and sum (2.4) in any fashion.

So, let n_0 denote the ground node of possibly infinite degree. Let n_k, where $k = 1, 2, 3, \cdots$, denote all the other nodes. Let W_k be the node voltages measured with respect to n_0 and resulting from the branch voltages w_j of $\mathbf{w}' = \sum w_j \mathbf{B}_j' \in \mathscr{V}$. Consider any node n_k other then n_0. Let B_{j_m}, where $m = 1, \cdots, p$, be the branches incident to n_k. Corresponding to each B_{j_m}, there is a term $\pm (W_k - W_q, h_j - g_j v_j)$ in the right-hand side of (2.4), where W_k and W_q are the voltages at the nodes of B_{j_m}; q is a nonnegative integer. We have $W_0 = 0$. The $+(-)$ sign is chosen if B_{j_m} is oriented away from (toward) n_k. Now, gather all the terms in (2.4) having the factor W_k; we get

$$\sum_{m=1}^{p} \pm (W_k, h_{j_m} - g_{j_m} v_{j_m}). \tag{2.5}$$

By Kirchhoff's node law, (2.5) equals zero. (2.4) may be rearranged to get a sum of

terms like (2.5), with one such term for each node other than n_0. We can thus conclude that (2.4) equals zero; that is, (2.3) holds. This completes the proof.

THEOREM 2.2. *Let N satisfy Conditions* A *and* B. *Assume that* $\{h_j\} \in l_2(H_r)$. *Then, there exists a unique vector* $v \in l_2(H_r)$ *of branch voltages* v_j *such that Kirchhoff's node and loop laws and Ohm's law are satisfied. Those branch voltages* v_j *are identical to the branch voltages dictated by Theorem* 2.1.

Proof. By Condition B(ii), $g_j^{-1} \leqq Q$ for all j. Therefore, with regard to the current sources, we may write

$$\infty > Q\sum\|h_j\|^2 \geqq \sum\|g_j^{-1}\|\|h_j\|^2 \geqq \sum(g_j^{-1}h_j, h_j).$$

So (2.2) is satisfied. By Theorem 2.1, there is a $\mathbf{v}' \in \mathcal{V}$ satisfying (2.3). Lemma 2.1 now implies that Kirchhoff's loop law is satisfied and that the vector $\mathbf{v}$ of branch voltages v_j is in $l_2(H_r)$. (2.3) implies that Kirchhoff's node law and Ohm's law are satisfied. There cannot be two such voltage vectors in $l_2(H_r)$, for, if there were, then by Lemmas 2.1 and 2.2 the corresponding members of $\mathcal{V}$ would both satisfy (2.3), in contradiction to the uniqueness assertion of Theorem 2.1. This completes the proof.

As an application of these results, we can specify a unique solution for the half-plane grid of positive resistances indicated in Fig. 1. First note that that grid satisfies Conditions A and B. We now let l_{2r} denote Hilbert's coordinate space of doubly infinite vectors; that is, the vector $[\cdots, a_{-1}, a_0, a_1, \cdots]^T$ (the superscript T denotes matrix transpose) is a member of l_{2r} if all its entries a_j are real numbers and $\sum a_j^2 < \infty$. Let us assume that the vector of current sources in Fig. 1 is a member of l_{2r}:

$$H = [\cdots, H_{-1}, H_0, H_1, \cdots]^T \in l_{2r}. \tag{2.6}$$

Since these are the only sources in Fig. 1, the quantity $\sum(g_j^{-1}h_j, h_j)$ in Theorem 2.1 becomes $Y_c^{-1}\sum_{k=-\infty}^{\infty} H_k^2$, which is finite by virtue of (2.6). Upon numbering the branches of the entire grid appropriately, we can conclude from Theorem 2.2 that there is a unique branch-voltage vector in l_{2r} for which Kirchhoff's node and loop laws and Ohm's law are satisfied. This branch-voltage vector determines and is determined by a unique node-voltage vector in l_{2r}.

3. A ladder network of Hilbert ports. Our objective is to derive a set of equations from which the currents and voltages in the infinite grid of Fig. 1 can be numerically computed. This will be accomplished by replacing the infinite grid of Fig. 1 by a ladder network each of whose branches is a Hilbert port [17] with respect to l_{2r} obtained from a subnetwork of the grid. We need, in fact, only two different Hilbert ports, those shown in Figs. 4(b) and 4(c). Fig. 4(a) is a vector-valued current source whose value is given by (2.6). As we shall see, the Hilbert port admittance Y of Fig. 4(b) and the Hilbert-port impedance Z of Fig. 4(c) are continuous linear mappings of l_{2r} into l_{2r} given by certain Laurent matrices. A Laurent matrix is an infinite matrix of the form $A = [A_{jk}]$, where $j, k = \cdots, -1, 0, 1, \cdots$, such that $A_{j,k} = A_{j+1,k+1}$ for all j, k [1], [6, p. 135]. Thus, all its rows are the same except for horizontal shifts. To specify a Laurent matrix, it suffices to specify its row for $j = 0$; we call that row the principal row and denote it by

$$A^+ = [\cdots, \quad A_{0,-1}, \quad (A_{0,0}), \quad A_{0,1} \cdots]$$

where the parentheses are used to identify the $j = 0$, $k = 0$ entry in A.

Because of the disconnected form of the Hilbert port of Fig. 4(c), only one current vector i can respond when a voltage-source vector in l_{2r} is connected to that Hilbert port. Moreover, i will also be in l_{2r} because of the constant value of Z_b of the impedance therein. Thus, the impedance Z of that Hilbert port is truly a Laurent matrix, in fact, a

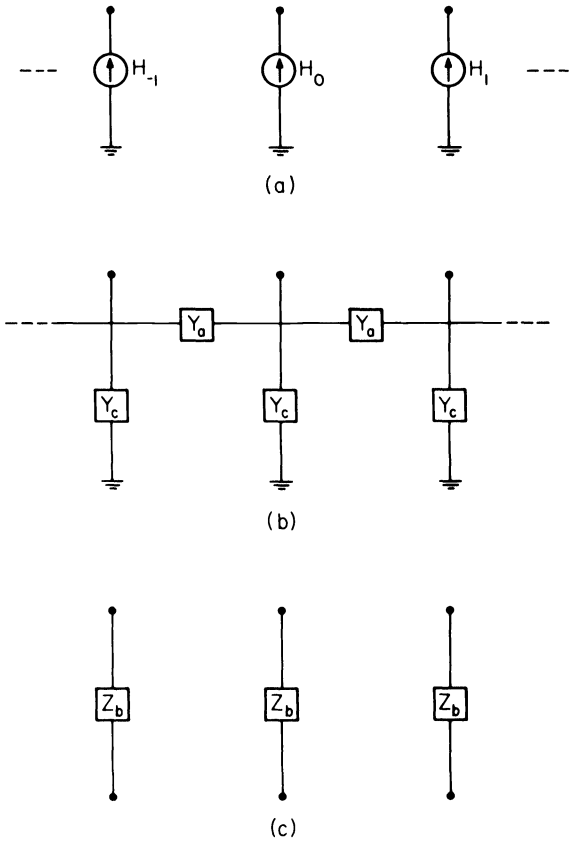


FIG. 4.

diagonal matrix with the principal row

$$Z^{+}=[\cdots, 0, \quad 0, \quad (Z_b), \quad 0, \quad 0, \quad \cdots]. \tag{3.1}$$

On the other hand, many different responding (node) voltage vectors are possible when a current-source vector i in l_{2r} is impressed upon the Hilbert port of Fig. 4(b). However, since Fig. 4(b) with that current source satisfies Conditions A and B, it follows from Theorem 2.2 that only one of those vectors, say, v will be in l_{2r}. A nodal analysis (that is, the application of Kirchhoff's node law at every finite node) indicates that the admittance Y of the Hilbert port that maps $v \in l_{2r}$ into $i \in l_{2r}$ is the Laurent matrix whose principal row is

$$Y^{+}=[\cdots, \quad 0, \quad 0, \quad -Y_a, \quad (Y_c+2Y_a), \quad -Y_a, \quad 0, \quad 0, \quad \cdots]. \tag{3.2}$$

To verify that Y yields the correct behavior dictated by Theorem 2.2, we need merely show that Y is invertible on l_{2r}. It truly is: Y^{-1} is the Laurent matrix whose principal row is

$$(Y^{-1})^{+}=\frac{1}{Y_a\sqrt{\alpha^2-4}}[\cdots, \quad \lambda^3, \quad \lambda^2, \quad \lambda, \quad (1), \quad \lambda, \quad \lambda^2, \quad \lambda^3, \quad \cdots],$$

where

$$\alpha=\frac{Y_c}{Y_a}+2, \qquad \alpha>2,$$

and

$$\lambda = \tfrac{1}{2}(\alpha - \sqrt{\alpha^2 - 4}), \qquad 0 < \lambda < 1.$$

Since $\alpha > 2$, $0 < \lambda < 1$. That this expression for Y^{-1} is truly the inverse of Y can be verified by a direct multiplication of Y and Y^{-1}.

Because of the common ground and the disconnected form of Fig. 4(c), we can connect these Hilbert ports together, without violating the port conditions, to get the network of Fig. 1. The resulting ladder network of Hilbert ports is shown in Fig. 5. The

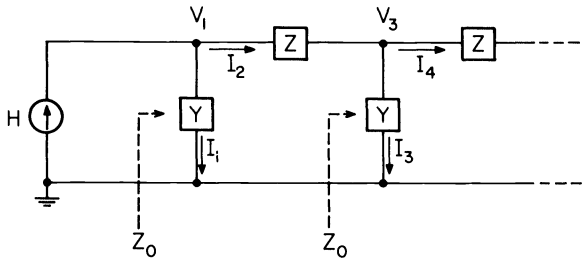


FIG. 5.

source current H, the node voltages V_1, V_3, V_5, $\cdots$ (measured with respect to ground), and the branch currents I_1, I_2, I_3, $\cdots$ are each members of l_{2r}. We can invoke Theorem 2.2, this time in its operator-network form, to conclude that there exists a unique set of node voltages, and thereby branch currents as well, dictated by the conditions of that theorem. We shall prove subsequently that the components of these l_{2r}-valued voltages and currents are identical to the voltages and currents of Fig. 1 dictated by the scalar version of Theorem 2.2. But, to do so, we will first develop some expressions for the voltages and currents of Fig. 5 based upon the characteristic-resistance approach to periodic structures.

4. The characteristic-resistance method. We wish to determine the l_{2r}-valued currents and voltages in the ladder network of Fig. 5, where H is a given l_{2r}-valued current source. We shall use the standard characteristic-resistance method for periodic transmission lines, which now requires our extension of it to the case where the line's parameters are operators rather than scalars.

We will show later on that the characteristic-resistance operator Z_0 indicated in Fig. 5 is invertible. Assume this to be true for the moment and set $Y_0 = Z_0^{-1}$. Then, by the usual argument that the characteristic resistance is not altered if a single Y and a single Z are removed from the beginning of the ladder structure, we have

$$(Y_0 - Y)(Z_0 + Z) = 1. \tag{4.1}$$

Here, 1 denotes the identity operator on l_{2r}.

This equation has more than one solution Z_0. Let us argue heuristically for the moment to see which solution we should seek. Z_0 should describe the behavior of the network of Fig. 1. By the symmetry of that network, a shift of the set of current sources H_k to the right or left should result in the same shift in the responding node voltages. That is, Z_0 should commute with the bilateral shift. This means that Z_0 should be a Laurent matrix [1, Theorem 2], [6, p. 135]. Moreover, the solution dictated by Theorem 2.1 disallows any energy being injected into the network from infinity. Moreover, the ladder network is passive. This suggests that Z_0 should be a positive operator.

To find such a Z_0, we shall make use of the natural isomorphism between l_{2r} and the space $L_2(0, 2\pi)$ of (equivalence classes of) quadratically integrable functions on the unit circle. Specifically, $a = \{a_k\}_{k=-\infty}^{\infty} \in l_{2r}$ corresponds to the Fourier series

$$\tilde{a}(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx},$$

where, convergence in $L_2(0, 2\pi)$ is understood.

Now, a Laurent matrix L defines a continuous linear mapping of l_{2r} into l_{2r} if and only if its principal row

$$L^{\dagger} = [\cdots, \quad l_{-1}, \quad (l_0), \quad l_1, \quad \cdots],$$

is such that

$$\tilde{L}(x) = \sum_{k=-\infty}^{\infty} l_k e^{-ikx},$$

is an essentially bounded integrable function of x [1, Theorem 3], [6, p. 135]. Such a Laurent matrix L acting on $a \in l_{2r}$ corresponds to the multiplication of $\tilde{a}(x)$ by the function $\tilde{L}(x)$. In addition, L is positive and invertible if and only if $\tilde{L}(x)$ is positive and uniformly bounded away from zero for almost all $x \in (0, 2\pi)$. Note that, because of the negative sign in the exponents of our expression for $\tilde{L}(x)$, the Fourier coefficients of $\tilde{L}(x)$ are the entries of $L^{\dagger}$ taken in reverse order. Of course, if $L^{\dagger}$ has even symmetry around its central entry, the order of those entries need not be reversed in order to get the Fourier coefficients of $\tilde{L}(x)$.

All of the aforementioned restrictions on L are satisfied by Y, Y^{-1}, and Z. Under the stated isomorphism, (4.1) transforms into

$$[\tilde{Y}_0(x) - \tilde{Y}(x)][\tilde{Z}_0(x) + \tilde{Z}(x)] = 1, \tag{4.2}$$

where by (3.1) and (3.2)

$$\tilde{Z}(x) = Z_b,$$

$$\tilde{Y}(x) = Y_c + 2\, Y_a(1 - \cos x) = Y_a(\alpha - 2\cos x),$$

$$\alpha = \frac{Y_c}{Y_a} + 2.$$

In view of the remarks of the preceding paragraph, we seek the positive solution $\tilde{Z}_0(x)$ of (4.2) where $\tilde{Y}_0(x) = \tilde{Z}_0(x)^{-1}$. It is

$$\tilde{Z}_0(x) = \frac{-\beta(\alpha - 2\cos x) + [\beta^2(\alpha - 2\cos x)^2 + 4\beta(\alpha - 2\cos x)]^{1/2}}{2\, Y_a(\alpha - 2\cos x)}, \tag{4.3}$$

where $\beta = Y_a Z_b$. Since $\alpha > 2$, $\tilde{Z}_0(x)$ is clearly positive, even, continuous, bounded, and bounded away from zero for all x. (The other root of (4.2) is negative for all x). Moreover, $\tilde{Y}_0(x)$ is given by

$$\tilde{Y}_0(x) = \frac{\beta(\alpha - 2\cos x) + [\beta^2(\alpha - 2\cos x)^2 + 4\beta(\alpha - 2\cos x)]^{1/2}}{2\, Z_b} \tag{4.4}$$

and has the same properties. Thus, Z_0 and Y_0 are the Laurent matrices whose principal rows are the Fourier coefficients of the functions $\tilde{Z}_0(x)$ and $\tilde{Y}_0(x)$, respectively. Therefore, upon reverting to the Laurent-matrix expressions corresponding to (4.3)

and (4.4), we obtain

$$Z_0 = (2Y)^{-1}[-YZ + (Y^2Z^2 + 4YZ)^{1/2}], \tag{4.5}$$

and

$$Y_0 = (2Z)^{-1}[YZ + (Y^2Z^2 + 4YZ)^{1/2}]. \tag{4.6}$$

By the aforementioned isomorphism, Z_0 and Y_0 are positive invertible operators on l_{2r}.

Laurent matrices commute. Also, the square root and the inverse of a positive invertible operator A commute with every operator that commutes with A. These facts allow us to manipulate Y and Z in much the same way as one might manipulate scalars. It follows from Fig. 5 that

$$V_1 = Z_0 H. \tag{4.7}$$

Thus,

$$I_2 = H - I_1 = H - YZ_0H = (1 - YZ_0)H.$$

Alternatively,

$$I_2 = (Z_0 + Z)^{-1}V_1 = (Z_0 + Z)^{-1}Z_0H = (1 + Y_0Z)^{-1}H.$$

We set

$$\theta = 1 - YZ_0 = (Z_0 + Z)^{-1}Z_0 = (1 + Y_0Z)^{-1}. \tag{4.8}$$

In general, for $n = 1, 2, 3, \cdots$ and $I_0 = H$,

$$I_{2n} = (Z_0 + Z)^{-1}V_{2n-1} = (Z_0 + Z)^{-1}Z_0I_{2n-2} = \theta I_{2n-2},$$

and

$$V_{2n+1} = Z_0I_{2n} = Z_0(Z_0 + Z)^{-1}V_{2n-1} = \theta V_{2n-1}.$$

By induction,

$$I_{2n} = \theta^n H, \tag{4.9}$$

$$V_{2n+1} = \theta^n V_1. \tag{4.10}$$

Equations (4.5) through (4.10) coupled with $I_{2n-1} = YV_{2n-1}$ determine all the branch currents and node voltages in Fig. 5 for a given $H \in l_{2r}$.

LEMMA 4.1. *θ is a positive, invertible, strongly contractive operator. Thus, there exist numbers ε and δ such that $0 < \varepsilon \leqq \delta < 1$ and $\varepsilon\|a\|^2 \leqq (\theta a, a) \leqq \delta\|a\|^2$ for every $a \in l_{2r}$. Moreover, $\|\theta\| \leqq \delta$.*

Proof. Z_0, Z, and $Z_0 + Z$ are positive and invertible. They also commute. Therefore, $\theta = Z_0(Z_0 + Z)^{-1}$ and $Z(Z_0 + Z)^{-1}$ are positive and invertible too. Hence, $(\theta a, a) \geqq \varepsilon\|a\|^2$ for some $\varepsilon > 0$ and $(Z(Z_0 + Z)^{-1}a, a) \geqq (1 - \delta)\|a\|^2$ for some $\delta < 1$. Moreover,

$$\begin{aligned}(\theta a, a) &= (Z_0(Z_0 + Z)^{-1}a, a)\\ &= ((-Z + Z + Z_0)(Z_0 + Z)^{-1}a, a)\\ &= -(Z(Z_0 + Z)^{-1}a, a) + \|a\|^2\\ &\leqq -(1 - \delta)\|a\|^2 + \|a\|^2\\ &= \delta\|a\|^2.\end{aligned}$$

Finally, the norm of a positive operator equals its numerical radius [11, p. 145], and therefore, $\|\theta\| \leqq \delta$.

This lemma allows us to prove

THEOREM 4.1. *The solution for the network of Fig.* 5 *given by* (4.5) *through* (4.10) *is precisely the solution dictated by Theorem* 2.2.

Proof. We shall show that the solution given by (4.5) through (4.10) satisfies the conditions of Theorem 2.2. First note that the network of Fig. 5 satisfies Conditions A and B. Number and orient the branches in accordance with the currents in Fig. 5 and let v_j be the branch voltages (i.e., voltage drops). Since the V_{2n+1} are node voltages, Kirchhoff's loop law is automatically satisfied by the corresponding v_j. The branch currents, as determined by Ohm's law in its operator form, satisfy Kirchhoff's node law. Indeed, for every n,

$$I_{2n+2} + I_{2n+1} = \theta^{n+1}H + YV_{2n+1} = \theta^{n+1}H + Y\theta^n V_1$$
$$= \theta^n(\theta H + YZ_0 H) = \theta^n(\theta + YZ_0)H.$$

By (4.8), $\theta + YZ_0 = 1$. Hence,

$$I_{2n+2} + I_{2n+1} = \theta^n H = I_{2n},$$

which fulfills Kirchhoff's node law.

Finally, we have to show that the vector of branch voltages is in $l_2(H_r)$, where $H_r = l_{2r}$. For the vertical branches in Fig. 5, we may invoke Lemma 4.1 to write

$$\sum_{n=0}^{\infty} \|v_{2n+1}\|^2 \leqq \sum_{n=0}^{\infty} \|\theta^n V_1\|^2 \leqq \|V_1\|^2 \sum_{n=0}^{\infty} \delta^{2n}.$$

The right-hand side is a finite quantity because $0 < \delta < 1$. Similarly, for the horizontal branches of Fig. 5 we have

$$\sum_{n=1}^{\infty} \|ZI_{2n}\|^2 \leqq \|Z\|^2 \sum_{n=1}^{\infty} \|\theta^n H\|^2 \leqq \|Z\|^2 \|H\|^2 \sum_{n=1}^{\infty} \delta^{2n} < \infty.$$

This completes the proof.

It is worth noting at this point a commonly occurring lacuna in the many expositions of the characteristic-resistance method for infinite resistive ladder networks. Namely, the total power is tacitly assumed to be finite; that this assumption eliminates all but the one solution dictated by the characteristic-resistance method is not obvious. Theorem 4.1 shows that this is so. Theorem 8.3 below implies a similar result for infinite ladder networks of impedances.

5. Verification of the two-step procedure. So far, we have developed a two-step procedure for determining all the currents and voltages in the half-plane grid of Fig. 1. First, we decomposed that grid into a ladder network of Hilbert ports (Fig. 5) whose port-voltage and port-current vectors are members of l_{2r}, and noted that within each Hilbert port the voltages and currents are those indicated by Theorem 2.2. The second step was to determine the l_{2r}-valued voltages and currents in that ladder network by using the characteristic-resistance method. These vector-valued voltages and currents were shown by Theorem 4.1 to be identical to the voltages and currents dictated by Theorem 2.2. Once l_{2r}-valued port-voltage and port-current vectors are so obtained, the scalar voltages and currents within each Hilbert port, and thereby on each branch in Fig. 1, are immediately determined.

But a lacuna remains. We have yet to show that the voltages and currents of the grid determined by this two-step procedure are identical to the voltages and currents dictated by Theorem 2.2. This is asserted by

THEOREM 5.1. *The voltages and currents in the grid of Fig.* 1 *determined from the components of the voltage and current vectors specified by equations* (4.5) *through* (4.10) *and the given vector* $H = \{H_k\}_{k=-\infty}^{\infty} \in l_{2r}$ *are identical to the voltages and currents dictated by Theorem* 2.2.

The proof of this theorem is quite straightforward and is therefore omitted. Its details are given in [24; § 5].

6. The computation of the voltages and currents in the grid. We can use the Fourier series representation of the members of l_{2r} to convert the equations of §4 into a form that is convenient for numerical computation. We have already noted that the periodic function $\tilde{Z}_0$ corresponding to the Laurent matrix Z_0 is given by (4.3). Similarly, we can use (4.8) to compute the periodic function $\tilde{\theta}$, which is the multiplier corresponding to the Laurent matrix θ. We obtain

$$\tilde{\theta}(x) = 1 + \tfrac{1}{2}\beta(\alpha - 2\cos x) - \tfrac{1}{2}[\beta^2(\alpha - 2\cos x)^2 + 4\beta(\alpha - 2\cos x)]^{1/2}. \tag{6.1}$$

Moreover, (4.7) and (4.10) transform into

$$\tilde{V}_1(x) = \tilde{Z}_0(x)\tilde{H}(x), \tag{6.2}$$

and

$$\tilde{V}_{2n+1}(x) = [\tilde{\theta}(x)]^n \tilde{V}_1(x), \qquad n = 1, 2, 3, \cdots, \tag{6.3}$$

where

$$\tilde{H}(x) = \sum_{n=-\infty}^{\infty} H_n e^{inx}.$$

So, given the current sources H_n of Fig. 1 such that $\{H_n\}_{n=-\infty}^{\infty} \in l_{2r}$, we immediately have the functions $\tilde{V}_{2n+1}(x)$, where $n = 0, 1, 2, \cdots$. It follows then that the Fourier coefficients of $\tilde{V}_1(x)$,

$$\frac{1}{2\pi}\int_0^{2\pi} \tilde{V}_1(x)e^{-inx}dx,$$

are the node voltages of the nodes in the uppermost horizontal row of Fig. 1, when the nodes are numbered in accordance with the H_n. Similarly, the Fourier coefficients of the functions $\tilde{V}_{2n+1}(x)$ are the node voltages for the nodes in the $(n+1)$st horizontal row from the top of Fig. 1, when the same numbering system is used. Once these node voltages are determined, all the currents and voltages in the grid of Fig. 1 can be easily computed (except of course for the fact that limited computer time limits the number of currents and voltages one can compute). According to Theorem 5.1, these are precisely the currents and voltages specified by Theorem 2.2.

An alternative computation is the following. We need only compute the Fourier coefficients of (6.2) to get the node voltages in the uppermost row of nodes in Fig. 1. Then, by successively applying Kirchhoff's node and loop laws and Ohm's law, we can compute any desired branch voltages or branch current (within the capabilities of our computer budget) by progressing downward from row to row in Fig. 1. Actually, this amounts to the use of the first row of vertical branches (i.e., the parallel connections of the H_n and Y_c) as the "joints" of a limb analysis [22]. Our present analysis has shown

how those joint voltages must be assigned in order to obtain the finite-power-dissipation condition, a heretofore open problem.

Here are some numerical examples regarding the grid of Fig. 1. As before, V_{2n+1} denotes the vector of node voltages in the $(n+1)$st horizontal row of nodes for $n = 0, 1, 2, \cdots$. We display the 0th-indexed entry in each vector with parentheses. Because each of these vectors has even symmetry around the 0th-indexed entry, we need merely give their entries for nonnegative indices.

Example 6.1. Let $Y_a = 1$, $Y_c = 1$, and $Z_b = 1$. Also, let $H = [\cdots, 0, 0, (1), 0, 0, \cdots]^T$. Therefore, $\tilde{H}(x) = 1$. Moreover, $\alpha = 3$, $\beta = 1$. Then $Z_0(x)$ is given by (4.3), and $\tilde{\theta}(x)$ by (6.1). Upon computing the Fourier coefficients of (6.2) and (6.3) we obtain all the V_{2n+1} for $n = 0, 1, 2, \cdots$. Ohm's law then yields all the branch currents. Some results for the V_{2n+1} are:

$$
\begin{aligned}
V_1 &= [\cdots, (.3216), .0996, .0322, .0108, .0037, .0013, \cdots]^T,\\
V_3 &= [\cdots, (.0873), .0445, .0185, .0072, .0027, .0010, \cdots]^T\\
V_5 &= [\cdots, (.0259), .0170, .0086, .0038, .0016, .0006, \cdots]^T\\
V_7 &= [\cdots, (.0082), .0062, .0036, .0018, .0008, .0004, \cdots]^T,\\
V_9 &= [\cdots, (.0027), .0022, .0014, .0008, .0004, .0002, \cdots]^T,\\
V_{11} &= [\cdots, (.0009), .0007, .0005, .0003, .0002, .0001, \cdots]^T.
\end{aligned}
$$

Execution time on the computer was 2.278 seconds.

Example 6.2. Now set $Y_a = 1$, $Y_c = 2$, $Z_b = 10$ and $\tilde{H}(x) = 1$. Since the conductances Y_c to ground and the resistances Z_b are now larger, we should expect a more rapid decay in the node voltages as we progress into the grid. This is substantiated by the following results:

$$
\begin{aligned}
V_1 &= [\cdots, (.2797), .0730, .0190, .0050, .0013, .0003, \cdots]^T\\
V_3 &= [\cdots, (.0087), .0042, .0016, .0005, .0002, .0001, \cdots]^T,\\
V_5 &= [\cdots, (.0003), .0002, .0001, .0000, .0000, \cdots]^T,\\
V_7 &= [\cdots, (.0000), .0000, \cdots]^T.
\end{aligned}
$$

Computer execution time was 1.993 seconds.

Example 6.3. This time set $Y_a = 1$, $Y_c = 1$, $Z_b = 1$, and $H = [\cdots, 0, 0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, (1), \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, 0, 0, \cdots]^T$. Therefore, $\tilde{H}(x) = 1 + \cos x + \frac{1}{2}\cos 2x + \frac{1}{4}\cos 3x$. We now get

$$
\begin{aligned}
V_1 &= [\cdots, (.4400), .3086, .1813, .0942, .0303, .0101, .0035, \cdots]^T,\\
V_3 &= [\cdots, (.1428), .1129, .0725, .0401, .0171, .0067, .0025, \cdots]^T,\\
V_5 &= [\cdots, (.0428), .0408, .0281, .0166, .0082, .0036, .0015, \cdots]^T,\\
V_7 &= [\cdots, (.0167), .0147, .0107, .0067, .0037, .0018, .0008, \cdots]^T,\\
V_9 &= [\cdots, (.0059), .0053, .0040, .0027, .0016, .0008, .0004, \cdots]^T,\\
V_{11} &= [\cdots, (.0021), .0019, .0015, .0009, .0006, .0003, .0002, \cdots]^T.
\end{aligned}
$$

Computer execution time was 3.094 seconds.

Example 6.4. Finally, set $Y_a = 1$, $Y_c = 2$, $Z_b = 10$, and $\tilde{H}(x) = 1 + \cos x + \frac{1}{2}\cos 2x + \frac{1}{4}\cos 3x$. We obtain

$$V_1 = [\cdots, \ (.3634), \ .2494, \ .1374, \ .0685, \ .0179, \ .0047, \ .0012, \ \cdots]^T,$$

$$V_3 = [\cdots, \ (.0138), \ .0107, \ .0067, \ .0035, \ .0014, \ .0005, \ .0002, \ \cdots]^T,$$

$$V_5 = [\cdots, \ (.0005), \ .0005, \ .0003, \ .0002, \ .0001, \ .0000, \ .0000, \ \cdots]^T,$$

$$V_7 = [\cdots, \ (.0000), \ .0000, \ \cdots]^T.$$

Computer execution time was 2.728 seconds.

7. The three-dimensional grid. We have already mentioned that we can complicate the network of Fig. 1 by inserting branches in a consistent fashion between presently nonadjacent nodes within the same horizontal row and still use our method to compute the voltages and currents. So long as the graph remains uniform, we can apply a node analysis to compute the new conductance Y; the corresponding $\tilde{Y}(x)$ will again be a polynomial in $\cos x$, but now of higher degree. Also, a vertical periodicity of period greater than one branch can be allowed in the element values and graph of Fig. 1; this leads to a more complicated equation than (4.2) for the function $\tilde{Z}_0(x)$.

There is still another way of extending our method to more complicated grids, and that is to introduce three-dimensional ones. Fig. 6 illustrates the three-dimensional

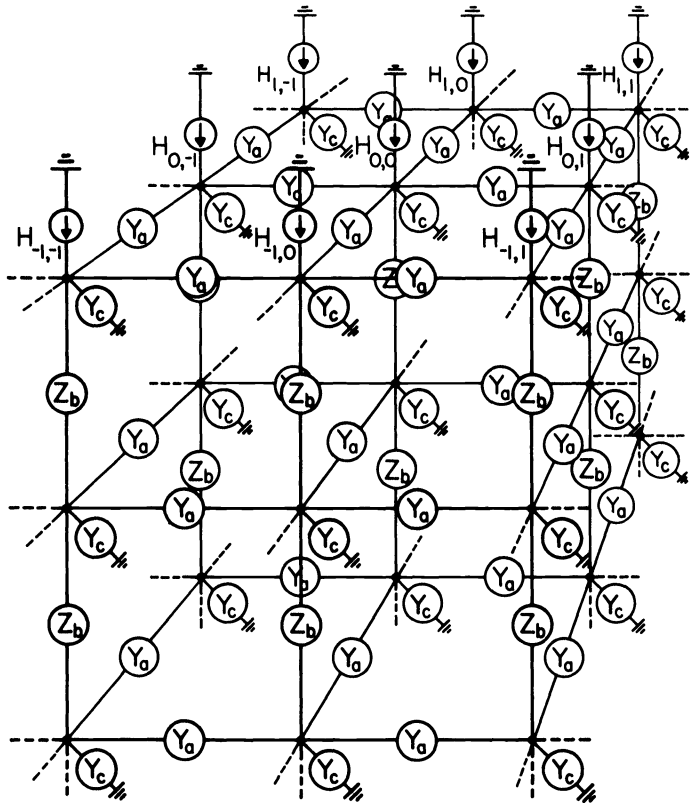


FIG. 6.

half-space analogue to Fig. 1. Let ξ, η and ζ represent the three coordinates of three-dimensional Euclidean space. Restrict ξ and η to the integers and ζ to the positive integers. Then, all such triplets (ξ, η, ζ) will be the locations of our grid's nodes, except for an additional ground node. For a fixed ζ and variable ξ and η, we have the ζth horizontal plane of nodes in Fig. 6. The positive ζ direction is the downward vertical direction. Every pair of nodes at a distance of one unit part is connected by a branch. Every horizontal branch has a positive conductance Y_a, and every vertical branch has a positive resistance Z_b. Moreover, every node is connected to ground through a positive conductance Y_c. Finally, every node of the form $(\xi, \eta, 1)$ has a current source $H_{\xi,\eta}$ connected to it from ground. $H_{\xi,\eta}$ is a real number, which we will also denote by H_k, where $k = (\xi, \eta) = (k_1, k_2)$ is a doublet of integers k_1 and k_2.

To analyze this configuration, it is convenient to alter our definition of the Hilbert space l_{2r} in a nonessential way. First of all, we let D denote the space of all ordered doublets whose entries are integers. Thus, $k \in D$. An element of l_{2r} is now taken to be a two-dimensional array $\{a_k : k \in D\}$ of real number a_k such that

$$\sum_{k \in D} a_k^2 = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} a_{(k_1,k_2)}^2 < \infty.$$

The inner product of two elements $a = \{a_k\}$ and $b = \{b_k\}$ in l_{2r} is now the double infinite series

$$(a, b) = \sum_{k \in D} a_k b_k.$$

Continuous linear mappings of l_{2r} into l_{2r} can be represented by a matrix-like notation, but that notation must now be interpreted appropriately since their indices are doublets. In particular, let F be such a mapping and let $a = \{a_k : k \in D\} \in l_{2r}$. Then there exists a four-dimensional array $[F_{jk}]$, where $j, k \in D$, such that

$$\sum_{k \in D} F_{jk} a_k \tag{7.1}$$

is the jth element of Fa. The proper interpretation of (7.1) is the following: For each fixed $j \in D$, $\{F_{jk} : k \in D\}$ is a two-dimensional array. Upon multiplying its entries with the corresponding ones in a and then summing the results, we get (7.1). Of course, not all arrays $[F_{jk}]$ can represent continuous linear mappings on l_{2r}. For example, conditions must be imposed to insure the convergence of (7.1) (see [7, p. 126]). However, all the matrices indicated below do represent continuous linear mappings of l_{2r} into l_{2r}.

Proceeding as before, we decompose Fig. 6 into a sequence of Hilbert ports. The first one consists of all the conductances Y_a and Y_c connected to the nodes in the first horizontal plane (and also to ground in the case of Y_c); the second Hilbert port consists of all the resistances Z_b connected between the first and second horizontal planes; the third Hilbert port consists of all the conductances Y_a and Y_c connected to the nodes of the third horizontal plane, and so forth. Finally, the array $H = \{H_k : k \in D\}$ of current sources connected to the nodes in the first horizontal plane is assumed to be in l_{2r}. This decomposition results in the ladder network of Hilbert ports shown in Fig. 5. Because of the grounded and disconnected structures of out Hilbert ports, the port conditions are not violated by the interconnections of Fig. 5.

A nodal analysis shows that the admittance operator Y for the Hilbert port corresponding to any horizontal plane is given by $Y = [Y_{jk}]$, where

$$Y_{jk} = \begin{cases} Y_c + 4Y_a & \text{for } j = k, \\ -Y_a & \text{for } |j_1 - k_1| + |j_2 - k_2| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

This is a four-dimensional analogue of a row-finite Laurent matrix.

Similarly, the impedance operator $Z = [Z_{jk}]$ of the Hilbert port of resistances Z_b is given by

$$Z_{jk} = \begin{cases} Z_b & \text{for } j = k, \\ 0 & \text{for } j \neq k. \end{cases}$$

We now have the analogue of a diagonal matrix with equal main diagonal terms.

Next, we exploit the isomorphism between our present space l_{2r} and the space of double Fourier series; that is, $a = \{a_k : k \in D\}$ corresponds to

$$\tilde{a}(x) = \sum_{k \in D} a_k e^{i(k,x)},$$

where $k = (k_1, k_2)$, $x = (x_1, x_2)$, and $(k, x) = k_1 x_1 + k_2 x_2$. Under this correspondence, the operator Y transforms into multiplication by the function

$$\tilde{Y}(x) = Y_a[\gamma - 2\cos x_1 - 2\cos x_2],$$

where

$$\gamma = \frac{Y_c}{Y_a} + 4 > 4.$$

The operator Z corresponds to multiplication by the constant function $\tilde{Z}(x) = Z_b$. Finally, the given current-source array H corresponds to

$$\tilde{H}(x) = \sum_{k \in D} H_k e^{i(k,x)}.$$

Now, the characteristic-resistance method of analyzing the operator network of Fig. 5 carries directly over to the present case. In particular, the characteristic resistance Z_0 is the Laurent matrix corresponding to multiplication by the function $\tilde{Z}_0(x)$, which is the solution of

$$[\tilde{Z}_0(x)^{-1} - \tilde{Y}(x)][\tilde{Z}_0(x) + \tilde{Z}(x)] = 1.$$

This yields

$$\tilde{Z}_0(x) = \frac{-Z_b\tilde{Y}(x) + \{[Z_b\tilde{Y}(x)]^2 + 4Z_b\tilde{Y}(x)\}^{1/2}}{2\tilde{Y}(x)}.$$

Similarly, the transmission operator θ corresponds to multiplication by the function

$$\tilde{\theta}(x) = 1 + \tfrac{1}{2}Z_b\tilde{Y}(x) - \tfrac{1}{2}\{[Z_b\tilde{Y}(x)]^2 + 4Z_b\tilde{Y}(x)\}^{1/2}.$$

Then

$$\tilde{V}_1(x) = \tilde{Z}_0(x)\tilde{H}(x)$$

and, for $n = 1, 2, 3, \cdots$,

$$\tilde{V}_{2n+1}(x) = [\tilde{\theta}(x)]^n \tilde{V}_1(x).$$

By expanding $\tilde{V}_1(x)$ into a double Fourier series and picking out its coefficients:

$$v_k = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \tilde{V}_1(x)\, e^{-i(k,x)}\, dx,$$

we obtain all the node voltages in the first horizontal plane of Fig. 6. From these, a simple limb analysis yields any desired voltage or current in Fig. 6 (at least in principle; limited computer time limits the number of branches one can consider). Alternatively, the coefficients of the double Fourier series expansion of $\tilde{V}_{2n+1}(x)$ yield the node

TABLE 1

$\zeta = 1$: η \ ξ	0	1	2	3	4
0	.20329	.04435	.01088	.00293	.00085
1		.01825	.00625	.00204	.00066
2			.00274	.00107	.00039
3				.00048	.00020
4					.00009

$\zeta = 2$: η \ ξ	0	1	2	3	4
0	.03938	.01593	.00542	.00178	.00058
1		.00864	.00357	.00133	.00047
2			.00175	.00074	.00029
3				.00036	.00016
4					.00007

$\zeta = 3$: η \ ξ	0	1	2	3	4
0	.00880	.00501	.00219	.00086	.00032
1		.00329	.00161	.00068	.00027
2			.00089	.00042	.00018
3				.00021	.00010
4					.00005

$\zeta = 4$: η \ ξ	0	1	2	3	4
0	.00223	.00155	.00081	.00037	.00016
1		.00115	.00065	.00031	.00014
2			.00039	.00020	.00009
3				.00011	.00006
4					.00003

$\zeta = 5$: η \ ξ	0	1	2	3	4
0	.00062	.00049	.00029	.00015	.00007
1		.00039	.00024	.00013	.00006
2			.00016	.00009	.00004
3				.00005	.00003
4					.00002

voltages in the nth horizontal plane of Fig. 6 for any n. From these, any voltage or current can be obtained once again (subject to the usual caveat about computer time).

What we have just described is a formal method of computing the voltages and currents in Fig. 6. Of equal importance is the fact that the existence and uniqueness discussions of the prior sections carry directly over to our present three-dimensional grid. Indeed, all the theorems, and in particular Theorems 2.2 and 5.1, hold once again for the grid of Fig. 6. Only a few minor modifications, primarily those of notation, need be made.

Finally, we point out once again that we can augment Fig. 6 by adding other horizontal branches to pairs of nodes further apart than one unit, so long as the uniformity of each horizontal plane is maintained. This merely alters the function $\tilde{Y}(x)$ by adding additional terms of the form $b_{\nu\mu} \cos \nu x_1 \cos \mu x_2$ where ν and μ are integers. Furthermore, variations in the graphs and element values in the vertical direction can also be allowed so long as periodicity in the vertical direction is maintained; this leads to the problem of solving more complicated equations for $\tilde{Z}_0(x)$.

Example 7.1. As an example, we have computed the node voltages for the grid of Fig. 6 for the case when $Y_a = Z_b = Y_c = 1$, $H_{0,0} = 1$, and $H_{\xi,\eta} = 0$ if $\xi \neq 0$ or $\eta \neq 0$. For the node voltages, we use the indexing system explained at the beginning of this section. The values of the node voltages for the first five horizontal planes of nodes (i.e., for $\zeta = 1, \cdots, 5$) are given in Table 1. Because of the symmetry of our grid and of the applied current sources, we need merely display the node voltages for $\xi = 1, 2, 3, \cdots$ and $\eta = 1, \cdots, \xi$. The other node values are obtained by interchanging ξ with η and by using even symmetry for negative indices. We terminate the tables at $\xi = 4$. Execution time on the computer was 1 minute, 56.0 seconds.

8. Grids of impedances. So far, we have restricted our attention to purely resistive grids. We now wish to allow capacitors, inductors, transformers, etc. as grid parameters. More specifically, we shall now generalize the grid of Fig. 1 by assuming that Y_a, Z_b, and Y_c are (scalar) positive real functions. Our objective is to show that, if the current sources in Fig. 1 are given as suitably restricted Laplace-transformable distributions on the time axis, then there exists a unique set of Laplace-transformable voltage and current distributions that satisfy Kirchhoff's and Ohm's laws and a certain form of the finite-power-dissipation condition. They will be determinable from our operator version of the characteristic-impedance method.

To accomplish this, we shall make use of the results and notations of [20]. C_+ will denote the open right half of the complex plane C:

$$C_+ = \{s \in C: \operatorname{Re} s > 0\}.$$

For $s \in C_+$, Ω_s is the closed cone

$$\Omega_s = \{z \in C: |\arg z| \leqq |\arg s|\},$$

where it is understood that the origin is a member of Ω_s. l_2 denotes the complexification of l_{2r}. By an "operator", we henceforth mean a continuous linear mapping of l_2 into l_2. For any operator F, $W[F]$ is the numerical range of F:

$$W[F] = \{(Wa, a): a \in l_2, \|a\| = 1\}.$$

P is the set of all analytic operator-valued functions F on C_+ such that, for every $s \in C_+$, $W[F(s)] \subset \Omega_s$. Thus, if $F \in P$, $F(\sigma)$ is a positive operator for each $\sigma > 0$. P_i is the set of all $F \in P$ such that, for every fixed $s \in C_+$, $W[F(s)]$ is bounded away from the origin; that is, there exists a $\delta > 0$ depending in general on s such that

$\mathrm{Re}(F(s)a, a) \geqq \delta\|a\|^2$ for all $a \in l_2$. Thus, for each $s \in C_+$, and $F \in P_i$, $F(s)$ is an invertible operator. It was shown in [20] that, if $F \in P_i$ and $G \in P$, then $F+G \in P_i$; also, if $F \in P_i$ and if F^{-1} denotes the function $s \to [F(s)]^{-1}$, then $F^{-1} \in P_i$.

Next, let $F_1, F_2, F_3, \cdots \in P_i$ and let $Z_n(s)$ be the operator-valued finite continued fraction

$$Z_n(s) = \frac{1}{F_1(s)} + \frac{1}{F_2(s)} + \cdots + \frac{1}{F_n(s)}. \tag{8.1}$$

Then, by what has just been stated, $Z_n \in P_i$. The following theorem is a somewhat simplified version of [20, Theorem 1, Corollary 1a].

THEOREM 8.1. *Assume the following three conditions*:

(i) $F_k \in P_i$ *for every* $k = 1, 2, 3, \cdots$.

(ii) *Given any compact set* $\Xi \subset C_+$, *there exists a constant* $\delta > 0$, *depending upon* Ξ, *such that* $\inf \mathrm{Re}\ W[F_k(s)] > \delta$, *for* $k = 1, 2$ *and* $s \in \Xi$.

(iii) *For each* $\sigma > 0$ *and all* $k = 1, 2, 3, \cdots$, *the operators* $F_k(\sigma)$ *commute with each other and* $W[F_k(\sigma)] > \delta_k(\sigma)$, *where the* $\delta_k(\sigma)$ *are positive numbers satisfying* $\sum_{k=1}^{\infty} \delta_k(\sigma) = \infty$.

Then, for every $s \in C_+$, *the sequence* $\{Z_n(s)\}_{n=1}^{\infty}$ *converges in the uniform operator topology, and the convergence is uniform with respect to s in any compact subset of* C_+. *Moreover, the limit function* $Z = \lim_{n\to\infty} Z_n$ *is a member of* P_i.

We shall now apply this theorem to the ladder network of Fig. 5. Assume that Y_a and Y_c are scalar positive real functions, neither of which is identically equal to zero. Let Y be the Laurent matrix defined by (3.2). By the argument in [18, p. 186], Y is an operator on l_2. Also, a short computation shows that, for each $a = \{a_k\}_{k=-\infty}^{\infty} \in l_2$ and $s \in C_+$,

$$(Y(s)a, a) = Y_c(s)\|a\|^2 + Y_a(s) \sum_{k=-\infty}^{\infty} |a_k - a_{k-1}|^2. \tag{8.2}$$

The polarization identity in conjunction with (8.2) shows that Y is weakly analytic on C_+ and therefore analytic on C_+ [17, p. 18, p. 197].

Now, a standard property of a scalar positive real function F is that $|\arg F(s)| \leqq |\arg s|$ for $s \in C_+$ [14, Theorem 5]. Since Y_a and Y_c are scalar positive real functions, this fact coupled with (8.2) implies that $Y \in P$. Since Y_c is not identically equal to zero, the minimax theorem for harmonic functions shows that, given any compact subset $\Xi \subset C_+$, there is a $\delta > 0$ for which

$$\mathrm{Re}\,(Y(s)a, a) \geqq \mathrm{Re}\ Y_c(s)\|a\|^2 \geqq \delta\|a\|^2,$$

for all $s \in \Xi$. Thus, $Y \in P_i$, and in addition Y satisfies Hypothesis (ii) of Theorem 8.1 when we set $F_k(s) = Y(s)$.

A simpler argument shows that, if Z_b is a scalar positive real function, then the Laurent matrix Z, as defined by (3.1), also is in P_i and satisfies Hypothesis (ii) of Theorem 8.1 when $F_k(s) = Z(s)$. Moreover, row-finite Laurent matrices commute, and therefore so do $Y(s)$ and $Z(s)$. Finally, upon setting $F_k(s) = Y(s)$ for k odd and $F_k(s) = Z(s)$ for k even and noting that the series and shunt elements of Fig. 5 remain invariant, we see that Hypothesis (iii) of Theorem 8.1 is also satisfied.

Thus, we may invoke Theorem 8.1 to conclude that, as $n \to \infty$, the Z_n defined by (8.1) converge to a member of P_i. But, the Z_n are the driving-point impedance operators of truncations of the ladder network of Fig. 5. Hence, their limit is the characteristic impedance Z_0 of that infinite ladder network. Thus, we have established

THEOREM 8.2. *Let Z be defined by* (3.1) *and Y by* (3.2), *where Z_b, Y_a, and Y_c are scalar positive real functions. Then, Z and Y are members of P_i, and so too is the characteristic-impedance operator Z_0 of the infinite ladder network of Fig.* 5.

P_i is a subset of the class of all positive* operator-valued functions [17, p. 178]. Moreover, an operator-valued positive* function is the Laplace transform of a right-sided operator-valued distribution, where C_+ is understood to be the region of definition for the transform. (A distribution on the real line is called right-sided if its support is bounded on the left). Also, the infimum of the support of that distribution is the origin [17, Theorem 8.12–1]. Thus, Y, Z_0, $\theta = 1 - YZ_0$, and, for every $k = 1, 2, 3, \cdots$, θ^k are all Laplace transforms of operator-valued distributions whose supports are bounded on the left at the origin, and C_+ is contained in all of the regions of definition for those transforms [17, Theorems 6.5–1, 8.11–2, and 8.12–1]. Moreover, these operators are real in the sense that for each $s = \sigma > 0$, the operator maps l_{2r} into l_{2r}.

Next, we assume that $h(t)$ is a doubly infinite vector of current sources in the time domain:

$$h(t) = [\cdots, \quad h_{-1}(t), \quad h_0(t), \quad h_1(t), \quad \cdots]^T. \tag{8.3}$$

Here, the superscript T denotes matrix transpose. Our next objective is to state conditions under which the Laplace transform H of h is l_2-valued for each point in C_+. To this end, we let $L_2(R, l_{2r})$ denote the Hilbert space of quadratically integrable (equivalence classes of) functions on the real line R taking their values in l_{2r}. If $a, b \in L_2(R, l_{2r})$, their inner product is

$$(a, b) = \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k(t) b_k(t) dt,$$

where a_k and b_k are the components of a and b and $t \in R$.

LEMMA 8.1. *Assume that the vector h is a member of $L_2(R, l_{2r})$ and that the support of h is bounded on the left. Let H be the Laplace transform of h, that is, the vector of Laplace transforms H_k of the components h_k. Then, for each $s \in C_+$, $H(s)$ exists and is a member of l_2. Also, for each $\sigma > 0$, $H(\sigma)$ is a member of l_{2r}.*

Proof. Every right-sided scalar function in L_2 has a Laplace transform on C_+. So, the components of $H(s)$ all exist for $s \in C_+$. We wish to show that $H(s) \in l_2$. Let τ be the infimum of the support of h, and let $\sigma = \operatorname{Re} s > 0$. Then,

$$|H_k(s)| = \left| \int_{\tau}^{\infty} h_k(t)\, e^{-st}\, dt \right| \leqq \int_{\tau}^{\infty} |h_k(t)|\, e^{-\sigma t}\, dt.$$

By Schwarz's inequality

$$|H_k(s)|^2 \leqq \frac{e^{-2\sigma\tau}}{2\sigma} \int_{\tau}^{\infty} |h_k(t)|^2\, dt.$$

Therefore,

$$\sum_{k=-\infty}^{\infty} |H_k(s)|^2 \leqq \frac{e^{-2\sigma\tau}}{2\sigma} \sum_{k=-\infty}^{\infty} \int_{\tau}^{\infty} |h_k(t)|^2\, dt. \tag{8.4}$$

Now, $\{\sum_{k=-r}^{r} |h_k(t)|^2 : r = 1, 2, 3, \cdots\}$ is a monotonic sequence of integrable functions which converge to $\sum_{k=-\infty}^{\infty} |h_k(t)|^2$ for almost all t. So, by Levi's theorem we can

interchange the summation with the integration to write

$$\sum_{k=-\infty}^{\infty} |H_k(s)|^2 \leqq \frac{e^{-2\sigma\tau}}{2\sigma} \int_{\tau}^{\infty} \sum_{k=-\infty}^{\infty} |h_k(t)|^2 \, dt.$$

The integral on the right-hand side is the square of the norm of $h \in L_2(R, l_{2r})$. This proves the first conclusion. The second conclusion now follows immediately from the reality of h.

Under the hypotheses of Theorem 8.2 and Lemma 8.1, $V_1(s)$, $V_{2n+1}(s)$, and $I_{2n}(s)$, as given by (4.7), (4.9), and (4.10), are all l_2-valued Laplace transforms with C_+ contained in their regions of definition [17, Theorem 6.3–2]. Moreover, their respective inverse Laplace transforms $v_1(t)$, $v_{2n+1}(t)$, and $i_{2n}(t)$ will be l_{2r}-valued distributions with supports bounded on the left at the origin. The reality of these distributions is a consequence of [17, Theorem 8.13–1] and the exchange formula [17, Theorem 6.3–2]. The various components of these distributions yield the transient voltages and currents of our infinite grid under the finite-power condition on the real positive axis of the s-domain.

To put this another way, we assume that each branch of Fig. 1 is a one-port of resistive and reactive elements such that Y_a, Z_b, and Y_c are scalar positive real functions. We apply current sources $h_k(t)$ and ask for the time-domain behavior of the grid's voltages and currents. To this end, we generalize Ohm's law by relating the time-domain voltage v and current i across any one-port through one of the distributional convolutions

$$i = y_a * v, \qquad v = z_b * i, \qquad i = y_c * v, \tag{8.4}$$

where y_a, z_b, and y_c are the inverse Laplace transforms of the scalar impedances Y_a, Z_b, and Y_c, with C_+ being at least part of the regions of definition for those transforms. The next theorem, which is our main time-domain existence and uniqueness assertion, follows immediately from Theorem 2.2, Theorem 5.1, and the uniqueness property of the Laplace transformation [17, Theorem 6.4–2].

THEOREM 8.3. *Let Y_a, Z_b, and Y_c be scalar positive real functions that are not identically equal to zero. Let the current-source vector* (8.3) *satisfy the hypothesis of Lemma* 8.1. *Then, there exists one and only one set of right-sided, Laplace-transformable distributions for the branch voltages v_j in the grid of Fig.* 1 *such that Kirchhoff's node and loop laws and the generalized Ohm's laws* (8.4) *are satisfied in the time domain and such that, for at least one $\sigma > 0$,*

$$\sum [V_j(\sigma)]^2 < \infty, \tag{8.5}$$

where V_j is the Laplace transform of v_j. In this case, (8.5) *holds for all $\sigma > 0$. For any given $\sigma > 0$, the $V_j(\sigma)$ can be determined by applying the characteristic-resistance method of* §§ 4 *and* 6.

This result extends directly to the three-dimensional grid of Fig. 6. Moreover, it can also be extended to more complicated grids such as those described in the penultimate paragraph of the preceding section.

9. The computation of transient responses. The last conclusion of Theorem 8.3 coupled with a method of Papoulis [12], [13] and its modification by Lerner and Lerner [10] provide a means of computing the transient behavior of our electrical grids. Those methods will work if the transitions are sufficiently well-behaved functions rather than distributions. This will be the case if the current sources and grid parameters are suitably restricted.

For example, consider the grid of Fig. 1. Let us assume that in the time domain a current source $f(t)$ is connected to the mth node in the first row and that all other sources are zero. Let $F(s)$ with $s \in C_+$ be the Laplace transform of $f(t)$ and assume that $F(s)$ is of order $O(|s|^{-j})$ as $s \to \infty$ in C_+. Then $\tilde{H}(x, s) = F(s)e^{imx}$, $s \in C_+$.

Let us also assume that Y_a, Z_b, and Y_c are rational positive real functions such that Y_a and Z_b behave resistively and Y_c behaves capacitively as $s \to \infty$. Then

$$\alpha(s) = 2 + \frac{Y_c(s)}{Y_a(s)} \sim Ks,$$

and

$$\beta(s) = Y_a(s)Z_b(s) \sim M \quad \text{as } s \to \infty,$$

where K and M are constants. Upon substituting $\alpha(s)$ and $\beta(s)$ into (4.3) and (6.1), we obtain $\tilde{Z}_0$ and $\tilde{\theta}$ as functions of both x and s. Thus the Fourier coefficients of (6.2) and (6.3) can be obtained as functions of s; these Fourier coefficients are the Laplace transforms of the time-dependent node voltages in our grid.

Let us examine the transient voltage $v(t)$ at an arbitrary node of our grid, say, the pth node in the $(n+1)$st row. Its Laplace transform $V(s)$ is the pth Fourier coefficient of $[\tilde{\theta}(x, s)]^n \tilde{Z}_0(x, s)\tilde{H}(x, s)$:

$$V(s) = \frac{F(s)}{2\pi} \int_0^{2\pi} [\tilde{\theta}(x, s)]^n \tilde{Z}_0(x, s)\, e^{i(m-p)x}\, dx.$$

It follows from the right-hand side of (4.4) and our assumptions on Y_a, Z_b, and Y_c that, as $s \to \infty$ in C_+, $\tilde{Z}_0(x, s) = [\tilde{Y}_0(x, s)]^{-1}$ is asymptotic to a constant times s^{-1} and is uniformly so for all x. As $s \to \infty$, $\tilde{\theta}(x, s)$ is asymptotic to a constant times s^{-1} uniformly for all x, as can be seen from the Fourier series analogue of the last expression in (4.8). Therefore, $V(s)$ is of order $O(|s|^{-n-j-1})$ as $s \to \infty$ in C_+. Consequently, $v(t)$ is a function whose first $n+j-1$ derivatives are continuous for all t; also, $v(t)$ is equal to zero for $t < 0$ [16, Lemma 3.6–1]. We can now use Papoulis' numerical method to compute $v(t)$. A particular example of such a computation is given in [24].

10. Nonlinear resistive grids. We can use some results of Dolezal [3] to analyze approximately the infinite grid of Fig. 1 (and many other grids as well) in the case where every branch parameter, other than the sources, is a single nonlinear monotonically increasing resistor $r: i \mapsto r(i)$. This is accomplished by replacing r by a linear resistor r_0: $i \mapsto r_0 i$, analyzing the resulting linear grid by our characteristic-resistance method, and then determining a bound on the error between the current responses of the nonlinear grid and its linear approximation.

More specifically, Dolezal's results are as follows. Let Y_a^{-1}, Z_b, and Y_c^{-1} all be the same nonlinear resistor r. Assume that $r(0) = 0$ and that the slope of r is bounded as follows:

$$\alpha(p-q)^2 \leqq [r(p) - r(q)](p-q) \leqq \beta(p-q)^2$$

for all p and q on the real line. Here, α and β are constants satisfying

$$0 < \alpha \leqq \beta < 3\alpha.$$

Furthermore, let us assume that each current source H_k and its parallel resistance Y_c^{-1} has been replaced by a voltage source e_k in series with r. Let

$$e = [\cdots, \quad e_{-1}, \quad e_0, \quad e_1, \quad \cdots]^T \in l_2.$$

A theorem of Dolezal [3, Theorem 3] asserts that, under these conditions, there exists a unique vector $i \in l_2$ of branch currents in this nonlinear network.

To obtain some estimation of what i is, Dolezal replaces every r by a linear resistance r_0 chosen as follows:

Let A be either a real positive number or ∞. Set

$$r_0 = \tfrac{1}{2}(S_A + J_A),$$

where

$$S_A = \sup_{p \in K_A} p^{-1} r(p), \qquad J_A = \inf_{p \in K_A} p^{-1} r(p),$$

$$K_A = [-A, A] - \{0\}.$$

Now, we let j denote the vector of branch currents in this linear approximation to the given nonlinear grid with the aforementioned voltage sources e_k. $j \in l_2$ can be computed by our characteristic-resistance method. (A change of voltage sources into current sources presents no difficulties). Finally, Dolezal has proven that, with $\|\cdot\|$ denoting the l_2-norm, we have

$$\|i - j\| = \frac{\beta - \alpha}{\alpha(3\alpha - \beta)} \|e\|,$$

for all $e \in l_2$ with $\|e\| \leqq \alpha A$.

Acknowledgment. The numerical computations for this work were performed by Pei-Hua Lo and Prasad Subramaniam.

REFERENCES

[1] ARLEN BROWN AND P. R. HALMOS, *Algebraic properties of Toeplitz operators*, J. für Mathematik, 213 (1963), pp. 89–102.
[2] V. DOLEZAL, *Nonlinear Networks*, Elsevier Scientific Publishing Co., New York, 1977.
[3] ———, *Linearization of monotone Hilbert networks*, this Journal, 10 (1979), pp. 523–531.
[4] H. FLANDERS, *Infinite networks: I—resistive networks*, IEEE Trans. Circuit Theory, CT-18 (1971), pp. 326–331.
[5] ———, *Infinite networks: II—resistance in an infinite grid*, J. Math. Anal. Appl., 40 (1972), pp. 30–35.
[6] P. R. HALMOS, *A Hilbert Space Problem Book*, D. Van Nostrand Co., Princeton, New Jersey, 1967.
[7] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.
[8] G. KRON, *Numerical solution of ordinary and partial differential equations by means of equivalent circuits*, J. Appl. Phys., 16 (1945), pp. 172–186.
[9] ———, *Electrical circuit models of partial differential equations*, Elec. Eng., 67 (1948), pp. 672–684.
[10] D. M. LERNER AND G. M. LERNER, *A simplified algorithm for the inversion of the Laplace transform*, Radiofizika, 13 (1970), pp. 619–621.
[11] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar Publ. Co., New York, 1961.
[12] A. PAPOULIS, *A new method of inversion of the Laplace transform*, Quart. Appl. Math., 14 (1956), pp. 405–414.
[13] ———, *A different approach to the analysis of tracer data*, SIAM J. Control, 11 (1973), pp. 466–474.
[14] S. SESHU AND N. BALABANIAN, *Transformations of positive real functions*, IEEE Trans. Circuit Theory, CT-4 (1957), pp. 306–312.
[15] S. M. SZE, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, 1969.
[16] A. H. ZEMANIAN, *Generalized Integral Transformations*, Wiley-Interscience, New York, 1968.
[17] ———, *Realizability Theory for Continuous Linear Systems*, Academic Press, New York, 1972.
[18] ———, *Passive operator networks*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 184–193.

[19] ———, *Countably infinite networks that need not be locally finite*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 274–277.
[20] ———, *Continued fractions of operator-valued analytic functions*, J. Approx. Theory, 11 (1974), pp. 319–326.
[21] ———, *Infinite electrical networks*, Proc. IEEE 64 (1976), pp. 6–17.
[22] ———, *The limb analysis of countably infinite electrical networks*, J. Combin. Theory Ser. B, 24 (1978), pp. 76–93.
[23] ———, *Countably infinite nonlinear time-varying active electrical networks*, this Journal, 10 (1979), to appear.
[24] ———, *The Characteristic-Resistance Method for Grounded Semi-infinite Grids*, State University of New York at Stony Brook, College of Engineering Technical Report No. 330, August 30, 1979.

SIAM J. MATH. ANAL.
Vol. 12, No. 1, January 1981

CORRIGENDUM: A NOTE ON THE RAYLEIGH POLYNOMIALS*

E. C. OBI†

There are three minor subscript and typographical errors in the article, to be simply amended as follows:

Page 825. In equation(*), the subscript n, in σ_n, should be k. In the sentence below equation(**), the expression

$$\frac{1}{4}(\nu+1) \text{ should read } \frac{1}{4(\nu+1)}.$$

Page 828. In lemma 2(b), the subscript $n-1$, in σ_{n-1}, should be $m-1$.

None of these corrections, however, affects anything else in the text.

* This Journal, 9 (1978), pp. 825-834.

† Department of Mathematics, University of Nigeria, Nsukka Campus, East Central State, Nigeria.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0001 $01.00/0

A UNIQUENESS THEOREM FOR ORDINARY DIFFERENTIAL EQUATIONS*

M. J. NORRIS† AND R. D. DRIVER‡

Abstract. The uniqueness theorem of this paper answers an open question for a system of differential equations arising in a certain n-body problem of classical electrodynamics. The essence of the result can be illustrated using the scalar prototype equation $x' = g_1(x) + g_2(t+x)$ with $x(0) = 0$. The solution of the latter will be unique provided g_1 and g_2 are continuous positive functions of bounded variation.

The theorem proved in this paper presents a criterion weaker than a Lipschitz condition which assures uniqueness of solutions of a system of ordinary differential equations. It was designed to resolve an open question in classical electrodyamics described at the end of the paper.

Before stating the theorem, let us illustrate it with two scalar examples typifying the problems we had in mind. These examples are easily treated with the theorem which follows. We are unaware of any previous uniqueness theorem which would handle them or the electrodynamics problem of Example 3.

Example 1. If g_1 and g_2 are continuous positive functions of bounded variation on an open interval containing 0, then the equation

$$x' = g_1(x) + g_2(t+x) \quad \text{with } x(0) = 0$$

has a unique solution on some open interval containing 0.

Example 2. The equation

$$x' = (t + x^{5/3})^{1/3} \quad \text{for } t \geqq 0 \text{ with } x(0) = 0$$

has a unique solution.

The theorem itself treats a system of n ordinary differential equations

$$x' = f(t, x) \tag{1}$$

with initial conditions

$$x(t_0) = x_0. \tag{2}$$

Let S be a subset (not necessarily open) of R^{n+1}, and let $f: S \to R^n$. Then, given $(t_0, x_0) \in S$, a solution of (1) and (2) is defined as any differentiable function x on an interval J such that $(t, x(t)) \in S$ and $x' = f(t, x(t))$ for $t \in J$, while $t_0 \in J$ and $x(t_0) = x_0$. (If J contains either of its endpoints, $x'(t)$ is a one-sided derivative there).

The norm used in this paper for a vector $\xi \in R^n$ is $\|\xi\| = \sum_{i=1}^{n} |\xi_i|$.

THEOREM. *Let $f: S \to R^n$ be continuous and satisfy the following condition. Each point in S has an open neighborhood U, a constant $K > 0$, an integer $m \geqq 0$, and functions h_j and g_j for $j = 1, \cdots, m$ such that*

$$\|f(t, \xi) - f(t, \eta)\| \leqq K\|\xi - \eta\| + K \sum_{j=1}^{m} |g_j(h_j(t, \xi)) - g_j(h_j(t, \eta))| \tag{3}$$

---

* Received by the editors May 30, 1980.

† Applied Mathematics Department 5640, Sandia National Laboratories, Albuquerque, New Mexico 87115. The work of this author was supported in part by the U.S. Department of Energy under contract DE-AC04-76DP00789.

‡ Department of Mathematics, University of Rhode Island, Kingston, Rhode Island 02881. The work of this author was supported in part by the U.S. Air Force Office of Scientific Research under contract F49620-79-C-0129.

on $U \cap S$, where $h_j : U \to R$ is continuously differentiable with

$$\frac{\partial h_j(t, \xi)}{\partial t} + \sum_{i=1}^{n} \frac{\partial h_j(t, \xi)}{\partial \xi_i} f_i(t, \xi) \neq 0 \quad \text{on } U \cap S, \tag{4}$$

and each $g_j : R \to R$ is continuous and of bounded variation on bounded subintervals. Then (1) *and* (2) *with any point $(t_0, x_0) \in S$ have at most one solution on any interval J.*

Remarks. The theorem of course does not guarantee the existence of a solution on a nontrivial interval J. Existence would follow, for example, if S were open.

To treat Example 1, define $h_1(t, \xi) = \xi$ and $h_2(t, \xi) = t + \xi$. For Example 2, let $h(t, \xi) = t + \xi^{5/3}$ and $g(\xi) = \xi^{1/3}$.

Proof of the theorem. Suppose there were two different solutions, x and y, on some interval $J = [t_0, b)$ where $b > t_0$. (The case $J = (b, t_0]$ is handled similarly.) Let

$$a \equiv \inf \{t \in (t_0, b) : x(t) \neq y(t)\}.$$

Then $x(a) = y(a)$.

For the point $(a, x(a)) \in S$ let U, K, m, h_j and g_j be as described in the hypotheses of the theorem. Without loss of generality, assume that for each j the expression in (4) is positive at $(a, x(a))$. Then, reducing U if necessary, the continuity of the derivatives of h_j assures that there exist positive constants p and M such that, for $j = 1, \cdots, m$,

$$\frac{\partial h_j(t, \xi)}{\partial t} + \sum_{i=1}^{n} \frac{\partial h_j(t, \xi)}{\partial \xi_i} f_i(t, \xi) \geqq p \quad \text{on } U \cap S \tag{5}$$

and

$$|h_j(t, \xi) - h_j(t, \eta)| \leqq M \|\xi - \eta\| \quad \text{on } U. \tag{6}$$

Choose a bounded interval $[\alpha_j, \beta_j]$ which contains $h_j(U \cap S)$, reducing U if necessary. Then g_j is the difference of two continuous nondecreasing functions on $[\alpha_j, \beta_j]$, and each of the latter can be extended to a continuous nondecreasing function on R by defining it to be constant on $(-\infty, \alpha_j]$ and constant on $[\beta_j, \infty)$. Without loss of generality, we shall assume that each g_j is itself nondecreasing on R and that

$$g_j(h_j(a, x(a))) = 0. \tag{7}$$

Define

$$z(t) = \int_a^t \|x'(s) - y'(s)\| \, ds \quad \text{for } a \leqq t < b.$$

Then $z(a) = 0$, $z'(a) = 0$, z and z' are continuous, $z'(t) \geqq 0$ and $\|x(t) - y(t)\| \leqq z(t)$ on $[a, b)$.

Choose $c \in (a, b]$ sufficiently small so that $(s, x(s))$ and $(s, y(s))$ remain in U for $a \leqq s < c$. Then, from (6),

$$h_j(s, x(s)) - Mz(s) \leqq h_j(s, y(s)) \leqq h_j(s, x(s)) + Mz(s),$$

and, from (5),

$$\frac{d}{ds} h_j(s, x(s)) \geqq p$$

for $a \leqq s < c$ and $j = 1, \cdots, m$.

Thus for $a \leqq t < c$, using (3) and the monotonicity of each g_j gives

$$z(t) \leqq K\int_a^t \left\{ \|x(s)-y(s)\| + \sum_{j=1}^{m} |g_j(h_j(s, x(s))) - g_j(h_j(s, y(s)))| \right\} ds$$

$$\leqq K(t-a)z(t) + \frac{K}{p}\sum_{j=1}^{m}\int_a^t [g_j(h_j(s, x(s)) + Mz(s))$$

$$-g_j(h_j(s, x(s)) - Mz(s))]\frac{d}{ds}h_j(s, x(s))\, ds$$

$$= K(t-a)z(t) + \frac{K}{p}\sum_{j=1}^{m}\int_{h_j(t,x(t))-Mz(t)}^{h_j(t,x(t))+Mz(t)} g_j(u)\, du$$

$$-\frac{K}{p}\sum_{j=1}^{m}\int_a^t [g_j(h_j(s, x(s)) + Mz(s)) + g_j(h_j(s, x(s)) - Mz(s))]Mz'(s)\, ds.$$

Choose $\delta_1 > 0$ such that for each j

$$|g_j(u)| < \frac{p}{6mKM} \quad \text{when } |u - h_j(a, x(a))| < \delta_1.$$

Then choose $\delta \in (0, 1/6K)$ such that $a + \delta \leqq c$ and, for each j,

$$|h_j(t, x(t)) - h_j(a, x(a))| + Mz(t) < \delta_1 \quad \text{when } a \leqq t < a + \delta.$$

Now for $a < t < a + \delta$ one finds $z(t) \leqq 5z(t)/6$. This contradiction completes the proof.

The motivation for this paper was the following problem from classical electrodynamics.

Example 3. Consider n electrically charged point particles moving along the x-axis at distinct positions, $x_1(t), x_2(t), \cdots, x_n(t)$. Assume that the motion of particle j depends only on the electromagnetic fields produced by the other $n-1$ particles, with these fields traveling to particle j at the speed of light, c.

The required fields are calculated in terms of the trajectories of the other particles from the retarded Liénard–Wiechert potentials; they are substituted into the Lorentz force law for particle j. Introducing $v_i = x_i'/c$ for the velocity of particle i as a multiple of c, one obtains a system of delay differential equations with state-dependent delays:

$$\frac{v_j'}{(1-v_j^2)^{3/2}} = \sum_{i \neq j} \frac{K_{ij}}{r_{ij}^2} \frac{\sigma_{ij} + v_i(t - r_{ij})}{\sigma_{ij} - v_i(t - r_{ij})}, \tag{8}$$

where each K_{ij} is a constant, $\sigma_{ij} \equiv \operatorname{sgn}[x_j(0) - x_i(0)]$, and where $r_{ij} > 0$ satisfies

$$r_{ij}' = \frac{v_j - v_i(t - r_{ij})}{\sigma_{ij} - v_i(t - r_{ij})}, \quad \text{for } i \neq j. \tag{9}$$

In these equations, v_j and r_{ij} without an argument stand for $v_j(t)$ and $r_{ij}(t)$.

In order to solve the system of n^2 equations represented by (8) and (9) when $t \geqq 0$, one should know not only

$$v_j(0) \text{ and } r_{ij}(0) \quad \text{for all } j \text{ and all } i \neq j, \tag{10}$$

but also the values of $v_i(t)$ for $t \leqq 0$, $i = 1, \cdots, n$.

Now, consideration of the problem in three-dimensional motion has led to the conclusion that accelerations should not be assumed continuous, but only integrable [2]. Thus it seems reasonable even in the case of one-dimensional motion to assume

that the given past history of v_i, say

$$v_i(t) = g_i(t) \quad \text{for } t \leqq 0, \quad i = 1. \cdots, n, \tag{11}$$

is merely absolutely continuous, not, in general, locally Lipschitzian.

Substituting (11) into the right-hand sides of (8) and (9), one gets a system of ordinary differential equations which satisfies the uniqueness criterion of the present paper. Thus, a unique solution exists at least as long as each $t - r_{ij}(t) \leqq 0$ and each $|v_j(t)| < 1$. (Further extension of the solution would use a "method-of-steps" argument which is not relevant to this paper.)

The above uniqueness problem was solved earlier for the case of two particles in one-dimensional motion [1]. But the method used did not seem to extend to the n-body problem.

REFERENCES

[1] R. D. DRIVER AND M. J. NORRIS, *Note on uniqueness for a one-dimensional two-body problem of classical electrodynamics*, Ann. Physics, 42 (1967), pp. 347–351.

[2] R. D. DRIVER, *Topologies for equations of neutral type and classical electrodynamics*, in Differencial'nye Uravnenija s Otklonjajuščimsja Argumentom, Naukova Dumka Kiev, 1977, pp. 113–127 (in Russian); transl. as Tech. Rept. 60, Dept. of Mathematics, University of Rhode Island, Kingston, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0002 $01.00/0

A STOCHASTIC MATRIX AND BILINEAR SUMS FOR RACAH-WILSON POLYNOMIALS*

MIZAN RAHMAN†

Abstract. A connection relation for the Racah-Wilson polynomials $W_n(x;\alpha_1+\beta_k-1,\ \alpha_2+\beta_2+\beta_3-\beta_k-1,\ \gamma+\beta_k-\beta_2,\ N)$, $k=1,2$, is obtained for a wide range of values of the parameters. Under certain restrictions on these parameters the corresponding matrix $K_N(x,y;\alpha_1,\beta_1,\beta_2,\beta_3,\alpha_2,\gamma,N)$ is shown to have stochastic properties. By using the spectral representation of the matrix K_N a set of bilinear sums is obtained for the W_n's.

1. Introduction. For a fixed positive integer N, the Racah-Wilson [8], [11], [13] polynomials $W_n(x)$ are defined by the balanced ${}_4F_3$ series

$$W_n(x)\equiv W_n(x;\alpha,\beta,\gamma,N)={}_4F_3\left[\begin{matrix}-n, & n+\alpha+\beta+1, & -x, & x+\gamma-N\\ & \alpha+1, & -N, & \beta+\gamma+1\end{matrix}\right], \tag{1.1}$$

$x, n=0,1,2,\cdots,N$. The parameters α, β, γ can have any real or complex values, except that $-N$ is the largest negative integer that occurs as a denominator parameter in the ${}_4F_3$ series above.

It was shown by Wilson [13] that $W_n(x)$ satisfies the orthogonality relations

$$\begin{aligned}&\sum_{x=0}^{N}\rho(x;\alpha,\gamma-N-\alpha-1,N+\alpha+\beta+1,N)W_m(x)W_n(x)\\&\quad=\frac{(\alpha+\beta+2)_N(-\gamma)_N}{(\beta+1)_N(\alpha-\gamma+1)_N}\cdot\frac{\delta_{mn}}{\rho(n;\alpha,\beta,\gamma,N)},\end{aligned} \tag{1.2}$$

and

$$\begin{aligned}&\sum_{n=0}^{N}\rho(n;\alpha,\beta,\gamma,N)W_n(x)W_n(y)\\&\quad=\frac{(\alpha+\beta+2)_N(-\gamma)_N}{(\beta+1)_N(\alpha-\gamma+1)_N}\cdot\frac{\delta_{xy}}{\rho(x;\alpha,\gamma-N-\alpha-1,N+\alpha+\beta+1,N)},\end{aligned} \tag{1.3}$$

where $y=0,1,\cdots,N$, and the weight function $\rho(x;a,b,c,N)$ is defined by

$$\rho(x;a,b,c,N)=\frac{(a+b+1)_x\left(1+\dfrac{a+b+1}{2}\right)_x(a+1)_x(b+c+1)_x(-N)_x}{x!\left(\dfrac{a+b+1}{2}\right)_x(b+1)_x(a-c+1)_x(N+a+b+2)_x}. \tag{1.4}$$

Equations (1.2) and (1.3) follow by use of Dougall's summation theorem [2, p. 25] for a very well-poised ${}_5F_4$ series:

$$\sum_{x=0}^{N}\rho(x;a,b,c,N)=\frac{(a+b+2)_N(-c)_N}{(b+1)_N(a-c+1)_N}, \tag{1.5}$$

and Whipple's transformation formula [2, p. 55] for a balanced ${}_4F_3$ series:

$${}_4F_3\left[\begin{matrix}\xi, & \eta, & \zeta, & -n\\ & u, & v, & w\end{matrix}\right]=\frac{(v-\zeta)_n(w-\zeta)_n}{(v)_n(w)_n}\cdot{}_4F_3\left[\begin{matrix}u-\xi, & u-\eta, & \zeta, & -n\\ & u, & 1+\zeta-v-n, & 1+\zeta-w-n\end{matrix}\right], \tag{1.6}$$

where $\xi+\eta+\zeta+1=u+v+w+n$.

* Received by the editors May 25, 1979. This work was supported by the Canadian Natural Sciences and Engineering Research Council under Grant A6197.

† Department of Mathematics, Carleton University, Ottawa, Ontario, Canada, KIS 5B6.

In a recent communication [11] the author showed that some of the well-known results for Hahn polynomials [5], [7] extend easily to Racah–Wilson polynomials. In particular, a projection formula and a nonnegative Poisson kernel under certain restrictions on the parameters were obtained.

The purpose of the present paper is to construct a positive kernel for Racah–Wilson polynomials in much the same way as one was obtained by the author for Hahn polynomials [10]. Specifically, we establish the connection relation

$$\sum_{y=0}^{N} K_N(x, y) W_n^{(2)}(y) = \lambda_n W_n^{(1)}(x), \tag{1.7}$$

where the eigenvalue λ_n is given by a balanced ${}_4F_3$:

$$\lambda_n = {}_4F_3\left[\begin{matrix} -n, & n+\alpha_1+\alpha_2+\beta_2+\beta_3-1, & \beta_2, & \beta_2+\beta_3-\beta_1 \\ & \alpha_1+\beta_2, & \beta_2+\beta_3, & \alpha_2+\beta_2+\beta_3-\beta_1 \end{matrix}\right], \tag{1.8}$$

and the matrix elements $K_N(x, y)$ are defined by the double sum:

$$\begin{aligned} K_N(x, y) &\equiv K_N(x, y; \alpha_1, \beta_1, \beta_2, \beta_3, \alpha_2, \gamma) \\ &= A(x, y) \sum_{i=0}^{x \wedge y} \sigma_i(\gamma - N - \beta_2; \alpha_1, \gamma - N + \beta_1 - \beta_2 + x, \gamma - N + y, \\ &\qquad \cdot 1 - \beta_2 - \beta_3 - N, \gamma + 1 - \beta_2, -x, -y) \\ &\quad \cdot \sum_{j=0}^{N - x \vee y} \sigma_j(-\gamma - N - \beta_3; \alpha_2, -\gamma - y, \beta_2 - \beta_1 - \gamma - x, 1 - \gamma - \beta_3 - i, \\ &\qquad \cdot 1 - \beta_2 - \beta_3 - N + i, x - N, y - N), \end{aligned} \tag{1.9}$$

where

$$\begin{aligned} A(x, y) &= \frac{(\beta_3)_N (\beta_2 + \beta_3 - \beta_1)_N}{(\beta_2 + \beta_3)_N (\alpha_2 + \beta_2 + \beta_3 - \beta_1)_N} \frac{(\beta_2 - \gamma)_N (\gamma + \alpha_2 + \beta_3)_N}{(-\gamma)_N (\gamma + \beta_3)_N} \\ &\quad \cdot \rho(y; \beta_2 - 1, \gamma - N - \beta_2, N + \beta_2 + \beta_3 - 1, N)(\beta_1)_x (\gamma + \beta_3)_x (\gamma - N + 1 - \alpha_1 - \beta_2)_x \\ &\quad \cdot \frac{(1 - N + \beta_1 - \alpha_2 - \beta_2 - \beta_3)_x}{(\gamma - N + 1 - \beta_2)_x (1 + \beta_1 - \beta_2 - \beta_3 - N)_x (\alpha_1 + \beta_1)_x (\gamma + \alpha_2 + \beta_3)_x}, \end{aligned} \tag{1.10}$$

$$\begin{aligned} &\sigma_i(a; b, c, d, e, f, g, h) \\ &= \frac{(a)_i \left(1 + \frac{a}{2}\right)_i (b)_i (c)_i (d)_i (e)_i (f)_i (g)_i (h)_i}{i! \left(\frac{a}{2}\right)_i (1 + a - b)_i (1 + a - c)_i (1 + a - d)_i (1 + a - e)_i (1 + a - f)_i (1 + a - g)_i (1 + a - h)_i}, \end{aligned} \tag{1.11}$$

with $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$, and

$$W_n^{(1)}(x) = W_n(x; \alpha_1 + \beta_1 - 1, \alpha_2 + \beta_2 + \beta_3 - \beta_1 - 1, \gamma + \beta_1 - \beta_2, N), \tag{1.12}$$

$$W_n^{(2)}(y) = W_n(y; \alpha_1 + \beta_2 - 1, \alpha_2 + \beta_3 - 1, \gamma, N). \tag{1.13}$$

Note that the σ's appearing in (1.9) are the general terms in a very well-poised terminating ${}_9F_8$ series. Also, every term in $K_N(x, y)$ is positive under the conditions

$$\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \beta_2 + \beta_3 - \beta_1 > 0, \qquad \gamma - N > \max(\alpha_i + \beta_j), \tag{1.14}$$

or

$$\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \alpha_2 + \beta_2 + \beta_3 - \beta_1 < -N, \qquad \gamma + \alpha_i + \beta_j > 0, \quad i = 1, 2, \quad j = 1, 2, 3. \tag{1.15}$$

For other values of the parameters the connection relation (1.7) still remains valid provided none of the shifted factorials in the denominators of K_N, λ_n, $W_n^{(1)}(x)$ or

$W_n^{(2)}(y)$ vanishes. However, for values other than those implied in the inequalities (1.14) or (1.15) the kernel $K_N(x, y)$ may be negative.

Since $W_n^{(1)}(x)$, $W_n^{(2)}(y)$, λ_n all equal one when $n = 0$, (1.7) implies that

$$\sum_{y=0}^{N} K_n(x, y) = 1, \tag{1.16}$$

and so $K_N(x, y)$ is a stochastic matrix when the inequalities (1.14) or (1.15) are satisfied.

It can be shown that the kernel $K_N(x, y)$ reduces to the one found in [10, eq. (1.10)] in the limit $\gamma \to \infty$. It is interesting to note that the eigenvalue λ_n is exactly the same as in [10, eq. (2.19)].

In [9] we proved that

$$\lambda_n = \frac{n!(\alpha_1+\beta_1)_n}{(\alpha_1+\beta_2)_n(\beta_2+\beta_3)_n} \sum_{k=0}^{n} \frac{(\alpha_1)_{n-k}(\beta_2)_k(\beta_3)_{n-k}(\alpha_2)_k}{k!(n-k)!(\alpha_1+\beta_1)_{n-k}(\alpha_2+\beta_2+\beta_3-\beta_1)_k}. \tag{1.17}$$

From this, the positivity of λ_n is rather obvious under (1.14). If we write $(\beta_2)_k = (\beta_2)_n/(\beta_2+k)_{n-k}$, it is also obvious under (1.15).

Positive sums and integrals of special functions in general, and of Jacobi and Hahn polynomials in particular, are known to be very useful [1], [12]. A good many theorems depend crucially on the nonnegativity of expansion coefficients and the sums of products of orthogonal polynomials (for a recent review of the subject with particular reference to Hahn polynomials see [6]). We believe that the positivity of the kernel $K_N(x, y)$ and that of the other sums that we will obtain in the sequel will be of some aid in the harmonic analysis of Racah–Wilson polynomials, analogous to that for Hahn polynomials. Furthermore, $K_N(x, y)$ has a physical significance in that it represents the transition probability of an ergodic Markov process in which a normalized weight function $\rho(x)$ corresponds to the stationary distribution. A model of such a stochastic process was, in fact, constructed for negative hypergeometric distributions in [3].

In § 2 we first prove (1.7) and then rewrite it in a properly normalized form. In § 3 we consider various special cases of $K_N(x, y)$. In § 4 we consider the corresponding bilinear formulas with particular reference to a Bateman-type sum. In § 5 we derive the Biedenharn–Elliot identity [4, p. 96], well known in quantum mechanics literature, as an application of our results.

2. Derivation of the connection relation (1.7). First, we apply (1.6) to obtain

$$W_n^{(2)}(y) = \frac{(\alpha_2+\beta_3)_n(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_n}{(\alpha_1+\beta_2)_n(-N)_n} W_n'(y), \tag{2.1}$$

where

$$W_n'(y) = \sum_{r=0}^{n} \frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r(\gamma+\alpha_2+\beta_3+y)_r(N+\alpha_2+\beta_3-y)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\gamma+\alpha_2+\beta_3)_r}. \tag{2.2}$$

Let us now compute the sum

$$L_1(i, j) = \sum_{y=i}^{j} \left(\frac{\gamma-N+2y}{\gamma-N+2i}\right) \frac{(\gamma-N+i)_y(\gamma-N+\beta_3+j)_y}{(\gamma-N+1+j)_y(\gamma-N+1-\beta_2+i)_y} \cdot (y-i+1)_{\beta_2-1}(j-y+1)_{\beta_3-1} W_n'(y). \tag{2.3}$$

It is convenient to use the following representation of the y-dependent factors on the r.h.s. of (2.2):

$$(\gamma+\alpha_2+\beta_3+y)_r(N+\alpha_2+\beta_3-y)_r = (\gamma+\alpha_2+2\beta_3+j)_r \sum_{k=0}^{r} \binom{r}{k} (j-y+\beta_3)_k(N+\alpha_2-j)_{r-k} \frac{(\gamma-N+\beta_3+y+j)_k}{(\gamma+\alpha_2+2\beta_3+j)_k}. \tag{2.4}$$

One can see that (2.4) is an immediate consequence of the Pfaff–Saalschutz summation theorem [1] which, in its general form, can be written as

$$(2.5)\qquad {}_3F_2\left[\begin{matrix}-r, & a, & b\\ c, & 1+a+b-c-r\end{matrix}\right]=\frac{(c-a)_r(c-b)_r}{(c)_r(c-a-b)_r},\qquad r=0,1,2,\cdots.$$

Now

$$\begin{aligned}(2.6)\quad &\sum_{y=i}^{j}\left(\frac{\gamma-N+2y}{\gamma-N+2i}\right)\frac{(\gamma-N+i)_y(\gamma-N+\beta_3+j)_y}{(\gamma-N+1+j)_y(\gamma-N+1-\beta_2+i)_y}(y-i+1)_{\beta_2-1}(j-y+1)_{\beta_3-1}\\ &\qquad\cdot(j-y+\beta_3)_k(\gamma-N+\beta_3+y+j)_k\\ &=\frac{\Gamma(\beta_2)(\gamma-N+i)_i(\gamma-N+\beta_3+j)_{i+k}(j-i+1)_{\beta_3-1+k}}{(\gamma-N+1+j)_i(\gamma-N+1-\beta_2+i)_i}\\ &\quad\cdot{}_5F_4\left[\begin{matrix}\gamma-N+2i, & 1+(\gamma-N/2)+i, & \beta_2, & \gamma-N+\beta_3+j+i+k, & i-j\\ & (\gamma-N/2)+i, & \gamma-N+1-\beta_2+2i, & 1-\beta_3+i-j-k, & \gamma-N+1+i+j\end{matrix}\right].\end{aligned}$$

But the ${}_5F_4$ series on the right is very well-poised and so, applying (1.5) and then simplifying the shifted factorials, we obtain the following expression for the r.h.s. of (2.6):

$$C(i,j)(\beta_3)_k(\gamma-N+\beta_3+i+j)_k(j-i+\beta_2+\beta_3)_k/(\beta_2+\beta_3)_k,$$

where

$$(2.7)\qquad \begin{aligned}&C(i,j)=B(\beta_2,\beta_3)\left(\frac{\gamma-N+i}{\gamma-N+2i}\right)\frac{(\gamma-N+1)_j(\gamma-N+\beta_3+j)_i(j-i+1)_{\beta_2+\beta_3-1}}{(\gamma-N+1)_i(\gamma-N+1-\beta_2+i)_j},\\ &B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b).\end{aligned}$$

Hence

$$(2.8)\qquad \begin{aligned}L_1(i,j)=C(i,j)&\sum_{r=0}^{n}\frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r(\gamma+\alpha_2+2\beta_3+j)_r(N+\alpha_2-j)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\gamma+\alpha_2+\beta_3)_r}\\ &\cdot{}_4F_3\left[\begin{matrix}-r, & \beta_3, & \gamma-N+\beta_3+i+j, & j-i+\beta_2+\beta_3\\ & \beta_2+\beta_3, & \gamma+\alpha_2+2\beta_3+j, & j+1-\alpha_2-N-r\end{matrix}\right].\end{aligned}$$

Applying (1.6) to the balanced ${}_4F_3$ on the right we get

$$\begin{aligned}&{}_4F_3\left[\begin{matrix}-r, & \beta_3, & \gamma-N+\beta_3+i+j, & j-i+\beta_2+\beta_3\\ & \beta_2+\beta_3, & \gamma+\alpha_2+2\beta_3+j, & j+1-\alpha_2-N-r\end{matrix}\right]\\ &=\frac{(\gamma+\alpha_2+\beta_3+j)_r(\beta_2)_r}{(\gamma+\alpha_2+2\beta_3+j)_r(\beta_2+\beta_3)_r}\cdot{}_4F_3\left[\begin{matrix}-r, & i+1-N-\alpha_2-\beta_2-\beta_3, & 1-\gamma-\alpha_2-\beta_3-i-r, & \beta_3\\ & j+1-\alpha_2-N-r, & 1-\gamma-\alpha_2-\beta_3-j-r, & 1-\beta_2-r\end{matrix}\right].\end{aligned}$$

This leads to

$$(2.9)\qquad \begin{aligned}L_1(i,j)=C(i,j)&\sum_{r=0}^{n}\frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\gamma+\alpha_2+\beta_3)_r(\beta_2+\beta_3)_r}\\ &\cdot\sum_{k=0}^{r}\binom{r}{k}(i+1-N-\alpha_2-\beta_2-\beta_3-r)_{r-k}(1-\gamma-\alpha_2-\beta_3-i-r)_{r-k}\\ &\cdot(\beta_2)_k(\beta_3)_{r-k}(\gamma+\alpha_2+\beta_3+j)_k(N+\alpha_2-j)_k.\end{aligned}$$

In the next stage we compute the following sum:

$$(2.10)\qquad \begin{aligned}L_2(i)\equiv&\sum_{j=x}^{N}\frac{(j-x+1)_{\beta_4-1}(N-j+1)_{\alpha_2-1}(\gamma+\alpha_2+\beta_3)_j(\gamma-N+\beta_3+x)_j}{(\gamma+\beta_3+1)_j(\gamma-N+1+\beta_3-\beta_4+x)_j}\\ &\cdot\left(\frac{\gamma-N+\beta_3+2j}{\gamma-N+\beta_3}\right)C^{-1}(i,j)L_1(i,j).\end{aligned}$$

Substituting (2.9) in (2.10) and applying (1.5) once again we obtain, after some simplification,

$$
\begin{aligned}
L_2(i) = D(x) \sum_{r=0}^{n} & \frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r(\beta_3)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\gamma+\alpha_2+\beta_3)_r(\beta_2+\beta_3)_r} \\
& \cdot \sum_{k=0}^{r} \frac{(-r)_k(\alpha_2)_k(\beta_2)_k(\gamma+\alpha_2+\beta_3+x)_k(\alpha_2+\beta_4+N-x)_k}{k!(\alpha_2+\beta_4)_k(1-\beta_3-r)_k} \\
& \cdot (N+\alpha_2+\beta_2+\beta_3+k-i)_{r-k}(\gamma+\alpha_2+\beta_3+k+i)_{r-k},
\end{aligned} \tag{2.11}
$$

where

$$
D(x) = B(\alpha_2, \beta_4)\left(\frac{\gamma-N+\beta_3+x}{\gamma-N+\beta_3}\right)\frac{(\gamma+\beta_3+1-N+x)_{N-x}(\gamma+\alpha_2+\beta_3)_x}{(\gamma+\beta_3-\beta_4+1-N+x)_N}(N-x+1)_{\alpha_2+\beta_4-1}. \tag{2.12}
$$

Finally, we need to compute

$$
\begin{aligned}
L(x) = D^{-1}(x) & \sum_{i=0}^{x} \frac{(\gamma-N-\beta_2)_i\left(1+\frac{\gamma-N-\beta_2}{2}\right)_i(\gamma-N+\beta_1-\beta_2+x)_i}{\left(\frac{\gamma-N-\beta_2}{2}\right)_i(\gamma-N+1-\beta_2+x)_i(\gamma-N+1-\alpha_1-\beta_2)_i} \\
& \cdot (i+1)_{\alpha_1-1}(x-i+1)_{\beta_1-1}L_2(i) \\
= \frac{\Gamma(\alpha_1)\Gamma(\beta_1+x)}{x!} & \sum_{r=0}^{n} \frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r(\beta_3)_r(N+\alpha_2+\beta_2+\beta_3)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\beta_2+\beta_3)_r} \\
& \cdot \sum_{k=0}^{r} \frac{(-r)_k(\alpha_2)_k(\beta_2)_k(\gamma+\alpha_2+\beta_3+x)_k(\alpha_2+\beta_4+N-x)_k}{k!(\alpha_2+\beta_4)_k(1-\beta_3-r)_k(N+\alpha_2+\beta_2+\beta_3)_k(\gamma+\alpha_2+\beta_3)_k} \\
& \cdot {}_7F_6\left[\begin{matrix} \gamma-N-\beta_2, & 1+(\gamma-N-\beta_2)/2, & \alpha_1, & \gamma-N+\beta_1-\beta_2+x, \\ & (\gamma-N-\beta_2)/2, & \gamma-N+1-\alpha_1-\beta_2, & 1-\beta_1-x, \\ 1-N-\alpha_2-\beta_2-\beta_3-k, & \gamma+\alpha_2+\beta_3+r, & -x & \\ \gamma+\alpha_2+\beta_3+k, & 1-N-\alpha_2-\beta_2-\beta_3-r, & \gamma-N+1+x-\beta_2 & \end{matrix}\right].
\end{aligned} \tag{2.13}
$$

The ${}_7F_6$ series on the right is not exactly summable, but it can be expressed in terms of a balanced ${}_4F_3$ series by virtue of Whipple's formula [2, p. 25] for a very well-poised ${}_7F_6$:

$$
\begin{aligned}
& {}_7F_6\left[\begin{matrix} a, & 1+a/2, & b, & c, & d, & e, & -m \\ & a/2, & 1+a-b, & 1+a-c, & 1+a-d, & 1+a-e, & 1+a+m \end{matrix}\right] \\
& = \frac{(1+a)_m(1-a-d-e)_m}{(1+a-d)_m(1+a-e)_m} \cdot {}_4F_3\left[\begin{matrix} 1+a-b-c, & d, & e, & -m \\ & 1+a-b, & 1+a-c, & d+e-a-m \end{matrix}\right].
\end{aligned} \tag{2.14}
$$

Thus we obtain

$$
\begin{aligned}
L(x) = B(\alpha_1, \beta_1)(x+1)_{\alpha_1+\beta_1-1}(\gamma-N-\beta_2+1)_x/(\gamma-N+1-\alpha_1-\beta_2)_x & \\
\cdot \sum_{r=0}^{n} \frac{(-n)_r(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_r(\beta_3)_r(N+\alpha_2+\beta_2+\beta_3)_r}{r!(\alpha_2+\beta_3)_r(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\beta_2+\beta_3)_r} & \\
\cdot \sum_{k=0}^{r} \frac{(-r)_k(\gamma+\alpha_2+\beta_3+x)_k(\alpha_2)_k(\beta_2)_k(\alpha_2+\beta_4+N-x)_k}{k!(\alpha_2+\beta_4)_k(1-\beta_3-r)_k(N+\alpha_2+\beta_2+\beta_3)_k(\gamma+\alpha_2+\beta_3)_k} & \\
\cdot {}_4F_3\left[\begin{matrix} k-r, & \alpha_1, & \gamma-N+\beta_1-\beta_2+x, & -x \\ & \gamma+\alpha_2+\beta_3+k, & 1-N-\alpha_2-\beta_2-\beta_3-r, & \alpha_1+\beta_1 \end{matrix}\right]. &
\end{aligned} \tag{2.15}
$$

A change of variable followed by an application of (1.6) finally leads to the following sum:

$$M_n(x) \equiv L(x)(\gamma - N + 1 - \alpha_1 - \beta_2)_x / B(\alpha_1, \beta_1)(x+1)_{\alpha_1+\beta_1-1}(\gamma - N - \beta_2 + 1)_x$$

$$= \sum_{r=0}^{n} \sum_{s=0}^{n-r} \frac{(-n)_{r+s}(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_{r+s}(\alpha_2)_r(\beta_2)_r}{r!\,s!\,(\alpha_2+\beta_3)_{r+s}(\beta_2+\beta_3)_{r+s}(\alpha_2+\beta_4)_r}$$

$$\cdot \frac{(\alpha_2+\beta_4+N-x)_r(\gamma+\alpha_2+\beta_3+x)_r(\beta_1)_s(\beta_3)_s}{(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_r(\gamma+\alpha_2+\beta_3)_r(\alpha_1+\beta_1)_s} \tag{2.16}$$

$$\cdot {}_4F_3\left[\begin{matrix} \gamma+\alpha_2+\beta_3+x+r, & \alpha_2+\beta_2+\beta_3-\beta_1+N-x+r, & \alpha_1, & -s \\ & \gamma+\alpha_2+\beta_3+r, & 1-\beta_1-s, & N+\alpha_1+\alpha_2+\beta_1+\beta_3+r \end{matrix}\right].$$

We shall now prove that

$$M_n(x) = \frac{(-N)_n(\alpha_1+\beta_2)_n}{(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_n(\alpha_2+\beta_3)_n} \lambda_n W_n^{(1)}(x), \tag{2.17}$$

provided

$$\beta_4 = \beta_2 + \beta_3 - \beta_1. \tag{2.18}$$

In [11] the author worked out the product of two W_n's, namely, $W_n(x; \alpha, \beta, \gamma, N) W_n(y; \alpha, \beta, \gamma', N')$, in a form very similar to the r.h.s. of (2.16). In fact, by a slight modification of the argument, (2.17) can be seen as following directly from (2.16) by virtue of the results presented in [11]. However, for the sake of completeness, it seems desirable to give an alternative proof here.

First, let us assume, for the moment, that

$$z \equiv x + \beta_1 - \beta_2 - \beta_3 - \alpha_2 - N \tag{2.19}$$

is a positive integer. We know that, in general, this is not true; indeed, it is strictly negative under the conditions (1.14). However, our plan is to prove (2.17) with this assumption and then claim that (2.17) must be true for all other values of z since the r.h.s. of (2.16) is a rational function of z.

Note that (1.6) enables us to carry out another set of transformations on the ${}_4F_3$ series in (2.16). Thus we get

$$M_n(x) = (-x)_z(N-\gamma+\alpha_1+\beta_2-x)_z / (\gamma+\alpha_2+\beta_3)_z(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_z$$

$$\cdot \sum_{r=0}^{n} \sum_{s=0}^{n-r} \frac{(-n)_{r+s}(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_{r+s}(\alpha_2+\beta_2+\beta_3-\beta_1+N-x)_r}{r!\,s!\,(\alpha_2+\beta_3)_{r+s}(\beta_2+\beta_3)_{r+s}(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_r} \tag{2.20}$$

$$\cdot \frac{(\gamma+\alpha_2+\beta_3+x)_r(\alpha_2)_r(\beta_2)_r(\beta_1)_s(\beta_3)_s}{(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_r(\alpha_2+\beta_2+\beta_3-\beta_1)_r(\alpha_1+\beta_1)_s}$$

$$\cdot {}_4F_3\left[\begin{matrix} 1-\beta_1, & 1-\alpha_1-\beta_1-s, & \gamma+\alpha_2+\beta_3+x+r, & \alpha_2+\beta_2+\beta_3-\beta_1+N-x+r \\ & 1-\beta_1-s, & N+\alpha_2+\beta_2+\beta_3-\beta_1+1+r, & \gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1+r \end{matrix}\right].$$

However,

$$\frac{(\alpha_2+\beta_2+\beta_3-\beta_1+N-x)_r(\gamma+\alpha_2+\beta_3+x)_r(\beta_1)_s}{(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_r(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_r(\alpha_1+\beta_1)_s} {}_4F_3[\,\bullet\,]$$

$$= \sum_{q=0}^{z-r} \frac{(-z)_{q+r}(\gamma+\alpha_2+\beta_3+x)_{q+r}(\beta_1-q)_s(1-\alpha_1-\beta_1-s)_q}{q!(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_{q+r}(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_{q+r}(\alpha_1+\beta_1)_s} \tag{2.21}$$

$$= (-1)^r \sum_{q=r}^{z} \frac{(-1)^q(\alpha_2+\beta_2+\beta_3-\beta_1+N-x)_q(\gamma+\alpha_2+\beta_3+x)_q(\beta_1-q+r)_s}{(q-r)!(\alpha_1+\beta_1)_{r+s-q}(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_q(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_q},$$

and

$$\frac{(\beta_2)_r(\beta_3)_s(\beta_1-q+r)_s}{s!(\beta_2+\beta_3)_{r+s}}=\sum_{j=r}^{r+s}\frac{(\beta_2)_j(\beta_1-\beta_2-q)_{r+s-j}(\beta_2+\beta_3-\beta_1+q)_{j-r}}{(j-r)!(r+s-j)!(\beta_2+\beta_3)_j}. \tag{2.22}$$

Equation (2.22) is a consequence of (2.5), as are the following:

$$\begin{aligned}&\sum_{s=j-r}^{n-r}\frac{(-n)_{r+s}(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_{r+s}(\beta_1-\beta_2-q)_{r+s-j}}{(r+s-j)!(\alpha_1+\beta_1)_{r+s-q}(\alpha_2+\beta_2)_{r+s}}\\&=\frac{(-n)_j(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_j}{(\alpha_1+\beta_1)_{n-q}(\alpha_2+\beta_3)_j}\frac{(\alpha_1+\beta_2+j)_{n-j}(1-\alpha_2-\beta_2-\beta_3+\beta_1-q-n)_{n-j}}{(1-\alpha_2-\beta_3-n)_{n-j}}\\&=\frac{(\alpha_1+\beta_2)_n(\alpha_2+\beta_2+\beta_3-\beta_1)_n}{(\alpha_2+\beta_3)_n(\alpha_1+\beta_1)_n}\frac{(-n)_j(n+\alpha_1+\alpha_2+\beta_2+\beta_3-1)_j}{(\alpha_1+\beta_2)_j}\\&\quad\cdot\frac{(-1)^q(1-\alpha_1-\beta_1-n)_q(\alpha_2+\beta_2+\beta_3-\beta_1+n)_q}{(\alpha_2+\beta_2+\beta_3-\beta_1)_{q+j}},\end{aligned} \tag{2.23}$$

and

$$\sum_{r=0}^{q}\frac{(-1)^{r+q}(\alpha_2)_r(\beta_2+\beta_3-\beta_1+q)_{j-r}}{(j-r)!(q-r)!r!(\alpha_2+\beta_2+\beta_3-\beta_1)_r}=\frac{(-1)^q(\alpha_2+\beta_2+\beta_3-\beta_1+q)_j(\beta_2+\beta_3-\beta_1)_j}{q!j!(\alpha_2+\beta_2+\beta_3-\beta_1)_j}. \tag{2.24}$$

Using (2.21) through (2.24) in (2.20) we obtain

$$\begin{aligned}M_n(x)&=(-x)_z(N-\gamma+\alpha_1+\beta_2-x)_z/(\gamma+\alpha_2+\beta_3)_z(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_z\\&\quad\cdot\frac{(\alpha_1+\beta_2)_n(\alpha_2+\beta_2+\beta_3-\beta_1)_n}{(\alpha_2+\beta_3)_n(\alpha_1+\beta_1)_n}\\&\quad\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_1+\alpha_2+\beta_2+\beta_3-1, & \beta_2, & \beta_2+\beta_3-\beta_1\\ & \alpha_1+\beta_2, & \beta_2+\beta_3, & \alpha_2+\beta_2+\beta_3-\beta_1\end{matrix};1\right]\\&\quad\cdot{}_4F_3\left[\begin{matrix}1-\alpha_1-\beta_1-n, & n+\alpha_2+\beta_2+\beta_3-\beta_1, & \gamma+\alpha_2+\beta_3+x, & N+\alpha_2+\beta_2+\beta_3-\beta_1-x\\ & \alpha_2+\beta_2+\beta_3-\beta_1, & N+\alpha_2+\beta_2+\beta_3-\beta_1+1, & \gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1\end{matrix};1\right].\end{aligned} \tag{2.25}$$

Making use of the assumption (2.19) once again, we get from (1.6)

$$\begin{aligned}&{}_4F_3\left[\begin{matrix}1-\alpha_1-\beta_1-n, & n+\alpha_2+\beta_2+\beta_3-\beta_1, & \gamma+\alpha_3+\beta_3+x, & N+\alpha_2+\beta_2+\beta_3-\beta_1-x\\ & \alpha_2+\beta_2+\beta_3-\beta_1, & N+\alpha_2+\beta_2+\beta_3-\beta_1+1, & \gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1\end{matrix};1\right]\\&=\frac{(N-\gamma-x-\beta_1+\beta_2+1)_z(1-\alpha_1-\beta_1+x)_z}{(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_z(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_z}\\&\quad\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_1+\alpha_2+\beta_2+\beta_3-1, & \gamma+\alpha_2+\beta_3+x, & -z\\ & \alpha_2+\beta_2+\beta_3-\beta_1, & \alpha_2+\beta_3+\gamma, & N+\alpha_1+\alpha_2+\beta_2+\beta_3\end{matrix};1\right]\\&=\frac{(N-\gamma-x-\beta_1+\beta_2+1)_z(1-\alpha_1-\beta_1-x)_z}{(N+\alpha_2+\beta_2+\beta_3-\beta_1+1)_z(\gamma+\alpha_2+\beta_3-\alpha_1-\beta_1+1)_z}\\&\quad\cdot\frac{(-N)_n(\alpha_1+\beta_1)_n}{(N+\alpha_1+\alpha_2+\beta_2+\beta_3)_n(\alpha_2+\beta_2+\beta_3-\beta_1)_n}W_n^{(1)}(x).\end{aligned} \tag{2.26}$$

One final set of simplifications involving the product of shifted factorials in z leads to the desired (2.17).

The successive operations that led to (2.16) clearly suggest how the matrix element $K_N(x, y)$ can be written out. A straightforward calculation yields

$$
\begin{aligned}
K_N(x,y) &= \{B(\alpha_1,\beta_1)B(\beta_2,\beta_3)B(\alpha_2,\beta_2+\beta_3-\beta_1)(x+1)_{\alpha_1+\beta_1-1}(N-x+1)_{\alpha_2+\beta_2+\beta_3-\beta_1-1}\}^{-1} \\
&\cdot \frac{\Gamma(\gamma+\alpha_2+\beta_3)\Gamma(\gamma-N-\beta_2+1)\Gamma(\gamma-N-\beta_2+1)\Gamma(\gamma+1+\beta_1-\beta_2+x)\Gamma(\gamma-N+1+x-\alpha_1-\beta_2)}{\Gamma(\gamma+\beta_3+1)\Gamma(\gamma-N+1-\alpha_1-\beta_2)\Gamma(\gamma-N+\beta_3)\Gamma(\gamma+\alpha_2+\beta_3+x)\Gamma(\gamma-N+\beta_1-\beta_2+x)} \\
&\cdot (\gamma-N+\beta_3)(\gamma-N+2y)\sum_{i=0}^{x\vee y}(i+1)_{\alpha_1-1}(x-i+1)_{\beta_1-1}(y-i+1)_{\beta_2-1}(\gamma-N-\beta_2)_i \\
&\cdot \frac{\left(1+\frac{\gamma-N-\beta_2}{2}\right)_i(\gamma-N+1-\beta_2+i+x)_{\beta_1-1}(\gamma-N+1-\beta_2+i+y)_{\beta_2-1}}{\left(\frac{\gamma-N-\beta_2}{2}\right)_i(\gamma-N+1-\alpha_1-\beta_2)_i} \\
&\cdot \sum_{j=x\vee y}^{N}(j-y+1)_{\beta_3-1}(j-x+1)_{\beta_2+\beta_3-\beta_1-1}(N-j+1)_{\alpha_2-1}(\gamma+\alpha_2+\beta_3)_j \\
&\cdot \frac{\left(1+\frac{\gamma-N+\beta_3}{2}\right)_j(\gamma-N+1+\beta_1-\beta_2+j+x)_{\beta_2+\beta_3-\beta_1-1}(\gamma-N+1+j+y)_{\beta_3-1}(\gamma-N+1-\beta_2)_{i+j}}{\left(\frac{\gamma-N+\beta_3}{2}\right)_j(\gamma+\beta_3+1)_j(\gamma-N+\beta_3)_{i+j}(j-i+1)_{\beta_2+\beta_3-1}}.
\end{aligned}
\tag{2.27}
$$

Written in this form, it is obvious how the kernel reduces, in the limit $\gamma \to \infty$, to the formulas (1.10) and (1.11) of [10]. Transforming this to the form (1.9) needs a bit of computation, but it is all very straightforward.

Having proved the connection relation (1.7) we shall now rewrite it in a properly normalized form. Let

$$
\begin{aligned}
\rho^{(1)}(x) &= \rho(x;\alpha_1+\beta_1-1,\gamma-N-\alpha_1-\beta_2,N+\alpha_1+\alpha_2+\beta_2+\beta_3-1,N), \\
\pi^{(1)}(n) &= \rho(n;\alpha_1+\beta_1-1,\alpha_2+\beta_2+\beta_3-\beta_1-1,\gamma+\beta_1-\beta_2,N), \\
\rho^{(2)}(x) &= \rho(x;\alpha_1+\beta_2-1,\gamma-N-\alpha_1-\beta_2,N+\alpha_1+\alpha_2+\beta_2+\beta_3-1,N), \\
\pi^{(2)}(n) &= \rho(n;\alpha_1+\beta_2-1,\alpha_2+\beta_3-1,\gamma,N).
\end{aligned}
\tag{2.28}
$$

Then the orthogonality relations for $W_n^{(1)}(x)$ and $W_n^{(2)}(x)$ take the form

$$
\sum_{x=0}^{N}\pi^{(1)}(n)\rho^{(1)}(x)W_m^{(1)}(x)W_n^{(1)}(x) = \frac{(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(\beta_2-\beta_1-\gamma)_N}{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\alpha_1+\beta_2-\gamma)_N}\delta_{mn},
\tag{2.29}
$$

and

$$
\sum_{x=0}^{N}\pi^{(2)}(n)\rho^{(2)}(x)W_m^{(2)}(x)W_n^{(2)}(x) = \frac{(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(-\gamma)_N}{(\alpha_2+\beta_3)_N(\alpha_1+\beta_2-\gamma)_N}\delta_{mn}.
\tag{2.30}
$$

Hence the normalized eigenfunctions

$$
\begin{aligned}
\phi_n^{(1)}(x) &= \left\{\frac{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\alpha_1+\beta_2-\gamma)_N}{(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(\beta_2-\beta_1-\gamma)_N}\rho^{(1)}(x)\pi^{(1)}(n)\right\}^{1/2}W_n^{(1)}(x), \\
\phi_n^{(2)}(x) &= \left\{\frac{(\alpha_2+\beta_3)_N(\alpha_1+\beta_2-\gamma)_N}{(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(-\gamma)_N}\rho^{(2)}(x)\pi^{(2)}(n)\right\}^{1/2}W_n^{(2)}(x)
\end{aligned}
\tag{2.31}
$$

have the property that

$$
\sum_{x=0}^{N}\phi_m^{(i)}(x)\phi_n^{(i)}(x) = \delta_{mn}, \qquad i=1,2,
\tag{2.32}
$$

and satisfy the connection relation

$$\sum_{y=0}^{N} G_N(x, y)\phi_n^{(2)}(y) = \mu_n \phi_n^{(1)}(x), \tag{2.33}$$

where

$$\mu_n = \{\pi^{(2)}(n)/\pi^{(1)}(n)\}^{1/2}\lambda_n = \left\{\frac{(\alpha_1+\beta_2)_n(\alpha_2+\beta_2+\beta_3-\beta_1)_n}{(\alpha_1+\beta_1)_n(\alpha_2+\beta_3)_n}\right\}^{1/2}\lambda_n, \tag{2.34}$$

and

$$\begin{aligned} G_N(x, y) &= \left\{\frac{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(-\gamma)_N}{(\alpha_2+\beta_3)_N(\beta_2-\beta_1-\gamma)_N}\frac{\rho^{(1)}(x)}{\rho^{(2)}(y)}\right\}^{1/2} K_N(x, y) \\ &= \omega_N f_N(x, y) \sum_{i=0}^{x\wedge y} \sigma_i(\gamma-N-\beta_2; \alpha_1, \gamma-N+\beta_1-\beta_2+x, \\ &\qquad \gamma-N+y, 1-\beta_2-\beta_3-N, \gamma+1-\beta_2, -x, -y) \\ &\quad \cdot \sum_{j=0}^{N-x\vee y} \sigma_j(-\gamma-N-\beta_3; \alpha_2, -\gamma-y, \beta_2-\beta_1-\gamma-x, \\ &\qquad 1-\gamma-\beta_3-i, 1-\beta_2-\beta_3-N+i, x-N, y-N), \end{aligned} \tag{2.35}$$

where

$$\omega_N = \left\{\frac{(\beta_3)_N(\beta_2+\beta_3-\beta_1)_N(\gamma+\alpha_2+\beta_3)_N(\beta_2-\gamma)_N}{(\beta_2+\beta_3)_N(\gamma+\beta_3)_N}\right\} \cdot\{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\alpha_2+\beta_3)_N(-\gamma)_N(\beta_2-\beta_1-\gamma)_N\}^{-1/2}, \tag{2.36}$$

and

$$\begin{aligned} f_N(x, y) &\equiv f_N(x, y; \alpha_1, \beta_1, \beta_2, \beta_3, \alpha_2, \gamma) \\ &= \{\rho(x; \gamma-N-\alpha_1-\beta_2, \alpha_1+\beta_1-1, 1-\alpha_1-\alpha_2-\beta_2-\beta_3-N, N) \\ &\quad \cdot \rho(y; \gamma-N-\alpha_1-\beta_2, \alpha_1+\beta_2-1, 1-\alpha_1-\alpha_2-\beta_2-\beta_3-N, N)\}^{1/2} \\ &\quad \cdot (\beta_1)_x(\beta_2)_y(\gamma+\beta_3)_x(\gamma+\beta_3)_y[(\gamma-N+1-\beta_2)_x(\gamma-N+1-\beta_2)_y \\ &\qquad \cdot (1+\beta_1-\beta_2-\beta_3-N)_x(1-\beta_3-N)_y]^{-1}. \end{aligned} \tag{2.37}$$

It is clear that the kernel $G_N(x, y)$ becomes symmetric in the special case $\beta_1 = \beta_2$, while the functions $\rho^{(1)}(x)$, $W_n^{(1)}(x)$, $\pi^{(1)}(n)$, μ_n coincide with $\rho^{(2)}(x)$, $W_n^{(2)}(x)$, $\pi^{(2)}(n)$ and λ_n, respectively.

3. Some special cases. Apart from the properties that we have mentioned in § 1, the kernels $K_N(x, y)$ and $G_N(x, y)$ have other interesting features. For example, they connect the balanced and terminating ${}_4F_3$ series with the ${}_9F_8$ functions in a very natural way. This becomes more explicit in some of the special cases we shall consider now.

Case I. $\alpha_1 \to 0$.

The double series on the right of (2.35) reduces to a single sum in this limit since

$$\lim_{\alpha_1\to 0} \sigma_i(\gamma-N-\beta_2; \alpha_1, \cdots, -y) = \delta_{i,0}.$$

Thus we get

$$\begin{aligned} &G_N(x, y); 0, \beta_1, \beta_2, \beta_3, \alpha_2, \gamma) \\ &= \omega_N f_N(x, y; 0, \beta_1, \beta_2, \beta_3, \alpha_2, \gamma) \\ &\quad \cdot {}_9F_8\left[\begin{matrix} -\gamma-N-\beta_3, & 1-(\gamma+N+\beta_3)/2, & \alpha_2, & -\gamma-y, \\ & -(\gamma+N+\beta_3)/2, & 1-\alpha_2-\beta_3-\gamma-N, & 1-\beta_3-N+y, \\ \beta_2-\beta_1-\gamma-x, & 1-\gamma-\beta_3, & 1-\beta_2-\beta_3-N, & x-N, \quad y-N \\ 1+\beta_1-\beta_2-\beta_3-N+x, & -N, & \beta_2-\gamma, & 1-\gamma-\beta_3-x, \quad 1-\gamma-\beta_3-y \end{matrix}\right]. \end{aligned} \tag{3.1}$$

Also

$$\lim_{\alpha_1\to 0}\mu_n=\frac{(\alpha_2)_n}{(\beta_2+\beta_3)_n}\left\{\frac{(\beta_1)_n(\beta_2)_n}{(\alpha_2+\beta_2+\beta_3-\beta_1)_n(\alpha_2+\beta_3)_n}\right\}^{1/2}, \tag{3.2}$$

$$\begin{aligned}\lim_{\alpha_1\to 0}\phi_n^{(1)}(x)=\Big\{&\frac{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\beta_2-\gamma)_N}{(\alpha_2+\beta_2+\beta_3)_N(\beta_2-\beta_1-\gamma)_N}\\ &\cdot\rho(x;\beta_1-1,\gamma-N-\beta_2,N+\alpha_2+\beta_2+\beta_3-1,N)\\ &\cdot\rho(n;\beta_1-1,\alpha_2+\beta_2+\beta_3-\beta_1-1,\gamma+\beta_1-\beta_2,N)\Big\}^{1/2}\\ &\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_2+\beta_2+\beta_3-1, & -x, & x+\gamma-N+\beta_1-\beta_2\\ & \beta_1, & -N, & \alpha_2+\beta_3+\gamma\end{matrix}\right]\end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\lim_{\alpha_1\to 0}\phi_n^{(2)}(x)=\Big\{&\frac{(\alpha_2+\beta_3)_N(\beta_2-\gamma)_N}{(\alpha_2+\beta_2+\beta_3)_N(-\gamma)_N}\rho(x;\beta_2-1,\gamma-N-\beta_2,N+\alpha_2+\beta_2+\beta_3-1,N)\\ &\cdot\rho(n;\beta_2-1,\alpha_2+\beta_3-1,\gamma,N)\Big\}^{1/2}\\ &\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_2+\beta_2+\beta_3-1, & -x, & x+\gamma-N\\ & \beta_2, & -N, & \alpha_2+\beta_3+\gamma\end{matrix}\right].\end{aligned} \tag{3.4}$$

It can be seen that the sum of the denominator parameters in the ${}_9F_8$ series above exceeds that of the numerator parameters by $2-2(\alpha_2-\beta_1+\beta_3)$. This suggests that we should look at the following special case.

Case II. $\alpha_1\to 0$, $\alpha_2=\beta_1-\beta_3$.

Setting this value of α_2 in the formulas (3.1) through (3.4) we obtain

$$\lim_{\substack{\alpha_1\to 0\\ \alpha_2=\beta_1-\beta_3}}\mu_n=\frac{(\beta_1-\beta_3)_n}{(\beta_2+\beta_3)_n}, \tag{3.5}$$

$$\begin{aligned}\lim_{\substack{\alpha_1\to 0\\ \alpha_2=\beta_1-\beta_3}}\phi_n^{(1)}(x)=\Big\{&\frac{(\beta_2)_N(\beta_2-\gamma)_N}{(\beta_1+\beta_2)_N(\beta_2-\beta_1-\gamma)_N}\rho(x;\beta_1-1,\gamma-N-\beta_2,N+\beta_1+\beta_2-1,N)\\ &\cdot\rho(n;\beta_1-1,\beta_2-1,\gamma+\beta_1-\beta_2,N)\Big\}^{1/2}\\ &\cdot{}_4F_3\left[\begin{matrix}-n, & n+\beta_1+\beta_2-1, & -x, & x+\gamma-N+\beta_1-\beta_2\\ & \beta_1, & -N, & \beta_1+\gamma\end{matrix}\right],\end{aligned} \tag{3.6}$$

$$\begin{aligned}\lim_{\substack{\alpha_1\to 0\\ \alpha_2=\beta_1-\beta_3}}\phi_n^{(2)}(x)=\Big\{&\frac{(\beta_1)_N(\beta_2-\gamma)_N}{(\beta_1+\beta_2)_N(-\gamma)_N}\rho(x;\beta_2-1,\gamma-N-\beta_2,N+\beta_1+\beta_2-1,N)\\ &\cdot\rho(n;\beta_2-1,\beta_1-1,\gamma,N)\Big\}\\ &\cdot{}_4F_3\left[\begin{matrix}-n, & n+\beta_1+\beta_2-1, & -x, & x+\gamma-N\\ & \beta_2, & -N, & \beta_1+\gamma\end{matrix}\right],\end{aligned} \tag{3.7}$$

$$\begin{aligned}&G_N(x,y;0,\beta_1,\beta_2,\beta_3,\beta_1-\beta_3,\gamma)\\ &\quad=\{(\beta_3)_N(\beta_2+\beta_3-\beta_1)_N(\gamma+\beta_1)_N(\beta_2-\gamma)_N/(\beta_2+\beta_3)_N(\gamma+\beta_3)_N\}\\ &\quad\cdot\{(\beta_2)_N(\beta_1)_N(-\gamma)_N(\beta_2-\beta_1-\gamma)_N\}^{-1/2}f_N(x,y;0,\beta_1,\beta_2,\beta_3,\beta_1-\beta_3,\gamma)\\ &\quad\cdot{}_9F_8\left[\begin{matrix}-\gamma-N-\beta_3, & 1-(\gamma+N+\beta_3)/2, & \beta_1-\beta_3, & -\gamma-y,\\ & -(\gamma+N+\beta_3)/2, & 1-\beta_1-\gamma-N, & 1-\beta_3-N+y,\\ \beta_2-\beta_1-\gamma-x, & 1-\gamma-\beta_3, & 1-\beta_2-\beta_3-N, & x-N, & y-N\\ 1+\beta_1-\beta_2-\beta_3-N+x, & -N, & \beta_2-\gamma, & 1-\gamma-\beta_3-x, & 1-\gamma-\beta_3-y\end{matrix}\right].\end{aligned} \tag{3.8}$$

The interesting property of this ${}_9F_8$ series is that it is 2-balanced, and hence we can apply Bailey's transformation formula [2, p. 63]:

$$
{}_9F_8\left[\begin{matrix} a, & 1+\frac{1}{2}a, & b, & c, & d, & e, & f, & g, & -n, \\ & \frac{1}{2}a, & 1+a-b, & 1+a-c, & 1+a-d, & 1+a-e, & 1+a-f, & 1+a-g, & 1+a+n \end{matrix}\right]
$$

$$
= \frac{(1+a)_n(1+a-b-c)_n(1+a-b-d)_n(1+a-b-e)_n(1+a-b-f)_n(g)_n}{(1+a-b)_n(1+a-c)_n(1+a-d)_n(1+a-e)_n(1+a-f)_n(g-b)_n}
$$

$$
\cdot {}_9F_8\left[\begin{matrix} b-g-n, & 1+\frac{1}{2}(b-g-n), & b, & 1+a-c-g, & 1+a-d-g, \\ & \frac{1}{2}(b-g-n), & 1-g-n, & b+c-a-n, & b+d-a-n, \\ & & & 1+a-e-g, & 1+a-f-g, \\ & & & b+e-a-n, & b+f-a-n, \end{matrix}\right.
$$

$$
\left.\begin{matrix} b-a-n, & -n \\ 1+a-g, & 1+b-g \end{matrix}\right], \tag{3.9}
$$

where $3a+2=b+c+d+e+f+g-n$.

Choosing $b=\beta_1-\beta_3$, $g=1-\gamma-\beta_3$, $n=\min(N-x, N-y)$ we then obtain, after some simplifications,

$$
G_N(x, y; 0, \beta_1, \beta_2, \beta_3, \beta_1-\beta_3, \gamma)
$$

$$
= \{(\beta_1)_N(\beta_2)_N/(\beta_2-\beta_1-\gamma)_N(-\gamma)_N\}^{1/2}\{(\beta_2+\beta_2-\beta_1-\gamma)_N/(\beta_2+\beta_3)_N\}g_N(x, y)
$$

$$
\cdot {}_9F_8\left[\begin{matrix} \gamma-N+\beta_1-1+x\vee y, & 1+(\gamma-N+\beta_1-1+x\vee y)/2, & \beta_1-\beta_3, & \gamma+y-N, \\ & (\gamma-N+\beta_1-1+x\vee y)/2, & \gamma-N+\beta_3+x\vee y, & \beta_1+\eta, \end{matrix}\right.
$$

$$
\left.\begin{matrix} \gamma+x-N+\beta_1-\beta_2, & \beta_2+\beta_3-1, & \gamma+\beta_1+x\vee y, & -x\wedge y, & x\vee y-N \\ \beta_2+\xi, & \gamma-N+1+\beta_1-\beta_2-\beta_3+x\vee y, & -N, & \gamma-N+\beta_1+x+y, & \gamma+\beta_1 \end{matrix}\right], \tag{3.10}
$$

where

$$
\xi = \begin{cases} y-x, & y\geqq x, \\ 0, & y<x; \end{cases}
$$

$$
\eta = \begin{cases} x-y, & x\geqq y, \\ 0, & x<y; \end{cases} \tag{3.11}
$$

and

$$
g_N(x, y) \equiv g_N(x, y; \beta_1, \beta_2, \beta_3, \beta_1-\beta_3, \gamma)
$$

$$
= \{(\beta_1)_x(\beta_2)_y(\beta_1+\gamma)_x(\beta_1+\gamma)_y(\gamma-N+\beta_3+x\vee y)_{x\wedge y}(\beta_2+\beta_3-\beta_1)_\xi(\beta_3)_\eta\}
$$

$$
\cdot \{(1-\beta_2-N)_x(1-\beta_1-N)_y(\gamma-N+\beta_1+x\vee y)_{x\wedge y}(\gamma-N+1-\beta_2)_{x\wedge y}
$$

$$
\cdot (\gamma-N+1+\beta_1-\beta_2-\beta_3)_{x\vee y}(\beta_2)_\xi(\beta_1)_\eta\}^{-1}
$$

$$
\cdot \{\rho(x; \gamma-N-\beta_2, \beta_1-1, 1-\beta_1-\beta_2-N, N)\rho(y; \gamma-N-\beta_2, \beta_2-1, 1-\beta_1-\beta_2-N, N)\}^{1/2}. \tag{3.12}
$$

In the next section we shall see that a particularly interesting situation arises when we allow $\beta_1-\beta_3$ to take only negative integral values.

Case III. $\alpha_2\to 0$.

Since $\lim_{\alpha_2\to 0}\sigma_j(-\gamma-N-\beta_3; \cdots, y-N)=\delta_{j,0}$, we have

$$
G_N(x, y; \alpha_1, \beta_1, \beta_2, \beta_3, 0, \gamma)
$$

$$
= \frac{(\beta_2-\gamma)_N}{(\beta_2+\beta_3)_N}\left\{\frac{(\beta_3)_N(\beta_2+\beta_3-\beta_1)_N}{(-\gamma)_N(\beta_2-\beta_1-\gamma)_N}\right\}^{1/2} f_N(x, y; \alpha_1, \beta_1, \beta_2, \beta_3, 0, \gamma)
$$

$$
\cdot {}_9F_8\left[\begin{matrix} \gamma-N-\beta_2, & 1+(\gamma-N-\beta_2)/2, & \alpha_1, & \gamma-N+\beta_1-\beta_2+x, \\ & (\gamma-N-\beta_2)/2, & \gamma-N+1-\alpha_1-\beta_2, & 1-\beta_1-x, \end{matrix}\right.
$$

$$
\left.\begin{matrix} \gamma-N+y, & 1-\beta_2-\beta_3-N, & \gamma+1-\beta_2, & -x, & -y \\ 1-\beta_2-y, & \gamma+\beta_3, & -N, & \gamma-N+1-\beta_2+x, & \gamma-N+1-\beta_2+y \end{matrix}\right]. \tag{3.13}
$$

The ${}_9F_8$ series is 2-balanced if $\alpha_1=\beta_3-\beta_1$. Thus the cases I and III are, in a sense, complementary to each other. If we are interested in a positive sum, then we may use (3.8) if $\beta_1>\beta_3$, while (3.13) is more appropriate when $\alpha_1=\beta_3-\beta_1>0$.

The eigenvalue μ_n has the simple form

$$\lim_{\alpha_2\to 0}\mu_n=\frac{(\alpha_1)_n(\beta_3)_n}{(\alpha_1+\beta_2)_n(\beta_2+\beta_3)_n}\left[\frac{(\alpha_1+\beta_2)_n(\beta_2+\beta_3-\beta_1)_n}{(\alpha_1+\beta_1)_n(\beta_3)_n}\right]^{1/2} \tag{3.14}$$

and

$$\lim_{\substack{\alpha_2\to 0\\ \alpha_1=\beta_3-\beta_1}}=\frac{(\beta_3-\beta_1)_n}{(\beta_2+\beta_3)_n}. \tag{3.15}$$

Case IV. $|\alpha_1|\to\infty$.

This represents another interesting situation since the Racah–Wilson polynomials become dual Hahn polynomials [7] in this limit. The weight functions $\rho^{(1)}(x)$, $\rho^{(2)}(x)$ as well as $K_N(x,y)$, $G_N(x,y)$ remain nonnegative if $\alpha_1,\beta_1,\beta_2,\beta_3<-N$, $\beta_2\leqq\beta_1$, and $-\alpha_2-\beta_3<\gamma$. If we denote by $\Psi_n^{(1)}(x)$ and $\Psi_n^{(2)}(x)$ the limits of $W_n^{(1)}(x)$ and $W_n^{(2)}(x)$, respectively, then we have the connection relation

$$\sum_{y=0}^{N}G_N(x,y;\infty,\beta_1,\beta_2,\beta_3,\alpha_2,\gamma)\Psi_n^{(2)}(y)=\nu_n\Psi_n^{(1)}(x), \tag{3.16}$$

with

$$\nu_n=\left\{\frac{(\alpha_2+\beta_2+\beta_3-\beta_1)_n}{(\alpha_2+\beta_3)_n}\right\}^{1/2}{}_3F_2\left[\begin{matrix}-n, & \beta_2, & \beta_2+\beta_3-\beta_1\\ \beta_2+\beta_3, & \alpha_2+\beta_2+\beta_3-\beta_1\end{matrix}\right], \tag{3.17}$$

$$\begin{aligned}\Psi_n^{(1)}(x)=&\left[\frac{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(-N)_n(\gamma+\alpha_2+\beta_3)_n}{(\beta_2-\beta_2-\gamma)_N n!(\alpha_2+\beta_2+\beta_3-\beta_1)_n}\tau^{(1)}(x)\right]^{1/2}\\ &\cdot{}_3F_2\left[\begin{matrix}-n, & -x, & x+\gamma-N+\beta_1-\beta_2\\ & -N, & \alpha_2+\beta_3+\gamma\end{matrix}\right],\end{aligned} \tag{3.18}$$

$$\Psi_n^{(2)}(y)=\left[\frac{(\alpha_2+\beta_3)_N(-N)_n(\gamma+\alpha_2+\beta_3)_n}{(-\gamma)_N n!(\alpha_2+\beta_3)_n}\tau^{(2)}(y)\right]^{1/2}\cdot{}_3F_2\left[\begin{matrix}-n, & -y, & y+\gamma-N\\ & -N, & \alpha_2+\beta_3+\gamma\end{matrix}\right], \tag{3.19}$$

$$\begin{aligned}&G_N(x,y;\infty,\beta_1,\beta_2,\beta_3,\alpha_2,\gamma)\\ &=\omega_N f_N(x,y;\infty,\beta_1,\beta_2,\beta_3,\alpha_2,\gamma)\sum_{i=0}^{x\wedge y}\frac{(\gamma-N-\beta_2)_i\left(1+\frac{\gamma-N-\beta_2}{2}\right)_i(\gamma-N+\beta_1-\beta_2+x)_i}{i!\left(\frac{\gamma-N-\beta_2}{2}\right)_i(1-\beta_1-x)_i(1-\beta_2-y)_i}\\ &\cdot\frac{(\gamma-N+y)_i(1-\beta_2-\beta_3-N)_i(\gamma+1-\beta_2)_i(-x)_i(-y)_i(-1)^i}{(\gamma+\beta_3)_i(-N)_i(\gamma-N+1-\beta_2+x)_i(\gamma-N+1-\beta_2+y)_i}\\ &\cdot\sum_{j=0}^{N-x\vee y}\sigma_j(-\gamma-N-\beta_3;\alpha_2,-\gamma-y,\beta_2-\beta_1-\gamma-x,1-\gamma-\beta_3-i,1-\beta_2-\beta_3-N+i,x-N,y-N),\end{aligned} \tag{3.20}$$

where

$$\tau^{(1)}(x)=\binom{N}{x}\frac{(\gamma-N+\beta_1-\beta_2)_x(\gamma+\alpha_2+\beta_3)_x(2x+\gamma+\beta_1-\beta_2)}{(1+\beta_1-\beta_2-\alpha_2-\beta_3-N)_x(\gamma+1+\beta_1-\beta_2)_x)(\gamma-N+\beta_1-\beta_2)}, \tag{3.21}$$

and $\tau^{(2)}(x)$ is the same as $\tau^{(1)}(x)$ with $\beta_1-\beta_2$ replaced by 0.

4. The bilinear sums. For finite N the orthonormal systems $\{\phi_n^{(k)}(x)\}$, $n=0,1,\cdots,N$, $k=1,2$, constitute complete bases for the $(N+1)$-dimensional vector space. The corresponding eigenvalues μ_n are simple and nonvanishing. Hence we have

the spectral representation of the matrix $G_N(x, y)$:

$$G_N(x, y; \alpha_1, \beta_1, \beta_2, \beta_3, \alpha_2, \gamma) = \sum_{n=0}^{N} \mu_n \phi_n^{(1)}(x)\phi_n^{(2)}(y). \tag{4.1}$$

Using (2.31) and (2.35) this can be written out as a very general bilinear sum for Racah–Wilson polynomials. However, it is the special cases considered in the previous section that seem to be more interesting at the moment, and so we shall attempt to write out the bilinear sums only for these limiting kernels.

Thus we have the following sum for the kernel of Case I:

$$\begin{aligned}
&{}_9F_8\left[\begin{matrix} -\gamma-N-\beta_3, & 1-(\gamma+N+\beta_3)/2, & \alpha_2, & -\gamma-y, & \beta_2-\beta_1-\gamma-x, \\ & -(\gamma+N+\beta_3)/2, & 1-\alpha_2-\beta_3-\gamma-N, & 1-\beta_3-N+y, & 1+\beta_1-\beta_2-\beta_3-N+x, \end{matrix}\right. \\
&\qquad\left.\begin{matrix} 1-\gamma-\beta_3, & 1-\beta_2-\beta_3-N, & x-N, & y-N \\ -N, & \beta_2-\gamma, & 1-\gamma-\beta_3-x, & 1-\gamma-\beta_3-y \end{matrix}\right] \\
&= \frac{(\beta_2+\beta_3)_N(\alpha_2+\beta_3)_N(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\gamma+\beta_3)_N}{(\beta_3)_N(\alpha_2+\beta_2+\beta_3)_N(\beta_2+\beta_3-\beta_1)_N(\gamma+\alpha_2+\beta_3)_N} \\
&\quad\cdot \frac{(\gamma+\alpha_2+\beta_3)_x(1+\beta_1-\beta_2-\beta_3-N)_x(1-\beta_3-N)_y(\gamma+\alpha_2+\beta_3)_y}{(\gamma+\beta_3)_x(1+\beta_1-\alpha_2-\beta_2-\beta_3-N)_x(1-\alpha_2-\beta_1-N)_y(\gamma+\beta_3)_y} \\
&\quad\cdot \sum_{n=0}^{N} \frac{(\alpha_2+\beta_2+\beta_3-1)_n\left(1+\frac{\alpha_2+\beta_2+\beta_3-1}{2}\right)_n (\beta_1)_n(\beta_2)_n(\alpha_2)_n(\gamma+\alpha_2+\beta_3)_n(-N)_n}{n!\left(\frac{\alpha_2+\beta_2+\beta_3-1}{2}\right)_n (\alpha_2+\beta_2+\beta_3-\beta_1)_n(\alpha_2+\beta_3)_n(\beta_2+\beta_3)_n(\beta_2-\gamma)_n(N+\alpha_2+\beta_2+\beta_3)_n} \\
&\qquad\cdot W_n(x; \beta_1-1, \alpha_2+\beta_2+\beta_3-\beta_1-1, \gamma+\beta_1-\beta_2, N) W_n(y; \beta_2-1, \alpha_2+\beta_3-1, \gamma, N).
\end{aligned} \tag{4.2}$$

When $\alpha_2 = \beta_1 - \beta_3$ this assumes a simpler form:

$$\begin{aligned}
&{}_9F_8\left[\begin{matrix} -\gamma-N-\beta_3, & 1-(\gamma+N+\beta_3)/2, & \beta_1-\beta_3, & -\gamma-y, & \beta_2-\beta_1-\gamma-x, \\ & -(\gamma+N+\beta_3)/2, & 1-\beta_1-\gamma-N, & 1-\beta_3-N+y, & 1+\beta_1-\beta_2-\beta_3-N+x, \end{matrix}\right. \\
&\qquad\left.\begin{matrix} 1-\gamma-\beta_3, & 1-\beta_2-\beta_3-N, & x-N, & y-N \\ -N, & \beta_2-\gamma, & 1-\gamma-\beta_3-x, & 1-\gamma-\beta_3-y \end{matrix}\right] \\
&= \frac{(\beta_2+\beta_3)_N(\gamma+\beta_3)_N}{(\beta_1+\beta_2)_N(\gamma+\beta_1)_N} \cdot \frac{(\beta_2)_{N-x}(\beta_1)_{N-y}(\gamma+\beta_1)_x(\gamma+\beta_1)_y}{(\beta_2+\beta_3-\beta_1)_{N-x}(\beta_3)_{N-y}(\gamma+\beta_3)_x(\gamma+\beta_3)_y} \\
&\quad\cdot \sum_{n=0}^{N} \frac{(\beta_1+\beta_2-1)_n\left(1+\frac{\beta_1+\beta_2-1}{2}\right)_n (\beta_1-\beta_3)_n(\gamma+\beta_1)_n(-N)_n}{n!\left(\frac{\beta_1+\beta_2-1}{2}\right)_n (\beta_2+\beta_3)_n(\beta_2-\gamma)_n(N+\beta_1+\beta_2)_n} \\
&\qquad\cdot W_n(x; \beta_1-1, \beta_2-1, \gamma+\beta_1-\beta_2, N) W_n(y; \beta_2-1, \beta_1-1, \gamma, N).
\end{aligned} \tag{4.3}$$

Setting $\beta_3 = \beta_1 + m$, $m \leqq N$, a nonnegative integer in (3.10), we also obtain

$$\begin{aligned}
&{}_9F_8\left[\begin{matrix} \gamma-N+\beta_1-1+x\vee y, & 1+(\gamma-N+\beta_1-1+x\vee y)/2, & -m, & m+\beta_1+\beta_2-1, \\ & (\gamma-N+\beta_1-1+x\vee y)/2, & \gamma-N+\beta_1+m+x\vee y, & \gamma-N+1-\beta_2-m+x\vee y, \end{matrix}\right. \\
&\qquad\left.\begin{matrix} \gamma+y-N, & \gamma+x-N+\beta_1-\beta_2, & \gamma+\beta_1+x\vee y, & -x\wedge y, & x\vee y-N \\ \beta_1+\eta, & \beta_2+\xi, & -N, & \gamma-N+\beta_1+x+y, & \gamma+\beta_1 \end{matrix}\right] \\
&= \frac{(\beta_1+\beta_2+m)_N(\beta_2-\gamma)_N}{(\beta_1+\beta_2)_N(\beta_2+m-\gamma)_N} \frac{(\beta_2)_\xi(\beta_1)_\eta}{(\beta_2+m)_\xi(\beta_1+m)_\eta} \frac{(\gamma-N+\beta_1+x\vee y)_{x\wedge y}(\gamma-N+1-\beta_2-m)_{x\vee y}}{(\gamma-N+1-\beta_2)_{x\vee y}(\gamma-N+\beta_1+m+x\vee y)_{x\wedge y}} \\
&\quad\cdot \sum_{n=0}^{m} \frac{(-m)_n(\beta_1+\beta_2-1)_n\left(1+\frac{\beta_1+\beta_2-1}{2}\right)_n (\gamma+\beta_1)_n(-N)_n}{n!(\beta_1+\beta_2+m)_n\left(\frac{\beta_1+\beta_2-1}{2}\right)_n (\beta_2-\gamma)_n(N+\beta_1+\beta_2)_n} \\
&\qquad\cdot W_n(x; \beta_1-1, \beta_2-1, \gamma+\beta_1-\beta_2, N) W_n(y; \beta_2-1, \beta_1-1, \gamma, N).
\end{aligned} \tag{4.4}$$

If we replace $N-x \vee y$ by $x \vee y$ in this formula and take the limit as $\gamma \to \infty$, we get [10, (4.10)]. Thus (4.4) is a generalization of Bateman's formula for Jacobi polynomials [1], [9]. An equivalent formula is, of course, obtained by setting $\beta_3 = \beta_1 + m$ in (4.3). An inverse formula can easily be found by multiplying both sides by $(-k)_m(k+\beta_1+\beta_2-1)_m/(\beta_1+\beta_2)_m m!$ and summing over m from 0 to k, $k = 0, 1, \cdots, N$. Equation (4.3) gives the following:

$$
\begin{aligned}
&W_k(x;\beta_1-1,\beta_2-1,\gamma+\beta_1-\beta_2,N)W_k(y;\beta_2-1,\beta_1-1,\gamma,N)\\
&\quad=\frac{(N+\beta_1+\beta_2)_k(\beta_2-\gamma)_k}{(-N)_k(\beta_1+\gamma)_k}\\
(4.5)\quad&\cdot\sum_{m=0}^{k}\frac{(-k)_m(k+\beta_1+\beta_2-1)_m(N-x+\beta_2)_m(N-y+\beta_1)_m(\gamma-\beta_1+x)_m(\gamma+\beta_1+y)_m}{m!(N+\beta_1+\beta_2)_m(\beta_1)_m(\beta_2)_m(\gamma+\beta_1)_m(N+\beta_1+\gamma)_m}\\
&\cdot{}_9F_8\left[\begin{matrix}-\gamma-N-\beta_1-m, & 1-(\gamma+N+\beta_1+m)/2, & -m, & -\gamma-y,\\ & -(\gamma+N+\beta_1+m)/2, & 1-\beta_1-\gamma-N, & 1-\beta_1-N+y-m,\\ 1-\beta_1-\beta_2-m-N, & \beta_2-\beta_1-\gamma-x, & 1-\gamma-\beta_1-m, & x-N, \quad y-N\\ \beta_2-\gamma, & 1-\beta_2-m-N+x, & -N, & 1-\gamma-\beta_1-x-m, \quad 1-\gamma-\beta_1-y-m\end{matrix}\right].
\end{aligned}
$$

5. An application. Starting from the general bilinear sum (4.1) we shall now derive an important identity which is known in the quantum mechanics literature as the Biedenharn–Elliot identity [4]. But first, we need some transformations. Whipple's formula (1.6) gives

$$
\begin{aligned}
&W_n^{(1)}(x)\equiv W_n(x;\alpha_1+\beta_1-1,\alpha_2+\beta_2+\beta_3-\beta_1-1,\gamma+\beta_1-\beta_2,N)\\
(5.1)\qquad&=(\gamma-N+1-\alpha_1-\beta_2)_x(1+\beta_1-\beta_2-\beta_3-\alpha_2-N)_x/(\alpha_1+\beta_1)_x(\gamma+\alpha_2+\beta_3)_x\\
&\quad\cdot W_{N-n}(x;-N-1,1-\alpha_1-\alpha_2-\beta_2-\beta_3-N,\gamma+\alpha_2+\beta_3-1,N+\alpha_2+\beta_2+\beta_2-\beta_1-1).
\end{aligned}
$$

We also need the projection formula [11]

$$
\begin{aligned}
&W_n(x;a+v,b-v,c+v,M)\\
(5.2)\qquad&=\frac{(c-M-a)_x(v)_x}{(a+v+1)_x(c-M+1)_x}\sum_{p=0}^{x}\rho(p;c-M-1+x+v,-x-v,a+x+v,x)W_n(p;a,b,c,M),
\end{aligned}
$$

where v is an arbitrary parameter. Thus

$$
\begin{aligned}
&W_{N-n}(x;v-N-1,1-\alpha_1-\alpha_2-\beta_2-\beta_3-N-v,\gamma+\alpha_2+\beta_3-1+v,N+\alpha_2+\beta_2+\beta_3-\beta_1-1)\\
(5.3)\qquad&=\frac{(\gamma+1+\beta_1-\beta_2)_x(v)_x}{(v-N)_x(\gamma-N+1+\beta_1-\beta_2)_x}\\
&\quad\cdot\sum_{p=0}^{x}\rho(p;\gamma-N+\beta_1-\beta_2-1+x+v,-x-v,x-N+v-1,x)\\
&\qquad\cdot W_{N-n}(p;-N-1,1-\alpha_1-\alpha_2-\beta_2-\beta_3-N,\gamma+\alpha_2+\beta_3-1,N+\alpha_2+\beta_2+\beta_3-\beta_1-1).
\end{aligned}
$$

We now replace x by p in (4.1), multiply both sides by

$$
\left\{\frac{(\gamma+1+\beta_1-\beta_2)_x(v)_x}{(v-N)_x(\gamma-N+1+\beta_1-\beta_2)_x}\right\}\rho(p;\gamma-N+\beta_1-\beta_2-1+x+v,-x-v,x-N+v-1,x),
$$

use (1.9) through (1.11) and (2.35), and sum over p from 0 to x. After some

simplifications we obtain

$$
\begin{aligned}
&\sum_{n=0}^{N} \rho(n;\alpha_1+\beta_2-1,\alpha_2+\beta_3-1,\gamma,N)\lambda_n W_n(y;\alpha_1+\beta_2-1,\alpha_2+\beta_3-1,\gamma,N)\\
&\quad\cdot W_{N-n}(x;v-N-1,1-\alpha_1-\alpha_2-\beta_2-\beta_3-N-v,\gamma+\alpha_2+\beta_3-1+v,N+\alpha_2+\beta_2+\beta_3-\beta_1-1)\\
&= a_N \frac{(v)_x(\gamma+1+\beta_1-\beta_2)_x(\beta_2)_y(\gamma+\beta_3)_y(1-\alpha_2-\beta_3-N)_y(\gamma-N+1-\alpha_1-\beta_2)_y}{(v-N)_x(\gamma-N+1+\beta_1-\beta_2)_x(\alpha_1+\beta_2)_y(\gamma-N+1-\beta_2)_y(1-\beta_3-N)_y(\gamma+\alpha_2+\beta_3)_y}\\
&\quad\cdot \sum_{i=0}^{y} \frac{(\gamma-N-\beta_2)_i\left(1+\frac{\gamma-N-\beta_2}{2}\right)_i (\alpha_1)_i(\gamma-N+y)_i(1-\beta_2-\beta_3-N)_i(\gamma+1-\beta_2)_i(-y)_i}{i!\left(\frac{\gamma-N-\beta_2}{2}\right)_i (\gamma-N+1-\alpha_1-\beta_2)_i(1-\beta_2-y)_i(\gamma+\beta_3)_i(-N)_i(\gamma-N+1-\beta_2+y)_i}\\
&\quad\cdot \frac{(\gamma-N+\beta_1-\beta_2)_{2i}\left(1+\frac{\gamma-N+\beta_1-\beta_2}{2}\right)_i (\gamma-N+\beta_1-\beta_2+x+v)_i(-N)_i(\gamma+\beta_3)_i(-x)_i}{\left(\frac{\gamma-N+\beta_1-\beta_2}{2}\right)_i (\gamma-N+1-\beta_2)_{2i}(1-x-v)_i(\gamma-N+1+\beta_1-\beta_2+x)_i}\\
&\qquad\qquad (1+\beta_1-\beta_2-\beta_3-N)_i(\gamma+1+\beta_1-\beta_2)_i\\
&\quad\cdot \sum_{j=0}^{N-y} \frac{(-\gamma-N-\beta_3)_j\left(1-\frac{\gamma+N+\beta_3}{2}\right)_j (\alpha_2)_j(-\gamma-y)_j(1-\gamma-\beta_3-i)_j(1-\beta_2-\beta_3-N+i)_j(y-N)_j}{j!\left(\frac{-\gamma-N-\beta_3}{2}\right)_j (1-\alpha_2-\beta_3-N-\gamma)_j(y-N+1-\beta_3)_j(i-N)_j(\beta_2-\gamma-i)_j(1-\gamma-\beta_3-y)_j}\\
&\qquad\cdot \frac{(i-N)_j(\beta_2-\beta_1-\gamma-i)_j}{(1+\beta_1-\beta_2-\beta_3-N+i)_j(1-\gamma-\beta_3-i)_j} h(i,j),
\end{aligned}
\tag{5.4}
$$

where

$$
h(i,j) = {}_7F_6\left[\begin{matrix}\gamma-N+\beta_1-\beta_2+2i, & 1+(\gamma+\beta_1-\beta_2+N)/2+i, & i+j-N, & \beta_1\\ & (\gamma+\beta_1-\beta_2+N)/2+i, & i-j+\gamma+1+\beta_1-\beta_2, & \gamma-N+1-\beta_2+2i,\\ \gamma+\beta_3+i-j, & \gamma-N+\beta_1-\beta_2+x+v+i, & i-x & \\ 1+\beta_1-\beta_2-\beta_3-N+i+j, & 1-v-x+i, & \gamma-N+1+\beta_1-\beta_2+x+i & \end{matrix}\right]
\tag{5.5}
$$

and

$$
a_N = \frac{(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(\beta_3)_N(\beta_2+\beta_3-\beta_1)_N(\alpha_2+\beta_3+\gamma)_N(\beta_2-\gamma)_N}{(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\alpha_2+\beta_3)_N(\beta_2+\beta_3)_N(\alpha_1+\beta_2-\gamma)_N(\beta_3+\gamma)_N}.
\tag{5.6}
$$

The sum of the denominator parameters in (5.5) minus that of the numerator parameters equal $4-2\beta_2-2\beta_3-2v$. Hence the ${}_7F_6$ series is 2-balanced if

$$
v = 1-\beta_2-\beta_3.
\tag{5.7}
$$

Setting this value of v in (5.5) we obtain, by Dougall's formula [2, p. 26],

$$
h(i,j) = \frac{(\gamma-N+1+\beta_1-\beta_2+2i)_{x-i}(\gamma+1-\beta_2+i-j)_{x-i}(1+\beta_1-\beta_2-\beta_3)_{x-i}(1-\beta_2-\beta_3-N+i+j)_{x-i}}{(\gamma+1+\beta_1-\beta_2+i-j)_{x-i}(\gamma-N+1-\beta_2+2i)_{x-i}(1+\beta_1-\beta_2-\beta_3-N+i+j)_{x-i}(1-\beta_2-\beta_3)_{x-i}},
\tag{5.8}
$$

Using this in (5.4) we find that the sums over i and j decouple and the individual sums become very well-poised ${}_7F_6$ series which are, therefore, expressible in terms of

balanced ${}_4F_3$'s by (2.14). Carrying out the necessary simplifications we get

$$
\begin{aligned}
&\left\{\frac{(\beta_1)_N(\beta_3)_N(\alpha_1+\alpha_2+\beta_2+\beta_3)_N(\gamma+\alpha_2+\beta_3)_N}{(\alpha_2+\beta_3)_N(\beta_2+\beta_3)_N(\alpha_2+\beta_2+\beta_3-\beta_1)_N(\alpha_1+\beta_2-\gamma)_N}\right\}\\
&\quad\cdot\left\{\frac{(\beta_2)_y(\beta_2+\beta_3-\beta_1)_y(1-\alpha_2-\beta_3-N)_y(\gamma+1-N-\alpha_1-\beta_2)_y}{(\alpha_1+\beta_2)_y(1-\beta_1-N)_y(1-\beta_3-N)_y(\gamma+\alpha_2+\beta_3)_y}\right\}\\
&\quad\cdot{}_4F_3\left[\begin{matrix}-x, & -y & 1-\alpha_1-\beta_2-y, & \gamma+1+\beta_1-2\beta_2-\beta_3-N+x\\ & 1-\beta_2-y, & 1+\beta_1-\beta_2-\beta_3-y, & \gamma-N+1-\alpha_1-\beta_2\end{matrix}\right]\\
&\quad\cdot{}_4F_3\left[\begin{matrix}y-N, & 1-\beta_2-\beta_3-N+x, & \beta_2-\beta_1-\gamma-x, & 1-\alpha_2-\beta_3-N+y\\ & 1-N-\beta_3+y, & 1-\alpha_2-\beta_3-N-\gamma, & 1-\beta_1-N+y\end{matrix}\right]\\
&\quad=\sum_{n=0}^{N}\rho(n;\alpha_1+\beta_2-1,\alpha_2+\beta_3-1,\gamma,N)\\
&\quad\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_1+\alpha_2+\beta_2+\beta_3-1, & -y, & y+\gamma-N\\ & \alpha_1+\beta_2, & -N, & \alpha_2+\beta_3+\gamma\end{matrix}\right]\\
&\quad\cdot{}_4F_3\left[\begin{matrix}n-N, & 1-\alpha_1-\alpha_2-\beta_2-\beta_3-N-n, & -x, & x+\gamma-N+1+\beta_1-2\beta_2-\beta_3\\ & 1-\beta_2-\beta_3-N, & 1+\beta_1-\alpha_2-\beta_2-\beta_3-N, & \gamma-N+1-\alpha_1-\beta_2\end{matrix}\right]\\
&\quad\cdot{}_4F_3\left[\begin{matrix}-n, & n+\alpha_1+\alpha_2+\beta_2+\beta_3-1, & \beta_2, & \beta_2+\beta_3-\beta_1\\ & \alpha_1+\beta_2, & \beta_2+\beta_3, & \alpha_2+\beta_2+\beta_3-\beta_1\end{matrix}\right].
\end{aligned}
\tag{5.9}
$$

This is the essential content of Biedenharn–Elliot identity which, in its original form, was written in the notation of Wigner's 6-j symbols [4].

REFERENCES

[1] R. ASKEY, *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] W. N. BAILEY, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[3] R. D. COOPER, M. R. HOARE AND MIZAN RAHMAN, *Stochastic processes and special functions: On the probabilistic origin of some positive kernels associated with classical orthogonal polynomials*, J. Math. Anal. Appl., 61 (1977), pp. 262–291.

[4] A. R. EDMONDS, *Angular Momentum in Quantum Mechanics*, 2nd edn., Princeton University Press, Princeton, NJ, 1960.

[5] G. GASPER, *Non-negativity of a discrete Poisson kernel for the Hahn polynomials*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.

[6] ———, *Positivity and special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375–434.

[7] S. KARLIN AND J. MCGREGOR, *The Hahn polynomials, formulas and an application*, Scripta Math., 26 (1961), pp. 33–46.

[8] G. RACAH, *Theory of complex spectra II*, Phys. Rev., 62 (1942), pp. 438–462.

[9] MIZAN RAHMAN, *A five-parameter family of positive kernels from Jacobi polynomials*, SIAM J. Math. Anal., 7 (1976), pp. 386–413.

[10] ———, *Some positive kernels and bilinear sums for Hahn polynomials*, this Journal, 7 (1976), pp. 414–435.

[11] ———, *A product formula and a non-negative Poisson kernel for Racah–Wilson polynomials*, Canad. J. Math., to appear.

[12] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications Vol. 23, 4th edn., Providence RI, 1975.

[13] J. A. WILSON, *Hypergeometric Series, Recurrence Relations and Some New Orthogonal Functions*, Ph.D. thesis, University of Wisconsin, Madison WI, 1978.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0003 $01.00/0

LINEARIZATION AND RELATED FORMULAS FOR q-ULTRASPHERICAL POLYNOMIALS*

D. M. BRESSOUD‡

Abstract. A new proof is given for Rogers' linearization formula for the q-ultraspherical polynomials. This proof leads to several new formulas relating q-ultraspherical polynomials. The principal result yields the following formula for the ultraspherical polynomials, $C_n^{\lambda}(x)$, when q approaches 1:

$$(1-2rx+r^2)^{-\lambda}(1-2sx+s^2)^{-\lambda}=\sum_{m,n=0}^{\infty}\binom{m+n}{n}\frac{\Gamma(\lambda+m)\Gamma(\lambda+n)}{\Gamma(\lambda)\Gamma(\lambda+m+n)}r^m s^n\,{}_2F_1\left[\begin{matrix}\lambda, 2\lambda+m+n\\ \lambda+m+n+1\end{matrix};rs\right]C_{m+n}^{\lambda}(x).$$

1. Introduction. The linearization problem for an arbitrary family of orthogonal polynomials, $\{P_n(x)\}$, is the problem of finding the coefficients $a(m, n, k)$ in the expansion:

$$P_m(x)P_n(x)=\sum_{k=0}^{m+n}a(m, n, k)P_k(x).$$

If $\{P_n(x)\}$ is orthonormal with respect to $\omega(x)$ on $(-1, 1)$, then

$$\int_{-1}^{1}P_k(x)P_m(x)P_n(x)\omega(x)\,dx=\begin{cases}a(m, n, k) & \text{if } k\leqq m+n,\\ 0 & \text{otherwise.}\end{cases}$$

Thus, the linearization problem is equivalent to evaluating this integral. This problem has appeared in several contexts, and is treated by Askey in [2, Chapter 5]. A use arising from a problem on convergence of solutions of finite difference approximations to the wave equation can be found in [3].

The solution to the linearization problem for the ultraspherical or Gegenbauer polynomials has generally been credited to Dougall [10]. It is not well known that twenty-four years earlier Rogers [17] had solved this problem. What is most remarkable about Rogers' result is that it is given not just for the ultraspherical polynomials, but for a much larger class of orthogonal polynomials which Askey and Ismail [4] have named the *continuous q-ultraspherical polynomials* (or, simply, *q-ultraspherical polynomials*). These polynomials in x with parameters β and q are denoted by $C_n(x;\beta|q)$ and can be defined by means of their recurrence relation:

$$C_{-1}(x;\beta|q)=0,\qquad C_0(x;\beta|q)=1,$$

$$2xC_n(x;\beta|q)=\frac{(1-q^{n+1})}{(1-\beta q^n)}C_{n+1}(x;\beta|q)+\frac{(1-\beta^2q^{n-1})}{(1-\beta q^n)}C_{n-1}(x;\beta|q). \tag{1.1}$$

If β is set equal to q^{λ} and then the limit is taken as q approaches 1, the recurrence relation becomes that of the ultraspherical polynomials, $C_n^{\lambda}(x)$, which are thus a special case of the q-ultraspherical polynomials.

The q-ultraspherical polynomials with $\beta=0$ are intimately connected with theta functions, and have been studied in that context by Szegö [18] and Carlitz [8], [9]. With general β, they yield natural explanations of the expansions of certain infinite products

* Received by the editors November 12, 1979, and in revised form July 17, 1980. This work was partially supported by the National Science Foundation under grants MCS 77-22992 and MCS 77-18723(02).

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802. This work was carried out while the author was at the School of Mathematics, Institute for Advanced Study, Princeton, New Jersey 08540.

(see Rogers [14], [15], [16] and [17], Bressoud [7]). They have also been studied by Fejér [11], Feldheim [12], Lanzewizky [13] and Szegö [19], whose interest in them arose from the fact that, up to certain normalizations, these are all of the orthogonal polynomials which satisfy

$$|f(re^{i\theta})|^2 = \sum_{n=0}^{\infty} P_n(\cos\theta) r^n,$$

where f is a function analytic in a neighborhood of the origin, having real coefficients in its power series expansion. A summary of some results on these polynomials can be found in [20, § 6.5]. It was Askey and Ismail [4] who drew attention to the fact that both groups of people were studying the same polynomials. Askey and Wilson [5] have found the weight functions for these polynomials.

The following notation will be used in this paper:

$$(a;q)_\infty = \prod_{i=0}^{\infty}(1-aq^i), \qquad (a;q)_n = \frac{(a;q)_\infty}{(aq^n;q)_\infty}.$$

(If n is a positive integer, then $(a;q)_n = (1-a)(1-aq)\cdots(1-aq^{n-1})$.)

$$ {}_2\varphi_1\left[\begin{matrix}a, b\\ c\end{matrix}; x\right] = \sum_{n=0}^{\infty} \frac{(a;q)_n (b;q)_n x^n}{(q;q)_n (c;q)_n}.$$

We also have the following limits (for n a positive integer).

$$\lim_{q\to 1} \frac{(q^\lambda;q)_n}{(1-q)^n} = (\lambda)(\lambda+1)\cdots(\lambda+n-1) \equiv (\lambda)_n = \frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}, \tag{1.2}$$

$$\lim_{q\to 1} {}_2\varphi_1\left[\begin{matrix}q^a, q^b\\ q^c\end{matrix}; x\right] = {}_2F_1\left[\begin{matrix}a, b\\ c\end{matrix}; x\right] \equiv \sum_{n=0}^{\infty} \frac{(a)_n (b)_n x^n}{n!(c)_n}. \tag{1.3}$$

By the q-binomial theorem [1, Thm. 2.1],

$$\sum_{n=0}^{\infty} \frac{(a;q)_n x^n}{(q;q)_n} = \frac{(ax;q)_\infty}{(x;q)_\infty}, \tag{1.4}$$

it is seen that

$$\lim_{q\to 1^-} \frac{(xq^\lambda;q)_\infty}{(x;q)_\infty} = (1-x)^{-\lambda}. \tag{1.5}$$

2. The linearization formula.

THEOREM 1.

$$\begin{aligned} &C_u(\cos\theta;\beta|q) C_v(\cos\theta;\beta|q) \\ &\quad = \sum_{p=0}^{\min(u,v)} \frac{(\beta;q)_{u-p}(\beta;q)_{v-p}(\beta;q)_p (q;q)_{u+v-2p} (\beta^2 q^{u+v-2p};q)_p}{(q;q)_{u-p}(q;q)_{v-p}(q;q)_p (\beta;q)_{u+v-2p} (\beta q^{u+v-2p+1};q)_p} \\ &\qquad \times C_{u+v-2p}(\cos\theta;\beta|q). \end{aligned} \tag{2.1}$$

This is the linearization formula for the q-ultraspherical polynomials, first stated by Rogers [16, p. 29] and proved by him using a "tedious" induction argument involving considerable manipulation of q-series. While our proof is nontrivial, it requires only the recurrence relation (1.1) and simple algebraic manipulations.

We first observe that (2.1) is equivalent to

$$\sum_{u,v=0}^{\infty} r^u s^v C_u(\cos\theta;\beta|q)C_v(\cos\theta;\beta|q)$$

$$= \sum_{m,n,p=0}^{\infty} \frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2 q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta p^{m+n+1};q)_p} \times r^{m+p}s^{n+p}C_{m+n}(\cos\theta;\beta|q). \tag{2.2}$$

It follows from the recurrence relation (see [4, § 2]) that

$$\frac{(r\beta e^{i\theta};q)_\infty(r\beta e^{-i\theta};q)_\infty}{(re^{i\theta};q)_\infty(re^{-i\theta};q)_\infty} = \sum_{m=0}^{\infty} C_m(\cos\theta;\beta|q)r^m. \tag{2.3}$$

Therefore, (2.1) is equivalent to the following:

$$\frac{(r\beta e^{i\theta};q)_\infty(r\beta e^{-i\theta};q)_\infty(s\beta e^{i\theta};q)_\infty(s\beta e^{-i\theta};q)_\infty}{(re^{i\theta};q)_\infty(re^{-i\theta};q)_\infty(se^{i\theta};q)_\infty(se^{-i\theta};q)_\infty}$$

$$= \sum_{m,n=0}^{\infty} \frac{(\beta;q)_m(\beta;q)_n(q;q)_{m+n}r^m s^n}{(q;q)_m(q;q)_n(\beta;q)_{m+n}} {}_2\varphi_1\left[\begin{matrix}\beta, \beta^2 q^{m+n} \\ \beta q^{m+n+1}\end{matrix}; rs\right] C_{m+n}(\cos\theta;\beta|q), \tag{2.4}$$

$$|q|<1.$$

This is the result we shall actually prove. If β is replaced by q^λ in (2.1) and (2.4) and then q approaches 1, we get the following formulas for ultraspherical polynomials:

COROLLARY 1.

$$C_u^\lambda(x)C_v^\lambda(x)$$

$$= \sum_{p=0}^{\min(u,v)} \binom{u+v-2p}{u-p} \frac{\Gamma(\lambda+u-p)\Gamma(\lambda+v-p)(\lambda)_p(2\lambda+u+v-2p)_p}{\Gamma(\lambda)\Gamma(\lambda+u+v-2p)p!(\lambda+u+v-2p+1)_p} C_{u+v-2p}^\lambda(x), \tag{2.5}$$

$$(1-2rx+r^2)^{-\lambda}(1-2sx+s^2)^{-\lambda}$$

$$= \sum_{m,n=0}^{\infty} \binom{m+n}{n} \frac{\Gamma(\lambda+m)\Gamma(\lambda+n)}{\Gamma(\lambda)\Gamma(\lambda+m+n)} r^m s^n {}_2F_1\left[\begin{matrix}\lambda, 2\lambda+m+n \\ \lambda+m+n+1\end{matrix}; rs\right] C_{m+n}^\lambda(x). \tag{2.6}$$

When $\beta = 0$, (2.4) becomes

COROLLARY 2.

$$\frac{(rs;q)_\infty}{(re^{i\theta};q)_\infty(re^{-i\theta};q)_\infty(se^{i\theta};q)_\infty(se^{-i\theta};q)_\infty} = \sum_{m,n=0}^{\infty} \frac{(q;q)_{m+n}r^m s^n}{(q;q)_m(q;q)_n} C_{m+n}(\cos\theta;0|q). \tag{2.7}$$

Equation (2.7) was known to and extensively used by Rogers. In [6], this author has given a very simple proof of (2.7) which uses only the recurrence relation for $C_n(\cos\theta;0|q)$. Our proof of (2.4) will precisely parallel this proof of (2.7).

3. Proof of Theorem 1. We define:

$$f(r,s)=\sum_{m,n=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(q;q)_{m+n}r^m s^n}{(q;q)_m(q;q)_n(\beta;q)_{m+n}}\times{}_2\varphi_1\left[\begin{matrix}\beta,\beta^2q^{m+n}\\ \beta q^{m+n+1}\end{matrix};rs\right]C_{m+n}(x;\beta|q)$$

$$=\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n+1});q)_p}\times r^{m+p}s^{n+p}C_{m+n}(x;\beta|q). \tag{3.1}$$

We then have

$$2x(f(r,s)-\beta f(rq,s))$$

$$=\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n+1};q)_p}r^{m+p}s^{n+p}\times(1-\beta q^{m+p})(2xC_{m+n}(x;\beta|q)). \tag{3.2}$$

From the recurrence relation (1.1), the right-hand side of (3.2) can be rewritten as

$$\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n+1};q)_p}r^{m+p}s^{n+p}\times\frac{(1-q^{m+n+1})}{(1-\beta q^{m+n})}(1-\beta q^{m+p})C_{m+n+1}(x;\beta|q)$$

$$+\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n+1};q)_p}r^{m+p}s^{n+p}\times\left(\frac{1-\beta^2q^{m+n-1}}{(1-\beta q^{m+n})}\right)(1-\beta q^{m+p})C_{m+n-1}(x;\beta|q)$$

$$=\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n+1}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n+1}(\beta q^{m+n+1};q)_p}r^{m+p}s^{n+p}$$

$$\times\left\{\frac{(1-\beta q^m)(1-\beta^2q^{m+n+p})(1-\beta q^{m+n+1})(1-q^{m+p+1})}{(1-q^{m+1})(1-\beta^2q^{m+n})(1-\beta q^{m+n+p+1})}\right.$$

$$\left.-q^{m+1}\frac{(1-\beta q^n)(1-q^p)(1-\beta/q)(1-\beta^2q^{2m+n+p+1})}{(1-q^{m+1})(1-\beta^2q^{m+n})(1-\beta q^{m+n+p+1})}\right\}$$

$$\times C_{m+n+1}(x;\beta|q)$$

$$+\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(\beta;q)_p(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n+1};q)_p}r^{m+p}s^{n+p}$$

$$\times\left\{\frac{(1-\beta q^{m+n-1})(1-q^m)(1-\beta^2q^{m+n-1})(1-\beta q^{m+n+p})(1-\beta^2q^{m+p-1})}{(1-q^{m+n})(1-\beta q^{m-1})(1-\beta^2q^{m+n+p-1})(1-\beta q^{m+n})}\right.$$

$$\left.+q^m\frac{(1-\beta^2q^{m+n-1})(1-q^n)(1-\beta q^p)(1-\beta/q)(1-\beta^2q^{2m+n+p-1})}{(1-\beta q^{m+n})(1-\beta^2q^{m+n+p-1})(1-q^{m+n})(1-\beta q^{m-1})}\right\}$$

$$+C_{m+n-1}(x;\beta|q). \tag{3.3}$$

This last line results from the algebraic identities

$$(1-\beta q^{m+p}) = \frac{(1-\beta q^m)(1-\beta^2 q^{m+n+p})(1-\beta q^{m+n+1})(1-q^{m+p+1})}{(1-q^{m+1})(1-\beta^2 q^{m+n})(1-\beta q^{m+n+p+1})}$$
$$-q^{m+1}\frac{(1-\beta q^n)(1-q^p)(1-\beta/q)(1-\beta^2 q^{2m+n+p+1})}{(1-q^{m+1})(1-\beta^2 q^{m+n})(1-\beta q^{m+n+p+1})}$$
$$=\frac{(1-\beta q^{m+n-1})(1-q^m)(1-\beta q^{m+n+p})(1-\beta^2 q^{m+p-1})}{(1-q^{m+n})(1-\beta q^{m-1})(1-\beta^2 q^{m+n+p-1})}$$
$$+q^m\frac{(1-q^n)(1-\beta q^p)(1-\beta/q)(1-\beta^2 q^{2m+n+p-1})}{(1-q^{m+n})(1-\beta q^{m-1})(1-\beta^2 q^{m+n+p-1})}.$$

Each sum of the right-hand side of (3.3) is now split into two sums. It can be observed that the second of these four sums is zero when $p=0$, the third is zero when $m=0$, and the fourth is zero when $n=0$. Furthermore, the first and second sums are of equal magnitude and opposite sign if m is set equal to -1. The second equality of (3.4) results when the first and third sums are each written as a sum of two sums,

(3.4)

$$\begin{aligned}
&2x(f(r,s)-\beta f(rq,s))\\
&=\sum_{\substack{n,p=0\\ m=-1}}^{\infty}\frac{(\beta;q)_{m+1}(\beta;q)_n(\beta;q)_p(q;q)_{m+n+1}(\beta^2q^{m+n+1};q)_p}{(q;q)_{m+1}(q;q)_n(q;q)_p(\beta;q)_{m+n+1}(\beta q^{m+n+2};q)_p}\\
&\qquad\times r^{m+p}s^{n+p}(1-q^{m+p+1})C_{m+n+1}(x;\beta|q)\\
&\quad-\sum_{\substack{m=-1\\ n=0\\ p=1}}^{\infty}\frac{(\beta;q)_m(\beta;q)_{n+1}(\beta;q)_p(q;q)_{m+n+1}(\beta^2q^{m+n+1};q)_{p-1}}{(q;q)_{m+1}(q;q)_n(q;q)_{p-1}(\beta;q)_{m+n+1}(\beta q^{m+n+1};q)_{p+1}}\\
&\qquad\times r^{m+p}s^{n+p}q^{m+1}(1-\beta/q)(1-\beta^2q^{2m+n+p+1})C_{m+n+1}(x;\beta|q)\\
&\quad+\sum_{\substack{n,p=0\\ m=1}}^{\infty}\frac{(\beta;q)_{m-1}(\beta;q)_n(\beta;q)_p(q;q)_{m+n-1}(\beta^2q^{m+n-1};q)_p}{(q;q)_{m-1}(q;q)_n(q;q)_p(\beta;q)_{m+n-1}(\beta q^{m+n};q)_p}\\
&\qquad\times r^{m+p}s^{n+p}(1-\beta^2q^{m+p-1})C_{m+n-1}(x;\beta|q)\\
&\quad+\sum_{\substack{m,p=0\\ n=1}}^{\infty}\frac{(\beta;q)_{m-1}(\beta;q)_n(\beta;q)_{p+1}(q;q)_{m+n-1}(\beta^2q^{m+n-1};q)_p}{(q;q)_m(q;q)_{n-1}(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n};q)_{p+1}}\\
&\qquad\times r^{m+p}s^{n+p}q^m(1-\beta/q)(1-\beta^2q^{2m+n+p-1})C_{m+n-1}(x;\beta|q)\\
&=r^{-1}f(r,s)-r^{-1}f(rq,s)\\
&\quad-s\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_{m-1}(\beta;q)_{n+1}(\beta;q)_{p+1}(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n};q)_{p+2}}\\
&\qquad\times r^{m+p}s^{n+p}q^m(1-\beta/q)(1-\beta^2q^{2m+n+p})C_{m+n}(x;\beta|q)\\
&\quad+rf(r,s)-r\beta^2f(rq,s)\\
&\quad+s\sum_{m,n,p=0}^{\infty}\frac{(\beta;q)_{m-1}(\beta;q)_{n+1}(\beta;q)_{p+1}(q;q)_{m+n}(\beta^2q^{m+n};q)_p}{(q;q)_m(q;q)_n(q;q)_p(\beta;q)_{m+n}(\beta q^{m+n};q)_{p+2}}\\
&\qquad\times r^{m+p}s^{n+p}q^m(1-\beta/q)(1-\beta^2q^{2m+n+p})C_{m+n}(x;\beta|q)\\
&=f(r,s)(r^{-1}+r)-f(rq,s)(r^{-1}+r\beta^2).
\end{aligned}$$

This equation can be rewritten as

$$f(r,s)=\frac{(1-2x\beta r+\beta^2r^2)}{(1-xr+r^2)}f(rq,s). \tag{3.5}$$

Since f is symmetric in its variables, we also have

$$f(rq,s)=\frac{(1-2x\beta x+\beta^2s^2)}{(1+xs+s^2)}f(rq,sq). \tag{3.6}$$

If (3.5) and (3.6) are combined and x is set equal to $\cos\theta$, then we get

$$\begin{aligned}
f(r,s)&=\frac{(1-2\beta r\cos\theta+r^2)(1-2\beta s\cos\theta+s^2)}{(1-2r\cos\theta+r^2)(1-2s\cos\theta-s^2)}f(rq,sq)\\
&=\frac{(1-r\beta e^{i\theta})(1-r\beta e^{-i\theta})(1-s\beta e^{i\theta})(1-s\beta e^{-i\theta})}{(1-re^{i\theta})(1-re^{-i\theta})(1-se^{i\theta})(1-se^{-i\theta})}f(rq,sq)\\
&=\frac{(r\beta e^{i\theta};q)_2(r\beta e^{-i\theta};q)_2(s\beta e^{i\theta};q)_2(s\beta e^{-i\theta};q)_2}{(re^{i\theta};q)^2(re^{-i\theta};q)_2(se^{i\theta};q)_2(se^{-i\theta};q)_2}f(rq^2,sq^2)\\
&\quad\vdots\\
&=\frac{(r\beta e^{i\theta};q)_n(r\beta e^{-i\theta};q)_n(s\beta e^{i\theta};q)_n(s\beta e^{-i\theta};q)_n}{(re^{i\theta};q)_n(re^{-i\theta};q)_n(se^{i\theta};q)_n(se^{-i\theta};q)_n}f(rq^n,sq^n)\\
&\quad\vdots\\
&=\frac{(r\beta e^{i\theta};q)_\infty(r\beta e^{-i\theta};q)_\infty(s\beta e^{i\theta};q)_\infty(s\beta e^{-i\theta};q)_\infty}{(re^{i\theta};q)_\infty(re^{-i\theta};q)_\infty(se^{i\theta};q)_\infty(se^{-i\theta};q)_\infty}f(0,0),
\end{aligned} \tag{3.7}$$

since $|q|<1$.

From the definition of $f(r,s)$, (3.1), it is clear that $f(0,0)=1$. This proves (2.4), which was shown to be equivalent to the theorem.

4. Additional corollaries. In (2.4), we set $s=rq^{1/2}$ and obtain

$$\begin{aligned}
&\frac{(r\beta e^{i\theta};q^{1/2})_\infty(r\beta e^{-i\theta};q^{1/2})_\infty}{(re^{i\theta};q^{1/2})_\infty(re^{-i\theta};q^{1/2})_\infty}+\sum_{n=0}^{\infty}C_n(\cos\theta;\beta|q^{1/2})r^n\\
&=\sum_{m,n=0}^{\infty}\frac{(\beta;q)_m(\beta;q)_n(q;q)_{m+n}r^{m+n}q^{1/2}}{(q;q)_m(q;q)_n(\beta;q)_{m+n}}\,{}_2\varphi_1\left[\begin{matrix}\beta,\beta^2q^{m+n}\\ \beta q^{m+n+1}\end{matrix};r^2q^{1/2}\right]C_{m+n}(\cos\theta;\beta|q),
\end{aligned} \tag{4.1}$$

which is equivalent to

COROLLARY 3.

$$\begin{aligned}
C_n(\cos\theta;\beta|q^{1/2})=\sum_{p=0}^{[n/2]}&\frac{(\beta;q^{1/2})_{n-2p}(q;q)_{n-2p}(\beta;q)_p(\beta^2q^{n-2p};q)_p}{(q^{1/2};q^{1/2})_{n-2p}(\beta;q)_{n-2p}(q;q)_p(\beta q^{n-2p+1};q)_p}\\
&\times q^{p-2}C_{n-2p}(\cos\theta;\beta|q).
\end{aligned} \tag{4.2}$$

In the derivation of (4.2) from (4.1), we note that

$$\sum_{t=0}^{\infty} x^t \sum_{m=0}^{t} \frac{(\beta;q)_m(\beta;q)_{t-m}q^{(t-m)/2}}{(q;q)_m(q;q)_{t-m}} = \sum_{m=0}^{\infty} \frac{(\beta;q)_m}{(q;q)_m} x^m \sum_{n=0}^{\infty} \frac{(\beta;q)_n}{(q;q)_n}(xq^{1/2})^n$$

$$= \frac{(\beta x;q)_\infty}{(x;q)_\infty} \frac{(\beta x q^{1/2};q)_\infty}{(xq^{1/2};q)_\infty} \qquad \text{(by (1.4))}$$

$$= \frac{(\beta x;q^{1/2})_\infty}{(x;q^{1/2})_\infty}$$

$$= \sum_{t=0}^{\infty} \frac{(\beta;q^{1/2})_t}{(q^{1/2};q^{1/2})_t} x^t,$$

and thus

$$\sum_{m=0}^{t} \frac{(\beta;q)_m(\beta;q)_{t-m}q^{(t-m)/2}}{(q;q_m(q;q)_{t-m}} = \frac{(\beta;q^{1/2})_t}{(q^{1/2};q^{1/2})_t}.$$

By a similar argument with $s = -r$, we get:

COROLLARY 4.

$$C_n(\cos 2\theta;\beta^2|q^2) = \sum_{p=0}^{n} \frac{(\beta^2;q^2)_{n-p}(\beta;q)_p(q;q)_{2n-2p}(\beta^2 q^{2n-2p};q)_p}{(q^2;q^2)_{n-p}(q;q)_p(\beta;q)_{2n-2p}(\beta q^{2n-2p+1})_p}$$

$$\times(-1)^p C_{2n-2p}(\cos\theta;\beta|q). \tag{4.3}$$

If $s = r\beta$, then (2.4) yields

COROLLARY 5.

$$C_n(\cos\theta;\beta^2|q) = \sum_{p=0}^{[n/2]} \frac{(\beta;q)_p(\beta^2;q)_{n-p}(1-\beta q^{n-2p})}{(q;q)_p(q;q)_{n-p}(1-\beta)}$$

$$\times\beta^p C_{n-2p}(\cos\theta;\beta|q). \tag{4.4}$$

(Note: This is also a special case of the general formula for connection coefficients given by Rogers in [17].)

Another corollary, one which bears some resemblance to Mehler's formula for Hermite polynomials, is obtained from (2.4) when $r = \rho e^{-i\varphi}$, $s = \rho e^{i\varphi}$. We observe that

$$\sum_{t=0}^{\infty} x^t \sum_{m=0}^{t} \frac{(\beta;q)_m(\beta;q)_{t-m}e^{i\varphi(t-2m)}}{(q;q)_m(q;q)_{t-m}} = \sum_{m=0}^{\infty} \frac{(\beta;q)_m}{(q;q)_m}(xe^{-i\varphi})^m \sum_{n=0}^{\infty} \frac{(\beta;q)_n}{(q;q)_n}(xe^{i\varphi})^n,$$

$$= \frac{(x\beta e^{-i\varphi};q)_\infty(x\beta e^{i\varphi};q)_\infty}{(xe^{-i\varphi};q)_\infty(xe^{i\varphi};q)_\infty}$$

$$= \sum_{t=0}^{\infty} C_t(\cos\varphi;\beta|q)x^t,$$

and thus

$$C_t(\cos\varphi,\beta|q) = \sum_{m=0}^{t} \frac{(\beta;q)_m(\beta;q)_{t-m}}{(q;q)_m(q;q)_{t-m}} e^{i\varphi(t-2m)}. \tag{4.5}$$

With the above-mentioned substitution, the following corollary is obtained:

COROLLARY 6.

$$\frac{(\rho\beta\, e^{i\theta-i\varphi};q)_\infty(\rho\beta\, e^{-i\theta-i\varphi};q)_\infty(\rho\beta\, e^{i\theta+i\varphi};q)_\infty(\rho\beta\, e^{-i\theta+i\varphi};q)_\infty}{(\rho\, e^{i\theta-i\varphi};q)_\infty(\rho\, e^{-i\theta-i\varphi})_\infty(\rho\, e^{i\theta+i\varphi};q)_\infty(\rho\, e^{-i\theta+i\varphi};q)_\infty} \tag{4.6}$$
$$= \sum_{t=0}^{\infty} \frac{(q;q)_t}{(\beta;q)_t}\rho^t\, {}_2\varphi_1\left[\begin{matrix}\beta, \beta^2 q^t \\ \beta q^{t+1}\end{matrix};\rho^2\right] C_t(\cos\theta;\beta|q)C_t(\cos\varphi\,\beta|q).$$

Again, if $\beta = q^\lambda$, $q \to 1$, then we get the corresponding identity for ultraspherical polynomials:

COROLLARY 7.

$$(1-2\rho\cos(\theta+\varphi)+\rho^2)^{-\lambda}(1-2\rho\cos(\theta-\varphi)+\rho^2)^{-\lambda} \tag{4.7}$$
$$= \sum_{t=0}^{\infty} \frac{t!}{(\lambda)_t}\rho^t\, {}_2F_1\left[\begin{matrix}\lambda, 2\lambda+t \\ \lambda+t+1\end{matrix};\rho^2\right] C_t^\lambda(\cos\theta)C_t^\lambda(\cos\varphi).$$

REFERENCES

[1] G. ANDREWS, *The theory of partitions*, Encyclopedia of Mathematics and its Applications, vol. 2, G.-C. Rota, ed., Addison-Wesley, Reading, MA, 1977.

[2] R. ASKEY, *Orthogonal Polynomials and Special Functions*, Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[3] R. ASKEY AND G. GASPER, *Certain rational functions whose power series have positive coefficients*, Amer. Math. Monthly, 79 (1972), pp. 327–341.

[4] R. ASKEY AND M. ISMAIL, *A generalization of ultraspherical polynomials*, MRC Technical Summary Report 1851, Mathematics Research Center, Madison, WI, 1978.

[5] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, to appear.

[6] D. BRESSOUD, *A simple proof of Mehler's formula for q-Hermite polynomials*, Indiana Univ. Math. J., 29 (1980), pp. 577–580.

[7] ———, *On partitions, orthogonal polynomials, and the expansion of certain infinite products*, Proc. London Math. Soc., to appear.

[8] L. CARLITZ, *Some polynomials related to theta functions*, Ann. Math. Pura Appl., Ser. 4, 41 (1955), pp. 359–373.

[9] ———, *Some polynomials related to theta functions*, Duke Math. J., 24 (1957), pp. 521–527.

[10] J. DOUGALL, *A theorem of Sonine in Bessel functions, with two extensions to spherical harmonics*, Proc. Edinburgh Math. Soc., 37 (1919), pp. 33–37.

[11] L. FEJER, *Abschätzungen für die Legendreschen und verwandte Polynome*, Math. Zeit., 24 (1925), pp. 285–294.

[12] E. FELDHEIM, *Sur les polynômes généralisés de Legendre*, Izv. Akad. Nauk. SSSR Ser. Mat., 5 (1941), pp. 241–248, Russian Transl., ibid., pp. 248–254.

[13] I. L. LANZEWIZKY, *Uber die Orthogonalitat de Féjèr-Szegöschen Polynome*, C.R. (Dokl.) Acad. Sci. SSSR, 31 (1941), pp. 199–200.

[14] L. J. ROGERS, *On a three-fold symmetry in the elements of Heine's series*, Proc. London Math. Soc., 24 (1893), pp. 171–179.

[15] ———, *On the expansion of some infinite products*, Proc. London Math. Soc., 24 (1893), pp. 337–352.

[16] ———, *Second memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 25 (1894), pp. 318–343.

[17] ———, *Third memoir on the expansion of certain infinite products*, Proc. London Math. Soc., 26 (1895), pp. 15–32.

[18] G. SZEGÖ, *Ein Beitrag zur Theorie der Thetafunktionen*, Sitz. Preuss. Akad. Wiss. Phys. Math. K1, XIX (1926), pp. 242–252.

[19] ———, *Inequalities for the zeros of Legendre polynomials and related functions*, Trans. Amer. Math. Soc., 39 (1936), pp. 1–17.

[20] ———, *Orthogonal Polynomials*, Amer. Math. Soc. Coll. Publ., XXIII, fourth ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0004 $01.00/0

AN EXTREMAL PROBLEM INVOLVING CURRENT FLOW THROUGH DISTRIBUTED RESISTANCE*

ANDREW ACKER†

Abstract. We treat essentially the following problem in the context of electrostatics: Given a compact, convex set $Q \subset \mathbb{R}^2$ (of positive area), which is perfectly conducting and held at potential 1, how should a total amount $A > 0$ of resistance be distributed in $\mathbb{R}^2 \backslash Q$ (subject to an upper bound on resistivity) in order that the flow of current from Q to ∞ (assumed to have potential 0) be minimized?

1. Introduction and main results. Let $\mathbb{X}$ be the set of all pairs (Ω, r), where $\Omega \subset \mathbb{R}^2$ is a doubly-connected region bounded by simple closed curves Γ and Γ^* (Γ the exterior boundary) and r denotes a strictly-positive, bounded, continuously-differentiable function $r(p): \Omega \to \mathbb{R}$, such that the boundary value problem

$$\nabla \cdot \left(\frac{\nabla U(p)}{r(p)} \right) = 0 \quad \text{in } \Omega, \qquad U = 1 \quad \text{on } \Gamma^*, \qquad U = 0 \quad \text{on } \Gamma \tag{1}$$

has a unique solution. If $r(p)$ is interpreted as electrical resistivity in Ω (for $(\Omega, r) \in \mathbb{X}$ given) and the boundaries Γ and Γ^* are held at potentials 0 and 1, then the solution $U(p)$ of (1) is the electric potential in Ω, and the rate of flow of electricity across Ω from Γ^* to Γ is defined by

$$I(\Omega, r) := \int_\gamma \frac{|\nabla U(p)|}{r(p)} |dp| = \iint_\Omega \frac{|\nabla U(p)|^2}{r(p)} \, dx \, dy, \tag{2}$$

where $\gamma \subset \Omega$ is an equipotential curve of U, the second integral in (2) is a generalized Dirichlet integral and equality of the two integrals follows from Green's identity. We will solve the following problem.

Minimization problem. Given $A > 0$ and a compact, convex set $Q \subset \mathbb{R}^2$ (whose boundary ∂Q is a simple closed curve), we seek a pair $(\tilde{\Omega}, \tilde{r})$ which minimizes $I(\Omega, r)$ in the set $\mathbb{Y}$ of all $(\Omega, r) \in \mathbb{X}$ for which (a) the interior complement of Ω contains Q, (b) $r(p) \leqq 1$ throughout Ω and (c) $\|(\Omega, r)\| \leqq A$, where $\|(\Omega, r)\| := \iint_\Omega r(p) \, dx \, dy$ is the total resistance in Ω, $(\Omega, r) \in \mathbb{X}$.

Remark 1. For each pair $(\Omega, r) \in \mathbb{Y}$, we assume implicitly that the resistivity $r(p) = 0$ in $\mathbb{R}^2 \backslash \Omega$ and the potential $U(p) = 1$ (resp. 0) in the interior (exterior) complement of Ω. Thus, the above problem essentially coincides with the physical problem outlined in the abstract, with the exception that we do not consider the most general possible distributions of resistance.

The above minimization problem is completely solved (in the sense of existence, uniqueness and characterization of the solution) by the following theorem. (Note: We use $(\Omega, 1)$ to denote pairs $(\Omega, r) \in \mathbb{Y}$ for which $r(p) = 1$ throughout Ω.)

THEOREM 1.

(a) *There exists a unique pair* $(\tilde{\Omega}, 1) \in \mathbb{Y}$ *such that*

$$\tilde{\Gamma}^* = \partial Q, \ \|(\tilde{\Omega}, 1)\| = A \text{ and } |\nabla \tilde{U}(p)| = \tilde{c} \text{ on } \tilde{\Gamma} \ (\text{for some constant } \tilde{c} > 0), \tag{3}$$

(*where* $\tilde{\Gamma}$, $\tilde{\Gamma}^*$ *are the boundaries of* $\tilde{\Omega}$, *and* $\tilde{U}(p)$ *solves* (1) *relative to the pair* $(\tilde{\Omega}, 1)$).

* Received by the editors November 16, 1979.

† Mathematisches Institut I, Universität Karlsruhe (TH), 75 Karlsruhe 1, Englerstrasse 2, Postfach 6380, Federal Republic of Germany.

(b) *For any pair* $(\Omega, r) \neq (\tilde{\Omega}, 1)$ *in* $\mathbb{Y}$, *we have*

$$I(\Omega, r) > I(\tilde{\Omega}, 1). \tag{4}$$

Thus $(\tilde{\Omega}, 1)$ *is the unique solution of the minimization problem.*

Remark 2. The existence aspect of Theorem 1, namely part (a), follows easily from a result of Tepper [6] (see also [1, Theorem 1]). Thus, our main concern here is the proof of the isoperimetric inequality stated in part (b). This inequality directly generalizes [1, Theorem 2 (Case 1)], which asserts in the present context that $I(\Omega, 1) > I(\tilde{\Omega}, 1)$ for any pair $(\Omega, 1) \neq (\tilde{\Omega}, 1)$ in $\mathbb{Y}$. In fact Theorem 1(b) was conjectured in [1, Remark 5].

Remark 3. The proof of [1, Theorem 2 (Case 1)], which appears not to generalize to the present context, was essentially based on a continuous deformation (with monotone increasing capacity) of $\tilde{\Omega}$ into any other admissible region Ω. (Note that $I(\Omega, 1) =$ capacity of Ω.) By contrast, our proof here of Theorem 1(b) is based on a more direct method due to J. Hersch [5] involving the reduction of our extremal problem to a class of auxiliary one-dimensional minimization problems.

2. Proof of Theorem 1(b). Let $(\tilde{\Omega}, 1)$ be the pair satisfying (3), and define $\tilde{S} = \mathbb{R}^2 \backslash (Q \cup \tilde{\Omega})$. Then $\tilde{\Gamma}$ is a convex analytic curve (by [1, Theorem 1(d) and Lemma 5(a)]), and $|\nabla \tilde{U}(p)| > \tilde{c}$ throughout $\tilde{\Omega}$ (by [1, Lemma 5(c)]). (Note: $\tilde{U}(p)$ is simply the harmonic measure of $\tilde{\Gamma}^*$ in $\tilde{\Omega}$.)

Let $p_0 \in \tilde{\Gamma}$ be fixed, and for any $\alpha \in \mathbb{R}$, let $p_\alpha \in \tilde{\Gamma}$ be the point attained by starting out at p_0 and proceeding a distance $|\alpha|$ along $\tilde{\Gamma}$ in the positive (negative) sense for $\alpha > (<) 0$. Clearly, $p_{\alpha+L} = p_\alpha$, $\alpha \in \mathbb{R}$, where L is the length of $\tilde{\Gamma}$. Since $\tilde{\Gamma}$ is analytic and $|\nabla \tilde{U}(p)| > 0$ in $\tilde{\Omega}$, we can define the family of curves γ_α, $\alpha \in \mathbb{R}$, by setting $\gamma_\alpha = \gamma'_\alpha \cup \gamma''_\alpha$, where $\gamma'_\alpha \subset \tilde{\Omega}$ is the curve of steepest ascent of the function $\tilde{U}(q)$ joining p_α to $\tilde{\Gamma}^* = \partial Q$ and $\gamma''_\alpha \subset \tilde{S}$ is the ray emanating from $p_\alpha \in \tilde{\Gamma}$ in the exterior normal direction (see Fig. 1). Clearly,

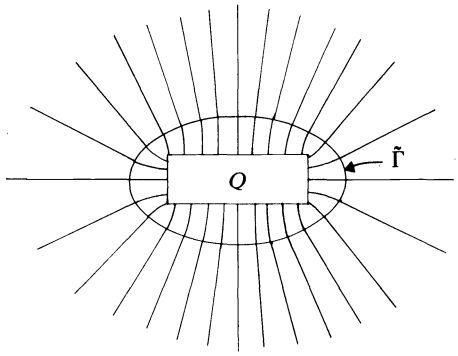


FIG. 1. *The family of curves* γ_α, $\alpha \in \mathbb{R}$.

$\gamma_{\alpha+L} = \gamma_\alpha$, $\alpha \in \mathbb{R}$, and $\cup_{0 \leqq \alpha < L} \gamma_\alpha = \mathbb{R}^2 \backslash Q$. Also, $\gamma_\alpha \cap \gamma_\beta = \varnothing$ for $0 \leqq \alpha < \beta < L$, since $\tilde{\Gamma}$ is a convex curve (i.e., $Q \cup \tilde{\Omega}$ is convex).

Using the fact that the function $\tilde{U}(q)$ can be harmonically continued across $\tilde{\Gamma}$ (see the proof of [1, Lemma 5(a)]), one easily sees that there is a unit-vector-valued function $T(p): \mathbb{R}^2 \backslash Q \to \mathbb{R}^2$, Lipschitz continuous in any closed subset of $\mathbb{R}^2 \backslash Q$, such that $T(p)$ is tangent to γ_α at each $p \in \gamma_\alpha$, $\alpha \in \mathbb{R}$. (One can set $T(p) = \nabla \tilde{U}(p)/|\nabla \tilde{U}(p)|$, $p \in \tilde{\Omega} \cup \tilde{\Gamma}$, and $T(p) = T(p_\alpha)$, $p \in \gamma''_\alpha$, $0 \leqq \alpha < L$.) Furthermore, we have

$$\lim_{\delta \to 0} \frac{\operatorname{dist}(p, \gamma_{\alpha+\delta})}{|\delta|} = m(p) \quad \text{for all } p \in \gamma_\alpha, \alpha \in \mathbb{R}, \tag{5}$$

where $m(p) = \tilde{c}/|\nabla \tilde{U}(p)|$ in $\tilde{\Omega} \cup \tilde{\Gamma}$, $m(p) = 1 + k(p_\alpha) \cdot |p - p_\alpha|$ in γ''_α, $\alpha \in \mathbb{R}$, and $k(p_\alpha) \geqq 0$

denotes the curvature of $\tilde{\Gamma}$ at p_α. Notice that $m(p)$ is a positive, continuous function in $\mathbb{R}^2\backslash Q$ and that the convergence in (5) is uniform over all p in any compact subset of $\mathbb{R}^2\backslash Q$. One can show, using these various properties of the curves γ_α, $\alpha \in \mathbb{R}$, that

$$\iint_D f(p)\,dx\,dy = \int_0^L \left(\int_{D\cap\gamma_\alpha} m(p)f(p)|dp| \right) d\alpha, \tag{6}$$

for any bounded region $D \subset \mathbb{R}^2\backslash Q$ and any function $f(p)$ which is continuous and absolutely integrable in D.

Now let a pair $(\Omega, r) \in \mathbb{Y}$ and a value $\alpha \in \mathbb{R}$ be fixed, and let $t \geqq 0$ measure arc length along γ_α from ∂Q. For each $t \geqq 0$, let $p(t) \in \gamma_\alpha$ be the point at arc-distance t from ∂Q, and define $m(t) = m(p(t))$, $a = \max\{t \geqq 0 : p(t) \in \Gamma^*\}$, $b = \min\{t \geqq 0 : p(t) \in \Gamma\}$, and $r(t) = r(p(t))$ in (a, b) (clearly, $a < b$). The functional $J(\phi) = \int_I (m(t)/r(t))\,(d\phi/dt)^2\,dt$ (where $I = [a, b]$) is easily seen to be minimized in the class Φ of all functions $\phi \in C^0[a, b] \cap C^1(a, b)$ satisfying $\phi(a) = 1$ and $\phi(b) = 0$ by the function $\phi_0(t) = (\int_{I(t)} r(t')\mu(t')\,dt' / \int_I r(t')\mu(t')\,dt')$, where $I(t) = [t, b]$, $t \in I$, $\mu(p) = 1/m(p)$ in $\mathbb{R}^2\backslash Q$, and $\mu(t) = \mu(p(t))$, $t \geqq 0$. It follows that

$$J(\phi) \geqq J(\phi_0) = \frac{1}{\int_I r(t)\mu(t)\,dt} \geqq \frac{1}{\int_{\Omega\cap\gamma_\alpha} r(p)\mu(p)\,|dp|} \tag{7}$$

for all $\phi \in \Phi$. Since the function $U(p(t)) : I \to \mathbb{R}$ (where $U(p)$ solves (1)) belongs to the class Φ, and since $|\nabla U(p(t))| \geqq |dU(p(t))/dt|$ in (a, b), we conclude using (7) that

$$\int_{\Omega\cap\gamma_\alpha} \left(\frac{m(p)}{r(p)} \right) |\nabla U(p)|^2 |dp| \geqq \frac{1}{\int_{\Omega\cap\gamma_\alpha} r(p)\mu(p)\,|dp|},$$

where $\alpha \in \mathbb{R}$ is arbitrary. Therefore, using (6), we obtain

$$\begin{aligned} I(\Omega, r) &= \iint_\Omega \left(\frac{|\nabla U(p)|^2}{r(p)} \right) dx\,dy \\ &= \int_0^L \left(\int_{\Omega\cap\gamma_\alpha} \left(\frac{m(p)}{r(p)} \right) |\nabla U(p)|^2 |dp| \right) d\alpha \\ &\geqq \int_0^L \left(\frac{1}{\int_{\Omega\cap\gamma_\alpha} r(p)\mu(p)|dp|} \right) d\alpha. \end{aligned}$$

We conclude, by applying the Schwarz inequality (in the form $L^2 \leqq \int_0^L f(\alpha)\,d\alpha \cdot \int_0^L (1/f(\alpha))\,d\alpha$, $f(\alpha)$ positive and continuous in $[0, L]$) and (6), that

$$I(\Omega, r) \geqq \frac{L^2}{\int_0^L (\int_{\Omega\cap\gamma_\alpha} r(p)\mu(p)|dp|)\,d\alpha} = \frac{L^2}{\iint_\Omega r(p)\mu^2(p)\,dx\,dy}$$

for any pair $(\Omega, r) \in \mathbb{Y}$. Now $\mu(p) = 1/m(p) = (|\nabla \tilde{U}(p)|/\tilde{c}) > 1$ throughout $\tilde{\Omega}$ by [1, Lemma 5(c)], and clearly, $0 < \mu(p) \leqq 1$ in $\tilde{S}$ (in fact $Q \cup \tilde{\Omega}$ is strictly convex and $\mu(p) < 1$ in the interior of $\tilde{S}$). Therefore, obviously,

$$\iint_\Omega r(p)\mu^2(p)\,dx\,dy < \iint_{\tilde{\Omega}} \mu^2(p)\,dx\,dy$$

for any pair $(\Omega, r) \neq (\tilde{\Omega}, 1)$ in $\mathbb{Y}$. It follows that

$$I(\Omega, r) > \frac{L^2}{\iint_{\tilde{\Omega}} \mu^2(p)\, dx\, dy} = \frac{\tilde{c}^2 L^2}{\iint_{\tilde{\Omega}} |\nabla \tilde{U}(p)|^2\, dx\, dy}$$
$$= \frac{\tilde{c}^2 L^2}{I(\tilde{\Omega}, 1)} = I(\tilde{\Omega}, 1)$$

for any pair $(\Omega, r) \neq (\tilde{\Omega}, 1)$ in $\mathbb{Y}$, which is exactly the assertion of Theorem 1(b).

3. Concluding remarks. It is not essential that Q be convex in the above argument. One can show just as before that $I(\Omega, r) \geqq I(\tilde{\Omega}, 1)$ for all $(\Omega, r) \in \mathbb{Y}$, where $(\tilde{\Omega}, 1)$ satisfies (3), provided only that $|\nabla \tilde{U}(p)| \geqq \tilde{c}$ throughout $\tilde{\Omega}$ ($\Rightarrow \tilde{\Gamma}$ is a convex curve). Moreover, $(\tilde{\Omega}, 1)$ does not minimize $I(\Omega, r)$ in $\mathbb{Y}$ unless this condition holds.

Clearly, the above proof can be extended to show that $(\tilde{\Omega}, 1)$ minimizes current flow from Q (at least in the sense of $\leqq$) within a more general class of distributions of resistance (subject to the same constraints).

REFERENCES

[1] A. ACKER, *A free boundary optimization problem*, this Journal, 9 (1978), pp. 1179–1191.
[2] ———, *A free boundary optimization problem. II*, this Journal, 11 (1980), pp. 201–209.
[3] ———, *An extremal problem involving current flow through distributed resistance*, Abstracts Amer. Math. Soc., 1 (1980), p. 340.
[4] ———, *An inequality involving the flow of current through distributed resistance*, Abstracts Amer. Math. Soc., 1 (1980), p. 225.
[5] J. HERSCH, *Sur un problème de potentiel avec conditions aux limites mixtes: une conjecture de A. Acker et sa démonstration à l'aide de problèmes auxiliares à une dimension*, Z. Angew. Math. Phys., 30 (1979), pp. 716–721.
[6] D. E. TEPPER, *Free boundary problem*, this Journal, 6 (1975), pp. 503–505.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0005 $01.00/0

EXISTENCE-UNIQUENESS FOR FOCAL-POINT BOUNDARY VALUE PROBLEMS*

ALLAN C. PETERSON†

Abstract. Our main concern is to prove uniqueness-existence theorems for the focal boundary value problem $y^{(n)}=f(x, y, \cdots, y^{(n-1)})$, $y^{(i)}(a)=A_i$, $y^{(j)}(b)=A_j$, $0\leqq i\leqq k-1$, $k\leqq j\leqq n-1$. The method of proof for existence is the shooting method. Similar results, but for the (p, q)-boundary value problem $y^{(n)}=f(x, y, \cdots, y^{(n-1)})$, $y^{(i)}(a)=A_i$, $y^{(j)}(b)=A_j$, $0\leqq i\leqq p-1$, $0\leqq j\leqq q-1$, $p+q=n$, were considered by the author in [J. Math. Anal. Appl., 55 (1976), pp. 773–784]. The last result in this paper extends the main result in the above-mentioned article.

At the outset of this paper we will be concerned with the general differential equation

$$y^{(n)}=f(x, y, \cdots, y^{(n-1)}). \tag{1}$$

Let I be a subinterval of the real numbers R. At various times we will assume f satisfies one or more of the following assumptions:

(A) The function f is continuous on $I\times R^n$.

(B) Solutions of initial value problems (IVP's) for (1) are unique and exist on I.

(C) If there is a nondegenerate subinterval $[\alpha, \beta]\subset I$ and a sequence of solutions $\{y_n\}$ which is uniformly bounded on $[\alpha, \beta]$, then there is a subsequence $\{y_{n_k}\}$ such that $\{y_{n_k}^{(j)}\}$ converges uniformly on each compact subinterval of I, $j=0, \cdots, n-1$.

(D) The functions $f_i\equiv \partial f/\partial y^{(i)}$, $i=0, \cdots, n-1$, are continuous on $I\times R^n$.

See [10] and the references given there concerning the compactness condition (C).

Our main concern (except for Theorem 10) is to prove uniqueness–existence theorems for the focal boundary value problem (1),

$$y^{(i)}(a)=A_i, \qquad 0\leqq i\leqq k-1, \tag{2}$$

$$y^{(j)}(b)=A_j, \qquad k\leqq j\leqq n-1, \tag{3}$$

where $a, b\in I$ with $a<b$. Our method of proof for existence will be the shooting method which has important implications in numerically approximating solutions of (1), (2), (3).

DEFINITION. Let $1\leqq k\leqq n-1$. We say that (1) is *k-disfocal* on $J\subset I$ provided the boundary value problem (BVP) (1), (2), (3), where $a<b$ are arbitrary in J and $A_i\in R$, $0\leqq i\leqq n-1$, has at most one solution.

Two questions of interest are when is (1) k-disfocal, and if (1) is k-disfocal does this imply that all BVP's (1), (2), (3) have solutions? First we give a local existence theorem for the BVP (1), (2), (3).

THEOREM 1. *Assume* (A) *and* (B) *hold and that* (1) *is k-disfocal on* $(\alpha, \beta)\subset I$; *then given* $\alpha<s<t<\beta$ *and* $y(x)$ *a solution of* (1) *there is an* $\varepsilon>0$ *such that if* $|s_1-s|<\varepsilon$, $|t_1-t|<\varepsilon$, t_1, $s_1\in(\alpha, \beta)$, $|y^{(i)}-y^{(i)}(s)|<\varepsilon$, $i=0, \cdots, k-1$, $|y^{(j)}-y^{(j)}(t)|<\varepsilon$, $j=k, \cdots, n-1$. *Then the* BVP (1)

$$y^{(i)}(s_1)=y^{(i)}, \qquad i=0, \cdots, k-1,$$

$$y^{(j)}(t_1)=y^{(j)}, \qquad j=k, \cdots, n-1,$$

* Received by the editors July 30, 1979, and in revised form August 4, 1980.

† Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588.

has a unique solution $u_\varepsilon(x)$. *Furthermore* $\lim_{\varepsilon\to 0} u_\varepsilon^{(i)}(x) = y^{(i)}(x)$ *uniformly on compact subsets of* I, $0 \leqq i \leqq n-1$.

Proof. The proof is a standard application (see [14] and references [1]–[3], [10] there) of Brouwer's invariance-of-domain theorem (see [14, Lemma 2.1,] and the reference there to [8]).

THEOREM 2. *Assume* (A), (B) *and* (C) *hold. If* (1) *is* k*-disfocal on* I, *then for any solution* $u(x)$ *of* (1) *and* $c < d$ *in* I, *the set*

$$A = \{y^{(k-1)}(c): y \text{ is a solution of (1) such that } y^{(i)}(c) = u^{(i)}(c),$$
$$i = 0, \cdots, k-2,\ y^{(j)}(d) = u^{(j)}(d),\ j = k, \cdots, n-1\}$$

is an open interval.

Proof. By Theorem 1, A is an open subset of the reals R. Assume A is not an interval; then either there is a $t_1 > u^{(k-1)}(c)$ such that $t_1 \in A$ but $[u^{(k-1)}(c), t_1] \not\subset A$, or there is a $t_2 < u^{(k-1)}(c)$ such that $t_2 \in A$ but $[t_2, u^{(k-1)}(c)] \not\subset A$. We will consider only the first case here. Set

$$\delta_0 = \sup\{s \geqq u^{(k-1)}(c): [u^{(k-1)}(c), s] \subset A\}.$$

So $\delta_0 \leqq t_1$, and by Theorem 1, $\delta_0 > u^{(k-1)}(c)$ and $\delta_0 \notin A$. Pick a sequence $\{s_n\}$ such that $u^{(k-1)}(c) < s_1 < \cdots < s_n \cdots$ and $\lim_{n\to\infty} s_n = \delta_0$. Let $y_n(x)$ be the solution of (1) satisfying $y_n^{(i)}(c) = u^{(i)}(c)$, $i = 0, \cdots, k-2$, $y_n^{(k-1)}(c) = s_n$, and $y_n^{(j)}(d) = u^{(j)}(d)$, $j = k, \cdots, n-1$. If there is an $\varepsilon > 0$ such that $\{y_n^{(k-1)}(x)\}$ is uniformly bounded on $[c, c+\varepsilon]$ then it follows easily that $\{y_n(x)\}$ is uniformly bounded on $[c, c+\varepsilon]$. But then by the compactness assumption (C) there is a subsequence $\{y_{n_l}(x)\}$ and a solution $y(x)$ of (1) such that

$$\lim_{l\to\infty} y_{n_l}^{(i)}(x) = y^{(i)}(x)$$

uniformly on compact subsets of I, $i = 0, \cdots, n-1$. This implies $\delta_0 \in A$, which is a contradiction.

Now since $t_1 \in A$ there is a solution $v(x)$ such that

$$v^{(i)}(c) = u^{(i)}(c), \qquad i = 0, \cdots, k-2,$$
$$v^{(k-1)}(c) = t_1,$$
$$v^{(j)}(d) = u^{(j)}(d), \qquad j = k, \cdots, n-1.$$

Since $\{y_n^{(k-1)}(x)\}$ or any subsequence is not uniformly bounded on $[c, c+\varepsilon]$ for any $\varepsilon > 0$, it follows that for all sufficiently large n, $y_n^{(k-1)}(x)$ crosses either $v^{(k-1)}(x)$ or $u^{(k-1)}(x)$. By relabeling sequences if necessary we can assume that $y_n^{(k-1)}(x)$ crosses $v^{(k-1)}(x)$ or $u^{(k-1)}(x)$ for all $n \geqq 1$. For each n, pick the first point x_n such that $y_n^{(k-1)}(x_n) \in \{u^{(k-1)}(x_n), v^{(k-1)}(x_n)\}$. For infinitely many values of n, either $y_n^{(k-1)}(x_n) = u^{(k-1)}(x_n)$ or $y_n^{(k-1)}(x_n) = v^{(k-1)}(x_n)$. Without loss of generality we will assume $y_n^{(k-1)}(x_n) = v^{(k-1)}(x_n)$ for $n \geqq 1$. Also we may as well assume $x_1 > x_2 > \cdots$ with $\lim_{n\to\infty} x_n = c$. Since

$$u^{(k-1)}(x) < y_n^{(k-1)}(x) < v^{(k-1)}(x) \quad \text{on } [c, x_n],$$

we get that

$$u^{(i)}(x) < y_n^{(i)}(x) < v^{(i)}(x)$$

on (c, x_n) for $i = 0, \cdots, k-1$. It follows from this that

$$\lim_{n\to\infty} y_n^{(i)}(x_n) = v^{(i)}(c), \qquad i = 0, \cdots, k-1.$$

Since

$$y_n^{(j)}(d) = v^{(j)}(d), \qquad j = k, \cdots, n-1,$$

we get from Theorem 1 that

$$\lim_{n\to\infty} y_n^{(i)}(x) = v^{(i)}(x)$$

uniformly on compact subsets for $i = 0, \cdots, n-1$, which again contradicts $\delta_0 \notin A$ and completes the proof.

Assume (D) holds and $y(x)$ is a solution of (1); then the linear differential equation

$$z^{(n)} = \sum_{i=0}^{n-1} f_i(x, y(x), \cdots, y^{(n-1)}(x))z^{(i)} \tag{4}$$

is called (see [1]) the variational equation of (1) along $y(x)$.

THEOREM 3. *Assume* (A), (B) *and* (D) *hold and* $c < d$. *Let* $1 \leqq k \leqq n-1$, *and assume* (1) *and the variational equation* (4) *along all solutions* $y(x)$ *of* (1) *are k-disfocal on I. If* $u(x)$ *is a solution of* (1), *then there is an open interval* (γ, δ), $\gamma < 0 < \delta$ *such that the focal* BVP (1)

$$\begin{aligned} &y^{(i)}(c) = u^{(i)}(c), \qquad i = 0, \cdots, k-2, \\ &y^{(k-1)}(c) = u^{(k-1)}(c) + s, \\ &y^{(j)}(d) = u^{(j)}(d), \qquad j = k, \cdots, n-1 \end{aligned}$$

has a unique solution $y(x, s)$ *for* $s \in (\gamma, \delta)$. *Furthermore* $\partial y/\partial s$ *exists for* $s \in (\gamma, \delta)$, *and* $z(x) = \partial y(x, s)/\partial s$ *is the solution of the focal* BVP (4) *with* $y(x) = y(x, s)$,

$$\begin{aligned} &z^{(i)}(c) = 0, \qquad i = 0, \cdots, k-2, \\ &z^{(k-1)}(c) = 1, \\ &z^{(j)}(d) = 0, \qquad j = k, \cdots, n-1. \end{aligned}$$

Proof. The first conclusion follows from Theorem 2. To show the last statement, fix $s \in (\gamma, \delta)$ and for $h \neq 0$ sufficiently small, set

$$z_h(x) = \frac{y(x, s+h) - y(x, s)}{h}.$$

Also set

$$A_i = y^{(i)}(c, s)$$

and

$$\delta_i(h) = y^{(i)}(c, s+h) - A_i, \qquad i = k, \cdots, n-1.$$

Note that by Theorem 1,

$$\lim_{h\to 0} \delta_i(h) = 0.$$

Let $y(x; u_1, \cdots, u_n)$ denote the solution of the IVP (1), $y^{(i-1)}(c) = u_i$, $i = 1, \cdots, n$.

Consider

$$z_h(x)=\frac{1}{h}[y(x,s+h)-y(x,h)]$$

$$=\frac{1}{h}[y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s+h,A_k+\delta_k,\cdots,A_{n-1}+\delta_{n-1})$$

$$-y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s,A_k,\cdots,A_{n-1})]$$

$$=\frac{1}{h}\{[y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s+h,A_k+\delta_k,\cdots,A_{n-1}+\delta_{n-1})$$

$$-y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s,A_k+\delta_k,\cdots,A_{n-1}+\delta_{n-1})]$$

$$+\cdots$$

$$+[y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s,A_k,\cdots,A_{n-2},A_{n-1}+\delta_{n-1})$$

$$-y(x;u(c),\cdots,u^{(k-2)}(c),u^{(k-1)}(c)+s,A_k,\cdots,A_{n-1})]\}.$$

By [1, Theorem V 3.1] we have, for $h\neq 0$,

$$z_h(x)=\frac{1}{h}\{hz_{k-1}(x;y(x;u(c),\cdots,u^{(k-1)}(c)+s+\bar{h},A_k+\delta_k,\cdots,A_{n-1}+\delta_{n-1}))$$

$$+\delta_k z_k(x;y(x;u(c),\cdots,u^{(k-1)}(c)+s,A_k+\bar{\delta}_k,A_{k+1}$$

$$+\delta_{k+1},\cdots,A_{n-1}+\delta_{n-1}))$$

$$+\cdots$$

$$+\delta_{n-1}z_{n-1}(x;y(x;u(c),\cdots,u^{(k-1)}(c)+s,A_{k-1},\cdots,A_{n-2},A_{n-1}+\bar{\delta}_{n-1})\}$$

where $z_j(x;y(x))$ denotes the solution of the IVP (4)

$$\begin{aligned} z_j^{(i)}(c;y(x))&=0,\\ z_j^{(j)}(c;y(x))&=1, \end{aligned}\qquad i=0,\cdots,n-1 \quad \text{but } i\neq j$$

and h is between 0 and h, $\bar{\delta}_j$ is between 0 and δ_j, $j=k,\cdots,n-1$. We want to show that $\lim_{h\to 0} z_h(x)$ exists for each x.

Since $y^{(j)}(d,s+h)=u^{(j)}(d)=y^{(j)}(d,s)$, $j=k,\cdots,n-1$, we get

$$0=z_{k-1}^{(j)}(d;y(x;u(c),\cdots,u^{(k-1)}(c)+s+\bar{h},A_k+\delta_k,\cdots,A_{n-1}+\delta_{n-1}))$$

$$+\frac{\delta_k}{h}z_k^{(j)}(d;y(x;u(c),\cdots,u^{(k-1)}(c)+s,A_k+\bar{\delta}_k,A_{k+1}$$

$$+\delta_{k+1},\cdots,A_{n-1}+\delta_{n-1}))$$

$$+\cdots \tag{6}$$

$$+\frac{\delta_{n-1}}{h}z_{n-1}^{(j)}(d;y(x;u(c),\cdots,u^{(k-1)}(c)+s,A_k,\cdots,A_{n-1}+\bar{\delta}_{n-1})),$$

$$j=k,\cdots,n-1.$$

By the k-disfocality of (4) along $y(x,s)$ we get

$$D\equiv\begin{vmatrix} z_k^{(k)}(d;y(x,s)) & \cdots & z_{n-1}^{(k)}(d;y(x,s))\\ \cdots & & \cdots\\ z_k^{(n-1)}(d;y(x,s)) & \cdots & z_{n-1}^{(n-1)}(d;y(x,s)) \end{vmatrix}\neq 0.$$

It follows that for $h \neq 0$, sufficiently small, we can uniquely solve the system (6) for δ_j/h, $j = k, \cdots, n-1$. For example (omitting the arguments for various functions),

$$\frac{\delta_k(h)}{h} = \frac{1}{D(h)} \begin{vmatrix} -z_{k-1}^{(k)} & z_{k+1}^{(k)} & \cdots & z_{n-1}^{(k)} \\ \cdots & & \cdots & \\ -z_{k-1}^{(n-1)} & z_{k+1}^{(n-1)} & \cdots & z_{n-1}^{(n-1)} \end{vmatrix}.$$

where $D(h)$ is suitably defined. It follows that

$$\lim_{h\to 0} \frac{\delta_k(h)}{h} = \frac{1}{D} \begin{vmatrix} -z_{k-1}^{(k)}(d; y(x,s)) & \cdots & z_{n-1}^{(k)}(d; y(x,s)) \\ \cdots & & \cdots \\ -z_{k-1}^{(n-1)}(d; y(x,s)) & \cdots & z_{n-1}^{(n-1)}(d; y(x,s)) \end{vmatrix}$$

Similarly we can evaluate $\lim_{h\to 0} \delta_j(h)/h$, $j = k+1, \cdots, n$.

It follows from (5) that

$$\lim_{h\to 0} z_h(x) = z_{k-1}(x; y(x,s)) + C_k z_k(x; y(x,s)) + \cdots + C_{n-1} z_{n-1}(x; y(x,s)),$$

where $C_j = \lim_{h\to 0} (\delta_j(h)/h)$, $k \leqq j \leqq n-1$.

It is now an easy matter to check that $\partial y(x,s)/\partial s = \lim_{h\to 0} z_h(x)$ is the solution of (4) along $y(x,s)$ satisfying the boundary conditions

$$z^{(i)}(c) = 0, \qquad i = 0, \cdots, k-2,$$
$$z^{(k-1)}(c) = 1,$$
$$z^{(j)}(d) = 0, \qquad j = k, \cdots, n-1.$$

We now give conditions under which (1) is k-disfocal.

THEOREM 4. *Assume* $1 \leqq p \leqq n-1$ *and that* (A)–(D) *hold. If the variational equations* (4) *are k-disfocal along all solutions of* (1) *on I for* $k = p, \cdots, n-1$, *then* (1) *is also k-disfocal on I for* $k = p, \cdots, n-1$.

Proof. We prove this theorem using finite mathematical induction. Assume $p = n-1$ and that (1) is not $(n-1)$-disfocal on I. Then there are distinct solutions y_1, y_2 and points $c < d$ in I such that

$$y_1^{(i)}(c) = y_2^{(i)}(c), \qquad i = 0, \cdots, n-2,$$
$$y_1^{(n-1)}(d) = y_2^{(n-1)}(d).$$

Let $y(x,s)$ be the solution of (1) such that $y^{(i)}(c) = y_1^{(i)}(c)$, $i = 0, \cdots, n-2$, $y^{(n-1)}(c) = s$. Then there are numbers $s_1 \neq s_2$ such that $y_1(x) = y(x, s_1)$, $y_2(x) = y(x, s_2)$. By [1, Theorem V, 3.1]

$$0 = y_1^{(n-1)}(d) - y_2^{(n-1)}(d)$$
$$= y^{(n-1)}(d, s_1) - y^{(n-1)}(d, s_2)$$
$$= (s_1 - s_2) \frac{\partial}{\partial s} y^{(n-1)}(d, \bar{s}),$$

where $\bar{s}$ is between s_1, and s_2. Let $z(x) = \partial y(x, \bar{s})/\partial s$; then $z(x)$ is the solution of (4) along $y(x) = y(x, \bar{s})$ satisfying $z^{(i)}(c) = 0$, $i = 0, \cdots, n-2$, $z^{(n-1)}(c) = 1$. By the last displayed equation we get $z^{(n-1)}(d) = 0$, which contradicts the fact that (4) is $(n-1)$-disfocal along $y(x, \bar{s})$. Hence this theorem is true for $p = n-1$.

Assume $1 \leqq p < n-1$, and that this theorem is true for $p+1, \cdots, n-1$. We want to show that (1) is p-disfocal on I. Assume not; then there are distinct solutions $y_1(x)$, $y_2(x)$ of (1) and points $c < d$ in I such that

$$y_1^{(i)}(c) = y_2^{(i)}(c), \qquad i = 0, \cdots, p-1,$$
$$y_1^{(j)}(d) = y_2^{(j)}(d), \qquad j = p, \cdots, n-1.$$

By the induction assumption (1) is $(p+1)$-disfocal on I, and so we can apply Theorem 2 to get

$$S \equiv \{s \in R: \text{the } (p+1)\text{-focal BVP (1), } y^{(i)}(c) = y_1^{(i)}(c),$$
$$i = 0, \cdots, p-1, \, y^{(p)}(c) = s, \, y^{(j)}(d) = y_1^{(i)}(d), \, j = p+1, \cdots,$$
$$n-1, \text{ has a solution } y(x, s)\}$$

is an open interval. Pick $s_1 \neq s_2$ such that $y_1(x) = y(x, s_1)$, $y_2(x) = y(x, s_2)$. Using the connectedness of S and Theorem 3 we have

$$\begin{aligned} 0 &= y_1^{(p)}(d) - y_2^{(p)}(d) \\ &= y^{(p)}(d, s_1) - y^{(p)}(d, s_2) \\ &= (s_1 - s_2)\frac{\partial}{\partial s} y^{(p)}(d, \bar{s}) \\ &= (s_1 - s_2) z^{(p)}(d), \end{aligned}$$

where $\bar{s}$ is between s_1 and s_2 and $z(x) = \partial y(x, \bar{s})/\partial s$ is the solution of (4) along $y(x, \bar{s})$ satisfying $z^{(i)}(c) = 0$, $i = 0, \cdots, p-1$, $z^{(p)}(c) = 1$, $z^{(j)}(d) = 0$, $j = p+1, \cdots, n-1$. But above we saw that $z^{(p)}(d) = 0$, which contradicts the fact that (4) is p-disfocal along $y(x, \bar{s})$ on I.

We now proceed to state and prove a theorem which is a completion of [11, Theorem 6]. This result will be concerned with the differential equation

$$\rho_{n+1}(x)\frac{d}{dx}\rho_n(x)\frac{d}{dx}\cdots\frac{d}{dx}\rho_1(x)y = \lambda p(x)y, \tag{7}$$

where $\lambda = \pm 1$ and p, $\rho_i(x)$, $1 \leqq i \leqq n-1$ are positive continuous functions on $[a, b]$. Define quasi-derivatives D_i, $0 \leqq i \leqq n$, by

$$D_0 y = \rho_1(x)y,$$
$$D_i y = \rho_{i+1}(x)(D_{i-1}y)', \qquad i = 1, \cdots, n.$$

We say that a solution $y(x)$ of (7) has a zero of order k at x_0 provided $D_i y(x_0) = 0$, $i = 0, \cdots, k-1$. Like Nehari [3] we say that (7) is disfocal on $[a, b]$ provided there is no nontrivial solution $y(x)$ of (7) such that there are points x_i, $0 \leqq i \leqq n-1$, in $[a, b]$ with $D_i y(x_i) = 0$, $i = 0, \cdots, n-1$. As expected, we say that (7) is k-disfocal on I, $1 \leqq k \leqq n-1$, means that there is no nontrivial solution $y(x)$ of (7) such that there are points $a \leqq c < d \leqq b$ such that $D_i y(c) = 0$, $0 \leqq i \leqq k-1$, $D_j y(d) = 0$, $k \leqq j \leqq n-1$.

If (7) is k-disfocal on $[a, b]$, then we will let $G_k(x, s)$ denote the Green's function for the k-focal BVP

$$D_n y - \lambda p(x)y = h(x), \qquad h \in C[a, b],$$
$$D_i y(a) = 0, \qquad i = 0, \cdots, k-1,$$
$$D_j y(b) = 0, \qquad j = k, \cdots, n-1,$$

(see [11, Lemma 3], where we replace regular derivatives d^i/dx^i with corresponding quasi-derivatives D_i).

In the following theorem we use the notation

$$D_0G_k(x, s)=\rho_1(x)G_k(x, s),$$

$$D_iG_k(x, s)=\rho_{i+1}(x)\frac{\partial}{\partial x}D_{i-1}G_k(x, s), \qquad i=1, \cdots, n.$$

THEOREM 5. *Assume* (7) *is disfocal on* $[a, b]$. *Then*

$$(-1)^{n-j}D_iG_k(x, s)>0 \tag{8}$$

on $(a, b)\times(a, b)$, *where* $j=k$ *for* $0\leqq i\leqq k-1$ *and* $j=i$ *for* $k\leqq i\leqq n-1$, *is true except when* $\lambda=1$, $i\geqq k$ *and* n *and* k *have opposite parity, or when* $\lambda=-1$, $i\geqq k$ *and* n *and* k *have the same parity. In the exceptional cases* $D_iG_k(x, s)$ *changes sign in* $(a, b)\times(a, b)$.

Proof. From [11, Theorem 6] we know that

$$(-1)^{n-k}D_iG_k(x, s)>0 \tag{9}$$

on $(a, b)\times(a, b)$ for $0\leqq i\leqq k-1$. Now let $u_j(x, \tau)$, $0\leqq j\leqq n-1$, $\tau\in[a, b]$ be the solution of (7) satisfying

$$D_iu_j(\tau, \tau)=\delta_{ij}, \qquad i=0, \cdots, n-1,$$

(δ_{ij} is the Kronecker delta). Then as in Lemma 3 and [11, Theorem 4] with derivatives replaced by quasi-derivatives, we get that

$$G_k(x, s)=\begin{cases}\dfrac{1}{D}u(x), & a\leqq x\leqq s\leqq b,\\[2ex] \dfrac{1}{D}v(x), & a\leqq s\leqq x\leqq b,\end{cases}$$

where

$$u(x)=\begin{vmatrix} 0 & u_k(x, a) & \cdots & u_{n-1}(x, a)\\ D_ku_{n-1}(b, s) & D_ku_k(b, a) & \cdots & D_ku_{n-1}(b, a)\\ \cdots & & & \cdots\\ D_{n-1}u_{n-1}(b, s) & D_{n-1}u_k(b, a) & \cdots & D_{n-1}u_{n-1}(b, a)\end{vmatrix},$$

$v(x)$ is this determinant with the upper left entry 0 replaced by $u_{n-1}(x, s)$, and D is given by

$$D=\begin{vmatrix} D_ku_k(b, a) & \cdots & D_ku_{n-1}(b, a)\\ \cdots & & \cdots\\ D_{n-1}u_k(b, a) & \cdots & D_{n-1}u_{n-1}(b, a)\end{vmatrix}.$$

Let s be a fixed but arbitrary point in (a, b). If $\lambda=1$ and n and k have the same parity, or if $\lambda=-1$ and n and k have the opposite parity, we will show that $\operatorname{sgn} D_ju(x)=(-1)^{n-j}$, $a<x\leqq s$ and $\operatorname{sgn} D_jv(x)=(-1)^{n-j}$, $s\leqq x<b$ for $j=k, \cdots, n-1$. If $\lambda=1$ and n and k have the same parity, then first, by (9), $u(x)>0$ on $(a, s]$ and $v(x)>0$ on $[s, b)$ and second, by (7), we get that $D_nu(x)>0$ on $(a, s]$ and $D_nv(x)>0$ on $[s, b)$. Similarly if $\lambda=-1$ and n and k have the opposite parity we can also argue that $D_nu(x)>0$ on $(a, s]$ and $D_nv(x)>0$ on $[s, b)$.

Now since $D_{n-1}v(b)=0$, $D_{n-1}v(x)<0$ on $[s, b)$. But $D_nu(x)>0$ on $(a, s]$ and $D_{n-1}u(s)=D_{n-1}v(s)-D<0$, so $D_{n-1}u(x)<0$ on $(a, s]$. Since $D_{n-1}v(x)<0$ on $[s, b)$

and $D_{n-2}v(b)=0$, $D_{n-2}v(x)>0$ on $[s, b)$. But $D_{n-1}u(x)<0$ on $(a, s]$ and $D_{n-2}u(s)=D_{n-2}v(s)>0$ implies $D_{n-2}u(x)>0$ on $(a, s]$. Proceeding in this fashion we obtain the desired results.

We now prove the last statement in this theorem. Assume that $\lambda=1$ and n and k have the opposite parity or $\lambda=-1$ and n and k have the same parity. By (9), we have that $u(x)<0$ on $(a, s]$, $v(x)<0$ on $[s, b)$ in the first case, while $u(x)>0$ on $(a, s]$, $v(x)>0$ on $[s, b)$ in the second case. By use of (7) we obtain in both cases that $D_nu(x)<0$ on $(a, s]$ and $D_nv(x)<0$ on $[s, b)$. But $D_nv(x)<0$ on $[s, b)$ and $D_{n-1}v(b)=0$ implies $D_{n-1}v(x)>0$ on $[s, b)$. By use of finite mathematical induction we easily get that

$$(-1)^jD_{n-j-1}v(x)>0 \quad \text{on } [s, b), \qquad j=0, \cdots, n-k-1.$$

Note that

$$D_ku(a)=-\begin{vmatrix} D_ku_{n-1}(b, s) & D_ku_{k+1}(b, a) & \cdots & D_ku_{n-1}(b, a) \\ \cdots & & & \cdots \\ D_{n-1}u_{n-1}(b, s) & D_{n-1}u_{k+1}(b, a) & \cdots & D_{n-1}u_{n-1}(b, a) \end{vmatrix}. \tag{10}$$

By [4, formula (1)] we get that

$$D_ku(a)=(-1)^{n-k}\begin{vmatrix} D_0^+z_{n-k-1}(s, b) & D_{n-k-2}^+z_{n-k-1}(a, b) & \cdots & D_{n-k-1}^+z_{n-k-1}(a, b) \\ \cdots & & & \cdots \\ D_0^+z_0(s, b) & D_{n-k-2}^+z_0(a, b) & \cdots & D_0^+z_0(a, b) \end{vmatrix},$$

with $D_i^+z_j(a, b)=\delta_{ij}$, $i=0, \cdots, n-1$.

This last determinant is zero if and only if there is a nontrivial solution $z(x)$ of the adjoint equation of (7) with

$$\begin{aligned} &D_iz(a)=0, \qquad 0\leqq i\leqq n-k-2, \\ &z(s)=0, \\ &D_jz(b)=0, \qquad n-k\leqq j\leqq n-1. \end{aligned}$$

By the use of Rolle's theorem this leads to a contradiction of the disfocality (see [3, Theorem 6]) of the adjoint equation of (7). Hence $D_ku(a)\neq 0$. This last inequality is true for $s\in(a, b)$. By setting $s=b$ in (10) we have that

$$D_ku(a)|_{s=b}=(-1)^{n-k}\begin{vmatrix} D_ku_{k+1}(b, a) & \cdots & D_ku_{n-1}(b, a) \\ \cdots & & \cdots \\ D_{n-2}u_{k+1}(b, a) & \cdots & D_{n-2}u_{n-1}(b, a) \end{vmatrix}.$$

Since this last determinant is positive (see [12, Lemma 3]),

$$(-1)^{n-k}D_ku(a)>0.$$

By very similar arguments we can show that

$$(-1)^{n-i}D_iu(a)>0,$$

for $k\leqq i\leqq n-1$. But for $k\leqq i\leqq n-2$,

$$(-1)^{n-i-1}D_iu(s)=(-1)^{n-i-1}D_iv(s)>0.$$

Hence $D_iu(x)$, $k\leqq i\leqq n-2$, changes sign in (a, s). Earlier we saw that $D_nu(x)<0$ on $(a, s]$. Since $D_{n-1}u(a)<0$, we have that $D_{n-1}u(x)<0$ on $[a, s]$. Since we have already seen that $D_{n-1}v(x)>0$ on $[s, b)$ we get that $D_{n-1}G_k(x, s)$ changes sign in $(a, b)\times(a, b)$.

Using an argument as in the proof of [12, Theorem 6] we can prove the next corollary.

COROLLARY 6. *If we assume* $p(x) \geqq 0$ *on* $[a, b]$ *in Theorem* 5, *then the inequalities* (8) *are true with* $>$ *replaced by* $\geqq$. *Also the last statement of that theorem is true.*

We would like now to state a comparison theorem for the differential equations

$$D_n y = \lambda q_1(x) y, \tag{11}$$

$$D_n y = \lambda q_2(x) y, \tag{12}$$

where we assume that

$$0 \leqq q_1(x) \leqq q_2(x) \quad \text{on } [a, b].$$

Indeed, it is a standard argument to use Corollary 6 to get the following comparison theorem.

COROLLARY 7. *Assume* (12) *is disfocal on* $[a, b]$. *Let* $y_1(x)$, $y_2(x)$ *be solutions of* (11) *and* (12) *respectively such that* $y_1(x) \geqq 0$ *on* $[a, b]$ *and* $D_i y_1(a) = D_i y_2(a)$, $0 \leqq i \leqq k-1$ *and* $D_j y_1(b) = D_j y_2(b)$, $k \leqq j \leqq n-1$. *Then*

$$(-1)^{n-j} \lambda D_i y_2(x) \geqq (-1)^{n-j} \lambda D_i y_1(x)$$

on $[a, b]$, *where* $j = k$ *for* $0 \leqq i \leqq k-1$ *and* $j = i$ *for* $k \leqq i \leqq n-1$, *where for* $k \leqq i \leqq n-1$ *we assume* $\lambda = 1$ *and* n *and* k *have the same parity or* $\lambda = -1$ *and* n *and* k *have the opposite parity.*

Assume (7) is disfocal on $[a, b]$ and let $y_k(x; p(x))$, $1 \leqq k \leqq n-1$, denote the solution of (7) satisfying

$$D_i y_k(a; p(x)) = 0, \qquad 0 \leqq i \leqq k-1,$$

$$D_k y_k(a; p(x)) = 1,$$

$$D_j y_k(b; p(x)) = 0, \qquad k+1 \leqq j \leqq n-1,$$

(In this definition we assume $p(x) \geqq 0$ on $[a, b]$).

An application of Corollary 7 gives us the following result.

COROLLARY 8. *Assume* (12) *is disfocal on* $[a, b]$. *Then*

$$(-1)^{n-j} \lambda D_i y_k(x; q_2(x)) \geqq (-1)^{n-j} \lambda D_i y_k(x; q_1(x))$$

on $[a, b]$, *where* $j = k+1$ *for* $0 \leqq i \leqq k$ *and* $j = i$ *for* $k+1 \leqq i \leqq n-1$, *where for* $k+1 \leqq i \leqq n-1$ *we assume* $\lambda = 1$ *and* n *and* $k+1$ *have the same parity or* $\lambda = -1$ *and* n *and* $k+1$ *have the opposite parity.*

Proof. Since (12) is disfocal, $y_k(x; q_2(x))$ is well defined. By [3, Theorem 6.2], (11) is also disfocal. Therefore $y_k(x; q_1(x))$ is well defined and

$$y_k(x; q_1(x)) > 0$$

on $(a, b]$. Hence by Corollary 7 with k replaced by $k+1$ we get the desired result.

Now consider the nonlinear differential equation

$$D_n y = \lambda f(x, y), \tag{13}$$

where $\lambda = \pm 1$ and f satisfies (A)–(D). We now prove the following existence–uniqueness theorem.

THEOREM 9. *Assume* $1 \leqq k \leqq n-1$, $0 \leqq f_y(x, y) \leqq p(x)$ *on* $[a, b] \times R$ *and* (7) *is disfocal on* $[a, b]$. *Then the* BVP (13), $D_i y(a) = A_i$, $D_j y(b) = B_j$, $0 \leqq i \leqq k-1$, $k \leqq j \leqq n-1$, *has a unique solution.*

Proof. Let $y(x)$ be a solution of (13). Then the variational equation along $y(x)$ is

$$D_n z = \lambda f_y(x, y(x)) z. \tag{14}$$

Since $0 \leqq f_y(x, y(x)) \leqq p(x)$ on $[a, b]$ we have by [3, Theorem 6.2] that (14) is disfocal on $[a, b]$. Hence by Theorem 4 we get the uniqueness part of this theorem. It remains to prove the existence part of this theorem. First we prove the existence for the $k = n-1$ case. To this end let $y(x, s)$ be the solution of the IVP (13), $D_i y(a) = A_i$, $0 \leqq i \leqq n-2$, $D_{n-1} y(a) = s$. By [1, Theorem V 3.1],

$$\begin{aligned} D_{n-1} y(b, s) - D_{n-1} y(b, s_0) &= (s - s_0) \frac{\partial}{\partial s} D_{n-1} y(b, \bar{s}) \\ &= (s - s_0) D_{n-1} z_{n-1}(b, \bar{s}), \end{aligned}$$

where $\bar{s}$ is between s_0 and s and $z_{n-1}(x, \bar{s}) = (\partial/\partial s) D_{n-1} y(x, \bar{s})$ is the solution of the IVP (14) with $y(x) = y(x, \bar{s})$, $D_i z(a) = 0$, $0 \leqq i \leqq n-2$, $D_{n-1} z(a) = 1$.

If $\lambda = 1$, let $\omega_{n-1}(x)$ be the solution of $D_n y = 0$ satisfying $D_j \omega_{n-1}(a) = 0$, $0 \leqq j \leqq n-2$, $D_{n-1} \omega_{n-1}(a) = 1$ while, if $\lambda = -1$, let $\omega_{n-1}(x)$ be the solution of $D_n y = -p(x) y$, satisfying the same initial conditions at a. It follows that (see [9, Lemma 2.1]

$$D_{n-1} z_{n-1}(x, \bar{s}) \geqq D_{n-1} \omega_{n-1}(x) \quad \text{on } [a, b].$$

Hence, for $s > s_0$,

$$D_{n-1} y(b, s) - D_{n-1} y(b, s_0) = (s - s_0) D_{n-1} z_{n-1}(b, \bar{s}) \geqq (s - s_0) D_{n-1} \omega_{n-1}(b).$$

Since $D_{n-1} \omega_{n-1}(b) > 0$,

$$\lim_{s \to \infty} D_{n-1} y(b, s) = \infty.$$

Similarly

$$\lim_{s \to -\infty} D_{n-1} y(b, s) = -\infty.$$

Since $\{D_{n-1} y(b, s) : s \in R\}$ is connected we have that every BVP of the form

$$\begin{aligned} D_n y &= \lambda f(x, y), \\ D_i y(a) &= A_i, \\ D_{n-1} y(b) &= B_{n-1} \end{aligned}$$

has a solution.

Assume we have existence for j-focal BVP's, $j = n-1, n-2, \cdots, k+1$. We want to show that the k-focal BVP (13)

$$\begin{aligned} D_i y(a) &= A_i, \qquad 0 \leqq i \leqq k-1, \\ D_j y(b) &= B_j, \qquad k \leqq j \leqq n-1 \end{aligned}$$

has a solution. To this end, this time let $y(x, s)$ be the solution, guaranteed by the induction hypothesis, of the $(k+1)$-focal BVP (13)

$$\begin{aligned} D_i y(a) &= A_i, \qquad 0 \leqq i \leqq k-1, \\ D_k y(a) &= s, \\ D_j y(b) &= B_j, \qquad k+1 \leqq j \leqq n-1. \end{aligned}$$

By Theorems 2 and 3,

$$D_k y(b, s) - D_k y(b, s_0) = (s - s_0) \frac{\partial}{\partial s} D_k y(b, \bar{s})$$

$$= (s - s_0) D_k z_k(b, \bar{s}),$$

where $\bar{s}$ is between s_0 and s and $z_k(x, \bar{s})$ is the solution of the variational equation (14), with $y(x) = y(x, \bar{s})$ satisfying

$$D_i z(a, \bar{s}) = 0, \qquad 0 \leqq i \leqq k-1,$$

$$D_k z(a, \bar{s}) = 1,$$

$$D_j z(b, \bar{s}) = 0, \qquad k+1 \leqq j \leqq n-1.$$

Using $0 \leqq f_y(x, y) \leqq p(x)$, $x \in [a, b]$, $y \in R$ and Corollary 8, we have that

$$(-1)^{n-k-1} \lambda D_k y_k(x; p(x)) \geqq (-1)^{n-k-1} \lambda D_k z_k(x, \bar{s}) \geqq (-1)^{n-k-1} \lambda D_k y_k(x; 0).$$

Letting $x = b$, we get that

$$(-1)^{n-k-1} \lambda D_k y_k(b; p(x)) \geqq (-1)^{n-k-1} \lambda D_k z_k(b, \bar{s}) \geqq (-1)^{n-k-1} \lambda D_k y_k(b; 0).$$

Let $\omega(x) = y_k(x; p(x))$ if $(-1)^{n-k-1}\lambda$ is negative, and let $\omega(x) = y_k(x; 0)$ if $(-1)^{n-k-1}\lambda$ is positive. Then

$$D_k z_k(b, \bar{s}) \geqq D_k \omega(b) > 0.$$

Hence for $s > s_0$

$$D_k y(b, s) - D_k y(b, s_0) = (s - s_0) D_k z_k(b, \bar{s})$$

$$\geqq (s - s_0) D_k \omega(b).$$

It follows that

$$\lim_{s \to \infty} D_k y(b, s) = \infty.$$

Similarly,

$$\lim_{s \to -\infty} D_k y(b, s) = -\infty.$$

It follows that all k-focal BVP's for (13) have solutions.

For the next theorem assume $1 \leqq p, q \leqq n-1$ and $p + q = n$. [7, Theorem 10] is the analogue of Theorem 9 for the (p, q)-BVP

$$D_n y = f(x, y),$$

$$D_i y(a) = 0, \qquad i = 0, \cdots, p-1,$$

$$D_j y(b) = 0, \qquad j = 0, \cdots, q-1.$$

We were not able to extend Theorem 9 to the Lipschitz case, but while trying to do so we were able to extend [7, Theorem 10] to the Lipschitz case. Although it may not be evident at first glance, we use the shooting method in this theorem (see the proof of [7, Theorem 12] which we use here). We now state and prove this result.

THEOREM 10. *Assume* (A) *holds and there are continuous functions* $k_1(x)$, $k_2(x)$ *such that*

$$k_1(x)[y - z] \leqq f(x, y) - f(x, z) \leqq k_2(x)[y - z]$$

for $x \in [a, b]$, $y \geqq z$. *If*

$$D_n y = k_1(x) y \tag{15}$$

and

$$D_n y = k_2(x) y \tag{16}$$

are disconjugate on $[a, b]$, *then the* BVP

$$\begin{aligned} &D_n y = f(x, y), \\ &D_i y(a) = A_i, \qquad 0 \leqq i \leqq p-1, \\ &D_j y(b) = B_j, \qquad 0 \leqq j \leqq q-1 \end{aligned}$$

has a unique solution.

Proof. By [5, Theorem 7] (see also [5, Theorem 6] and [9, Theorem 3.1]) all $(n-1, 1)$-BVP's on $[a, b]$ have unique solutions.

Extend definitions of k_1, k_2 and f by defining $k_1(x) = k_1(b)$, $k_2(x) = k_2(b)$, and $f(x, y) = f(b, y)$ for $x \geqq b$. We claim that (13) is disconjugate (see [13] for definition) on $[a, b]$. Let $\varepsilon > 0$ be given and assume (13) is not disconjugate on $[a, b]$. Then by [2, Theorem 2] there is an $\varepsilon > 0$ and there are distinct solutions $y_1(x)$, $y_2(x)$ such that $y_1(x) - y_2(x)$ has at least n distinct zeros on $[a, b+\varepsilon)$. Without loss of generality we can assume (15) and (16) are disconjugate on $[a, b+\varepsilon)$. Let $u_\delta(x)$ be the solution of the IVP (13), $D_i u_\delta(a) = D_i y_1(a)$, $0 \leqq i \leqq n-2$, $D_{n-1} u_\delta(a) = D_{n-1} y_1(a) + \delta$. Because of the $(n-1, 1)$-disconjugacy mentioned at the outset of the proof, $u_\delta(x) > y_1(x)$ on $(a, b+\varepsilon)$ if $\delta > 0$, while $u_\delta(x) < y_1(x)$ on $(a, b+\varepsilon)$ if $\delta < 0$. Depending on the zeros of $y_1(x) - y_2(x)$, there is either a positive δ or negative δ (call it δ_0) such that $u_{\delta_0}(x) - y_2(x)$ has at least n distinct odd-ordered zeros in $[a, b+\varepsilon)$. Define for each integer $k \geqq 1$ the integral mean

$$f_k(x, y) = \frac{k}{2} \int_{y-1/k}^{y+1/k} f(x, \tau)\, d\tau, \qquad (x, y) \in [a, b] \times R.$$

Note that $f_k(x, y)$ and $\partial f_k(x, y)/\partial y$ are continuous on $[a, b] \times R$, $k_1(x) \leqq \partial f_k(x, y)/\partial y \leqq k_2(x)$ on $[a, b] \times R$, and $\lim_{k \to \infty} f_k(x, y) = f(x, y)$ uniformly on compact subsets of $[a, b] \times R$. Let $\omega_k(x)$ and $v_k(x)$ be the solutions of $D_n y = f_k(x, y)$ satisfying the same initial conditions at a as $u_{\delta_0}(x)$ and $y_2(x)$ respectively. Then for k sufficiently large $\omega_k(x) - v_k(x)$ has at least n zeros on $[a, b+\varepsilon)$.

This contradicts the fact that $D_n y = f_k(x, y)$ is disconjugate on $[a, b+\varepsilon)$ (by [7, Corollary 6] the variational equations of $D_n y = f_k(x, y)$ are disconjugate on $[a, b+\varepsilon)$, so by [8, Theorem 1] $D_n y = f_k(x, y)$ is disconjugate on $[a, b+\varepsilon)$).

Since $D_n y = f(x, y)$ is disconjugate on $[a, b]$, we have by [7, Theorem 12] that every (p, q)-BVP for $D_n y = f(x, y)$ has a unique solution. Note that in [7, Theorem 12] the shooting method is used.

REFERENCES

[1] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[2] L. JACKSON, *Uniqueness of solutions of boundary value problems for ordinary differential equations*, SIAM J. Appl. Math., 24 (1973), pp. 535–538.

[3] Z. NEHARI, *Disconjugate linear differential operators*, Trans. Amer. Math. Soc., 129 (1967), pp. 500–516.

[4] A. PETERSON, *On the sign of the Green's function beyond the interval of disconjugacy*, Rocky Mountain J. Math., 3 (1973), pp. 41–51.

[5] A. PETERSON, *Comparison theorems for boundary value problems*, J. Math. Anal. Appl., 52 (1975), pp. 573–582.
[6] ———, *On the sign of Green's functions*, J. Differential Equations, 21 (1976), pp. 167–178.
[7] ———, *Comparison theorems and existence theorems for ordinary differential equations*, J. Math. Anal. Appl., 55 (1976), pp. 773–784.
[8] ———, *An expression for the first conjugate point for an n-th order nonlinear differential equation*, Proc. Amer. Math. Soc., 61 (1976), pp. 300–304.
[9] ———, *Existence-uniqueness for two-point boundary value problems for n-th order nonlinear differential equations*, Rocky Mountain J. Math., 7 (1977), pp. 103–109.
[10] ———, *Existence-uniqueness for ordinary differential equations*, J. Math. Anal. Appl., 64 (1978), pp. 166–172.
[11] ———, *Green's functions for focal type boundary value problems*, Rocky Mountain J. Math., 9 (1979), pp. 721–732.
[12] ———, *Focal Green's functions for fourth order differential equations*, J. Math. Anal. Appl., to appear.
[13] J. SPENCER, *Boundary value functions for nonlinear differential equations*, J. Differential Equations, 19 (1975), pp. 1–20.
[14] D. SUKUP, *On the existence of solutions to multipoint boundary value problems*, Rocky Mountain J. Math., 6 (1976), pp. 357–375.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0006 $01.00/0

INVERSE RELATIONS FOR CERTAIN SHEFFER SEQUENCES*

JAMES WARD BROWN† AND STEVEN M. ROMAN‡

Abstract. Let $s_n(x)$ $(n=0,1,2,\cdots)$ be a so-called Sheffer sequence of polynomials, and let a_n $(n=0,1,2,\cdots)$ be a sequence of the type $a_n = yn+z$ where y and z are constants. An expansion formula for each polynomial $s_n(x)$ in terms of the sequence $s_n(x+a_n)$ $(n=0,1,2,\cdots)$ is derived, and the formula is illustrated by applications to Laguerre, Hermite, and Gegenbauer polynomials.

1. Statement of main result. In 1939 Sheffer [13] initiated serious study of a class of polynomial sequences which have come to be known as Sheffer sequences. See, for example, [2], [11] and [12], where many additional references are given. These sequences have been characterized in a variety of ways, and we choose here to take as our starting point a generating function characterization that Sheffer himself originally gave. To be precise, a polynomial sequence $s_n(x)$ $(n=0,1,2,\cdots)$ is said to be a *Sheffer sequence* if it is generated by a relation of the form

$$G(t)\exp(xH(t)) = \sum_{n=0}^{\infty} s_n(x)\frac{t^n}{n!}, \tag{1.1}$$

where

$$G(t) = \sum_{n=0}^{\infty} g_n t^n \quad (t_0 \neq 0) \quad \text{and} \quad H(t) = \sum_{n=1}^{\infty} h_n t^n \quad (h_1 \neq 0). \tag{1.2}$$

All of the series here and in what follows are formal power series over the real or complex field.

Associated with any given Sheffer sequence $s_n(x)$ is a polynomial sequence $p_n(x)$ $(n=0,1,2,\cdots)$ of *binomial type* generated by

$$\exp(xH(t)) = \sum_{n=0}^{\infty} p_n(x)\frac{t^n}{n!}, \tag{1.3}$$

where the $H(t)$ is the same as in (1.1). In view of the additivity property of the exponential function, it is evident from (1.3) that the polynomials $p_n(x)$ satisfy the binomial-type identity

$$p_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} p_k(x)p_{n-k}(y), \qquad n=0,1,2,\cdots. \tag{1.4}$$

Note too that it follows from (1.1) and (1.3) that a similar relation,

$$s_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} s_k(x)p_{n-k}(y), \qquad n=0,1,2,\cdots, \tag{1.5}$$

relates any Sheffer sequence to the sequence of binomial type associated with it.

* Received by the editors March 19, 1979, and in final form September 16, 1980.

† Department of Mathematics and Statistics, The University of Michigan-Dearborn, Dearborn, Michigan 48128. The research of this author was partially supported by a Campus Grant from The University of Michigan-Dearborn.

‡ Department of Mathematics, California State University, Fullerton, California 92634. The research of this author was partially supported by the National Science Foundation under grant MCS 79-00911.

Suppose now that a_n $(n = 0, 1, 2, \cdots)$ is a sequence of the form

$$a_n = yn + z,$$

where y and z are constants, independent of x and n, and where $a_n \neq 0$ for any n. Very recently in [4] the first author showed that the sequence $s_n(x + a_n)$ $(n = 0, 1, 2, \cdots)$ is itself a Sheffer sequence, and we note here that the expansion

$$s_n(x + a_n) = \sum_{k=0}^{n} \binom{n}{k} p_{n-k}(a_n) s_k(x), \qquad n = 0, 1, 2, \cdots \tag{1.6}$$

of each polynomial $s_n(x + a_n)$ in terms of the sequence $s_n(x)$ is immediate from (1.5).

In the important special case of Appell sequences [1], occurring when $H(t) = t$ in (1.1) and (1.3) and therefore when $p_n(x) = x^n$ in (1.3), a pair of inverse relations obtained by Gould in [5] can be rewritten in such a fashion as to invert (1.6) and thus expand each polynomial $s_n(x)$ in terms of the sequence $s_n(x + a_n)$. To be precise, if we set $a = z$, $b = y$ and put

$$F(n) = \frac{(-1)^n s_n(x)}{n!\, a_n}, \qquad f(n) = \frac{s_n(x + a_n)}{a_n^{n+1}}$$

in Gould's

$$F(n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{(a + bk)^n}{n!} f(k), \tag{1.7}$$

$$\frac{(a + bn)^n}{n!} f(n) = \sum_{k=0}^{n} (-1)^k \frac{(a + bn)^{n-k}}{(n-k)!} F(k) \frac{a + bk}{a + bn}, \tag{1.8}$$

we find that (1.8) becomes (1.6) and (1.7) becomes

$$s_n(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{a_n}{a_k} p_{n-k}(-a_k) s_k(x + a_k), \qquad n = 0, 1, 2, \cdots. \tag{1.9}$$

Expansion (1.9) is valid, moreover, when $H(t) = \log(1 + t)$, in which case

$$p_n(x) = \binom{x}{n} n!.$$

It is readily obtained by setting $a = -z$, $b = 1 - y$ and writing

$$F(n) = \frac{s_n(x)}{n!\, a_n}, \qquad f(n) = \frac{(-1)^n s_n(x + a_n)}{\binom{n - a_n}{n} n!\, a_n}$$

in the inverse relations

$$F(n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{a + bk}{n} f(k), \tag{1.10}$$

$$\binom{a + bn}{n} f(n) = \sum_{k=0}^{n} (-1)^k \frac{a + bk - k}{a + bn - k} \binom{a + bn - k}{n - k} F(k), \tag{1.11}$$

also derived by Gould in [5]. Here (1.11) and (1.10) become (1.6) and (1.9), respectively.

The above suggests that expansion (1.9) may actually be valid for *any* Sheffer sequence, and our main object is to show that this is in fact the case. Once (1.9) has been established, we also have the following expansion, obtained by letting the z in $a_n = yn + z$ and $a_k = yk + z$ there tend to zero:

$$s_n(x) = \sum_{k=0}^{n} c_k s_k(x + yk), \qquad n = 1, 2, \cdots, \tag{1.12}$$

where

$$c_k = \begin{cases} -n!\, n y h_n, & k = 0, \\ \binom{n}{k} \frac{n}{k} p_{n-k}(-yk), & k = 1, 2, \cdots, n, \end{cases} \tag{1.13}$$

the h_n's being the coefficients in (1.2). To see this, we need to pay special attention to the first ($k = 0$) coefficient,

$$(yn + z)\frac{p_n(-z)}{z},$$

in (1.9) since it is undefined when $z = 0$. According to (1.3), however, $p_n(0) = 0$ and $p_n'(0) = h_n n!$ ($n = 1, 2, \cdots$); and l'Hôpital's rule reveals that

$$\lim_{z \to 0} \frac{p_n(-z)}{z} = -p_n'(0) = -h_n n!.$$

This gives c_0, the remaining coefficients in (1.9) being well defined when $z = 0$.

We shall derive (1.9), our main result, in two different ways. The first (§ 2) is more classical in nature and makes direct use of Lagrange's expansion formula. The second (§ 3) relies on the theory of Sheffer sequences from the more modern point of view of linear operators and linear functionals. That point of view has been intensively developed during the past decade and goes by the name *umbral calculus.*

Finally, in § 4 we illustrate the use of (1.9) in obtaining a variety of expansions, many of them evidently new, involving well-known special functions. We confine our illustrations to Laguerre, Hermite, and Gegenbauer polynomials. An extensive listing of other Sheffer sequences to which our main result can be applied is found, for example, in [2].

2. Derivation I. We begin our first derivation of (1.9) by writing the series

$$S = \sum_{n=0}^{\infty} \frac{s_n(x)}{a_n} \frac{t^n}{n!} \tag{2.1}$$

in the form

$$S = \sum_{n=0}^{\infty} [\exp(a_n H(t))] \frac{s_n(x) \exp(-a_n H(t))}{a_n} \frac{t^n}{n!}. \tag{2.2}$$

We then appeal to Lagrange's expansion formula [7, p. 145],

$$F(t) = F(0) + \sum_{k=1}^{\infty} \left\{ \frac{d^{k-1}}{dt^{k-1}} [F'(t)(f(t))^k] \right\}_{t=0} \frac{1}{k!} \left(\frac{t}{f(t)} \right)^k, \tag{2.3}$$

where $F(t)$ and $f(t)$ have formal Maclaurin series expansions and $f(0) \neq 0$. In that

formula we put

$$F(t) = \exp(a_n H(t)) \quad \text{and} \quad f(t) = \exp(yH(t)),$$

and also observe from (1.3), when viewed as a formal Maclaurin series with n replaced by k, that $p_0(x) = 1$ and

$$p_k(x) = \left\{ \frac{d^k}{dt^k} \exp(xH(t)) \right\}_{t=0}$$
$$= \left\{ \frac{d^{k-1}}{dt^{k-1}} [xH'(t) \exp(xH(t))] \right\}_{t=0}, \qquad k = 1, 2, \cdots.$$

Equation (2.3) then becomes

$$\exp(a_n H(t)) = \sum_{k=0}^{\infty} \frac{a_n}{a_{n+k}} p_k(a_{n+k}) \frac{[t \exp(-yH(t))]^k}{k!};$$

using this to substitute for the factor in square brackets in (2.2), we have

$$S = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{\infty} \binom{n+k}{k} p_k(a_{n+k}) s_n(x) \right\} \frac{\exp(-a_{n+k} H(t))}{a_{n+k}} \frac{t^{n+k}}{(n+k)!},$$

or

$$S = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{n} \binom{n}{k} p_k(a_n) s_{n-k}(x) \right\} \frac{\exp(-a_n H(t))}{a_n} \frac{t^n}{n!}. \tag{2.4}$$

Now, in view of (1.5), the factor in braces in (2.4) can be written $s_n(x + a_n)$; and so (2.4) becomes

$$S = \sum_{n=0}^{\infty} \frac{s_n(x + a_n)}{a_n} \frac{t^n}{n!} \exp(-a_n H(t)). \tag{2.5}$$

Replacing the variable of summation n here by k and then observing from (1.3) that

$$\exp(-a_k H(t)) = \sum_{n=0}^{\infty} p_n(-a_k) \frac{t^n}{n!},$$

we find that

$$S = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \binom{n+k}{k} \frac{p_n(-a_k)}{a_k} s_k(x + a_k) \frac{t^{n+k}}{(n+k)!},$$

or

$$S = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \binom{n}{k} \frac{p_{n-k}(-a_k)}{a_k} s_k(x + a_k) \frac{t^n}{n!}. \tag{2.6}$$

Finally, if we equate coefficients of $t^n/n!$ on the right-hand sides of (2.1) and (2.6), we arrive at (1.9).

3. Derivation II. We preface our second derivation of (1.9) with a summary of relevant results from the umbral calculus. In fact, most of this section is devoted to providing background to these recently developed methods, and our second derivation of (1.9) is actually shorter than the first. No proofs are given here; rather we refer the reader to [11]. For even more recent developments and generalizations of the umbral calculus, see [8], [9] and [10].

Let us start by defining three algebras. The first algebra P is the familiar algebra of polynomials in a single variable x over the real or complex field.

The second algebra P^* is the dual vector space of linear functionals on P endowed with the following product. Let L and M be linear functionals. We denote the action of a linear functional N on a polynomial $p(x)$ by $\langle N|p(x)\rangle$, and define the product LM by

$$\langle LM|x^n\rangle = \sum_{k=0}^{n} \binom{n}{k} \langle L|x^k\rangle\langle M|x^{n-k}\rangle.$$

It is easy to verify that P^* is an associative and commutative algebra with identity ε defined by

$$\langle \varepsilon|p(x)\rangle = p(0).$$

We call P^* the *umbral algebra.* A particularly important role is played by the *delta functionals*, namely those functionals L for which $\langle L|1\rangle = 0$ and $\langle L|x\rangle \neq 0$. Among these is the *generator* A defined by $\langle A|p(x)\rangle = p'(0)$, where $p'(x)$ is the derivative of $p(x)$. If a is a constant, the *evaluation functional* ε_a is defined by $\langle \varepsilon_a|p(x)\rangle = p(a)$. Note that $\varepsilon_0 = \varepsilon$ where ε is the identity defined above. Finally, we mention that a suitable topology can be put on P^*, allowing us to consider formal power series in a linear functional. It then holds that for any sequence of constants $a_k (k = 0, 1, 2, \cdots)$ the series $\sum_{k=0}^{\infty} a_k L^k$ converges if L is a delta functional. The umbral algebra becomes, moreover, the algebra of all formal power series in the generator A, or in any delta functional (see Theorem D below).

The third algebra S is the algebra of all linear operators on P, under composition, which commute with the derivative operator; that is, the elements of S are all linear operators T such that

$$TDp(x) = DTp(x)$$

for all $p(x) \in P$. We call S the algebra of *shift-invariant operators.* Again with a suitable topology, one may characterize S as the algebra of all formal power series in D.

Thus both P^* and S are isomorphic to the algebra of formal power series in a single variable, and so to each other. In fact, the map $\mu: P^* \to S$ sending the generator A to the derivative D can be extended to a continuous algebra isomorphism of P^* onto S. In other words, if $L = \sum_{k=0}^{\infty} a_k A^k$, then $\mu(L) = \sum_{k=0}^{\infty} a_k D^k$. A *delta operator* is the image of a delta functional under μ. In terms of formal power series, the adjective "delta" means zero constant term and nonzero linear term. The evaluation functional ε_a in P^* corresponds to the *shift operator*

$$E^a = \mu(\varepsilon_a): p(x) \to p(x + a)$$

in S.

A basic tool of the umbral calculus is the interplay between P^* and S that is described in the following theorem.

THEOREM A. *Let L and M be linear functionals. Then*

$$\langle LM|p(x)\rangle = \langle L|\mu(M)p(x)\rangle$$

for all $p(x) \in P$.

By a *sequence* of polynomials $p_n(x)$ $(n = 0, 1, 2, \cdots)$, we imply that $\deg p_n(x) = n$. A sequence $p_n(x)$ is of *binomial type* if

$$p_n(x + y) = \sum_{k=0}^{n} \binom{n}{k} p_k(x)p_{n-k}(y), \qquad n = 0, 1, 2, \cdots, \tag{3.1}$$

for all x and y. A polynomial sequence $s_n(x)$ $(n = 0, 1, 2, \cdots)$ is a *Sheffer sequence* if there is a sequence $p_n(x)$ of binomial type such that

$$s_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} s_k(x) p_{n-k}(y), \qquad n = 0, 1, 2, \cdots, \tag{3.2}$$

for all x and y. The reader will recall that this terminology was used in § 1. Characterizations (3.1) and (3.2) are, in fact, equivalent to the generating function characterizations (1.3) and (1.1), respectively, in that earlier section.

Now the key to the present theory is that sequences of Sheffer type (which includes binomial type) may be characterized by means of the algebras P^* and S.

THEOREM B. *A sequence $p_n(x)$ in P is of binomial type if and only if*

(i) *there exists a delta functional L such that*

$$\langle L^k | p_n(x) \rangle = n!\, \delta_{n,k},$$

or, in operator terms,

(ii) (a) $p_n(0) = \delta_{n,0}$,

(b) *there exists a delta operator $T(= \mu(L))$ such that*

$$Tp_n(x) = np_{n-1}(x), \qquad n = 1, 2, \cdots.$$

The sequence $p_n(x)$ is called the *associated sequence* for L (or T).

THEOREM C. *A sequence $s_n(x)$ in P is a Sheffer sequence if and only if*

(i) *there is an invertible linear functional N (i.e., $\langle N|1\rangle \neq 0$) and a delta functional L such that*

$$\langle NL^k | s_n(x) \rangle = n!\, \delta_{n,k},$$

or

(ii) *there exists an invertible shift-invariant operator T and a sequence $p_n(x)$ of binomial type such that*

$$s_n(x) = Tp_n(x),$$

or

(iii) *there exists a delta operator T such that*

$$Ts_n(x) = ns_{n-1}(x), \qquad n = 1, 2, \cdots.$$

The most useful result for our purposes is, however, the Expansion Theorem:

THEOREM D. *Let L be a delta functional with associated sequence $p_n(x)$. Then if M is any linear functional, we have*

$$M = \sum_{k=0}^{\infty} \frac{\langle M | p_k(x) \rangle}{k!} L^k.$$

In terms of shift-invariant operators, if $T = \mu(L)$ and $S = \mu(M)$, we obtain

$$S = \sum_{k=0}^{\infty} \frac{\langle M | p_k(x) \rangle}{k!} T^k.$$

We require one more result to complete our discussion. If T is a delta operator with associated sequence $p_n(x)$, then, for any constant a, E^aT is also a delta operator. Its associated sequence is given by

$$q_n(x) = xE^{-an}x^{-1}p_n(x) = \frac{x}{x-an} p_n(x-an). \tag{3.3}$$

We turn now to the derivation of expansion (1.9) using the umbral calculus. Actually, it is nothing more than a corollary of the Expansion Theorem. Let $s_n(x)$ be a Sheffer sequence and let T be the delta operator in part (iii) of Theorem C. Suppose that $p_n(x)$ is the associated sequence for T. Then, according to (3.3), the delta operator $E^{-y}T$ has the associated sequence

$$q_n(x) = \frac{x}{x+yn} p_n(x+yn).$$

If we write $a_n = yn + z$, the Expansion Theorem gives

$$E^{-a_n} = \sum_{k=0}^{\infty} \frac{\langle \varepsilon_{-a_n} | q_k(x) \rangle}{k!} (E^{-y}T)^k,$$

which may be written as

$$I = \sum_{k=0}^{\infty} \frac{q_k(-a_n)}{k!} E^{a_{n-k}} T^k.$$

Applying this to the polynomial $s_n(x)$, and noticing that

$$q_k(-a_n) = \frac{-a_n}{-a_n + yk} p_n(-a_n + yk) = \frac{a_n}{a_{n-k}} p_k(-a_{n-k}),$$

and

$$E^{a_{n-k}} T^k s_n(x) = \binom{n}{k} k! E^{a_{n-k}} s_{n-k}(x) = \binom{n}{k} k! s_{n-k}(x + a_{n-k}),$$

we obtain

$$s_n(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{a_n}{a_{n-k}} p_k(-a_{n-k}) s_{n-k}(x + a_{n-k}), \qquad n = 0, 1, 2, \cdots.$$

Replacing k by $n-k$ here finally gives (1.9).

4. Applications to special functions. The sequence of Laguerre polynomials

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n} \frac{n!}{k!} \binom{\alpha + n}{n-k} (-x)^k, \qquad n = 0, 1, 2, \cdots \tag{4.1}$$

generated by

$$(1-t)^{-1-\alpha} \exp\left(\frac{-xt}{1-t}\right) = \sum_{n=0}^{\infty} L_n^{(\alpha)}(x) \frac{t^n}{n!} \tag{4.2}$$

is a familiar Sheffer sequence. We follow Rota et al. [11], [12] here and in what follows immediately below, where we let $L_n(x)$ denote the basic Laguerre polynomials ($\alpha = -1$). It should be emphasized that other authors often do not include the $n!$ on the right-hand sides of (4.1) and (4.2) and use $L_n(x)$ for the case $\alpha = 0$.

It follows from (1.9) that

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n} \binom{n}{k} \frac{yn+z}{yk+z} L_{n-k}(-yk-z) L_k^{(\alpha)}(x+yk+z), \qquad n = 1, 2, \cdots, \tag{4.3}$$

and (1.12)–(1.13), with $h_n = -1$, tells us that the limiting case of (4.3) as $z \to 0$ is

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n} c_k L_k^{(\alpha)}(x+yk), \qquad n = 1, 2, \cdots, \tag{4.4}$$

where

$$c_k = \begin{cases} n!\, ny, & k = 0, \\ \binom{n}{k} \frac{n}{k} L_{n-k}(-yk), & k = 1, 2, \cdots, n. \end{cases} \tag{4.5}$$

As pointed out in [4], $L_n^{(\alpha)}(x)$ is also a Sheffer sequence in the parameter α. For (4.2) can be put into the form

$$(1-t)^{-1} \exp\left(\frac{-xt}{1-t}\right) \exp\left(-\alpha \log(1-t)\right) = \sum_{n=0}^{\infty} L_n^{(\alpha)}(x) \frac{t^n}{n!}.$$

Here $s_n(\alpha) = L_n^{(\alpha)}(x)$, and

$$\sum_{n=0}^{\infty} p_n(\alpha) \frac{t^n}{n!} = (1-t)^{-\alpha} = \sum_{n=0}^{\infty} (-1)^n \binom{-\alpha}{n} t^n.$$

Evidently, then,

$$p_n(\alpha) = (-1)^n n! \binom{-\alpha}{n},$$

and (1.9), with $a_n = \beta n + \gamma$, yields

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n} (-1)^{n-k} \frac{n!}{k!} \frac{\beta n + \gamma}{\beta k + \gamma} \binom{\beta k + \gamma}{n-k} L_k^{(\alpha+\beta k+\gamma)}(x), \qquad n = 1, 2, \cdots. \tag{4.6}$$

Note that $h_n = 1/n$, and, according to (1.12)–(1.13), the special case of (4.6) as $\gamma \to 0$ is

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n} c_k L_k^{(\alpha+\beta k)}(x), \qquad n = 1, 2, \cdots, \tag{4.7}$$

where

$$c_k = \begin{cases} -n!\, \beta, & k = 0, \\ (-1)^{n-k} \frac{n!}{k!} \frac{n}{k} \binom{\beta k}{n-k}, & k = 1, 2, \cdots, n. \end{cases} \tag{4.8}$$

Expansion (4.6) was obtained earlier in [3], where the limiting case as $\gamma \to 0$ was not noted and where the full generality of Sheffer sequences does not appear. That earlier paper treated only the special case when $H(t) = -\log(1-t)$. As pointed out in [3], (4.6) includes the interesting special case

$$x^n = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!} \frac{\alpha + \beta n + n}{\alpha + \beta k + n} \binom{\alpha + \beta k + n}{n-k} L_k^{(\alpha+\beta k)}(x), \tag{4.9}$$

obtained by putting $\alpha = -n$, then replacing γ by $\alpha + n$, and finally observing from (4.1) that $L_n^{(-n)}(x) = (-x)^n$.

The Hermite polynomials $H_n(x)$ form a Sheffer sequence generated by

$$\exp(2xt - t^2) = \sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!}, \tag{4.10}$$

and (1.9) is therefore applicable. For brevity, we note here only the following special case, obtained from (1.12)–(1.13) when $h_n = 0$ $(n = 2, 3, \cdots)$ and $p_n(x) = (2x)^n$:

$$H_n(x) = \sum_{k=1}^{n} \binom{n}{k} \frac{n}{k} (-2yk)^{n-k} H_k(x + yk), \qquad n = 2, 3, \cdots. \tag{4.11}$$

Finally, except for a factor of $n!$, the sequence of Gegenbauer polynomials $C_n^{\lambda}(x)$ is of binomial type in the parameter λ since it is generated by

$$(1-2xt+t^2)^{-\lambda} = \sum_{n=0}^{\infty} C_n^{\lambda}(x)t^n,$$

or

$$\exp[-\lambda \log(1-2xt+t^2)] = \sum_{n=0}^{\infty} C_n^{\lambda}(x)t^n.$$

Here

$$s_n(\lambda) = p_n(\lambda) = n!\, C_n^{\lambda}(x),$$

and if we write $a_n = \mu n + \nu$, (1.9) becomes

$$(4.12)\qquad \sum_{k=0}^{n} \frac{1}{\mu k+\nu} C_k^{\lambda+\mu k+\nu}(x) C_{n-k}^{-\mu k-\nu}(x) = \frac{1}{\mu n+\nu} C_n^{\lambda}(x), \qquad n = 0, 1, 2, \cdots.$$

Noticing, moreover, that [6, p. 259]

$$H(t) = -\log(1-2xt+t^2) = \sum_{n=1}^{\infty} \frac{2T_n(x)}{n} t^n,$$

where $T_n(x)$ are the Chebyshev polynomials of the first kind, we find from (1.12)–(1.13) that

$$(4.13)\qquad \sum_{k=1}^{n} \frac{n}{k} C_k^{\lambda+\mu k}(x) C_{n-k}^{-\mu k}(x) = C_n^{\lambda}(x) + 2\mu T_n(x), \qquad n = 1, 2, \cdots.$$

Of particular interest because of their symmetry are the identities

$$(4.14)\qquad \sum_{k=0}^{n} \frac{1}{\mu k+\nu} C_k^{\mu k+\nu}(x) C_{n-k}^{-\mu k-\nu}(x) = 0, \qquad n = 1, 2, \cdots$$

and

$$(4.15)\qquad \sum_{k=1}^{n} \frac{n}{k} C_k^{\mu k}(x) C_{n-k}^{-\mu k}(x) = 2\mu T_n(x), \qquad n = 1, 2, \cdots,$$

obtained by putting $\lambda = 0$ in (4.12) and (4.13), respectively.

Acknowledgment. The authors wish to thank Professor Richard A. Askey for bringing them together. Without his initiative, two one-dimensional papers would have been written rather than one two-dimensional paper.

Note added in proof. It has been brought to the authors' attention that the expansion (1.9) is obtained independently and in a somewhat different form by H. Niederhausen in an M.I.T. Technical Report of February 1979 entitled *Sheffer polynomials for computing exact Kolmogorov-Smirnov and Renyi type distributions*, which is to appear in Ann. Statist.

REFERENCES

[1] P. APPELL, *Sur une classe de polynômes*, Ann. Sci. École Norm. Sup., (2) 9 (1880), pp. 119–144.

[2] R. P. BOAS, JR. AND R. C. BUCK, *Polynomial Expansions of Analytic Functions*, rev. ed., Academic Press, New York, 1964.

[3] J. W. BROWN, *A new identity for Laguerre and related polynomials*, Glasnik Mat., 5 (1970), pp. 247–250.

[4] ———, *On multivariable Sheffer sequences*, J. Math. Anal. Appl., 69 (1979), pp. 398–410.
[5] H. W. GOULD, *A series transformation for finding convolution identities*, Duke Math. J., 28 (1961), pp. 193–202.
[6] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, 3rd ed., Springer-Verlag, New York, 1966.
[7] G. PÓLYA AND G. SZEGÖ, *Problems and Theorems in Analysis I*, rev. ed., Springer-Verlag, Berlin, 1972.
[8] S. M. ROMAN, *The algebra of formal series*, Advances in Math., 31 (1979), pp. 309–329. (Errata sheet required.)
[9] ———, *The algebra of formal series II: Sheffer sequences*, J. Math. Anal. Appl., 74 (1980), pp. 120–143.
[10] ———, *The algebra of formal series in several variables*, J. Approx. Theory, 26 (1979), pp. 340–381.
[11] S. M. ROMAN AND G.-C. ROTA, *The umbral calculus*, Advances in Math., 27 (1978), pp. 95–188.
[12] G.-C. ROTA, D. KAHANER, AND A. ODLYZKO, *On the foundations of combinatorial theory. VIII: Finite operator calculus*, J. Math. Anal. Appl., 42 (1973), pp. 684–760.
[13] I. M. SHEFFER, *Some properties of polynomial sets of type zero*, Duke Math. J., 5 (1939), pp. 590–622.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0007 $01.00/0

SUMMATION FORMULAS FOR BASIC HYPERGEOMETRIC SERIES*

GEORGE GASPER†

Abstract. Summation formulas for basic hypergeometric series are derived which are q-analogues of Minton's [J. Math. Phys., 11 (1970), pp. 1375–1376] and Karlsson's [J. Math. Phys., 12 (1971), pp. 270–271] summation formulas for generalized hypergeometric series, and some interesting limit cases are considered.

1. Introduction. In [6] Minton showed that if a is a negative integer and $m_1, \cdots, m_p$ are nonnegative integers such that $-a \geqq m_1 + \cdots + m_p$, then

$$(1) \qquad {}_{p+2}F_{p+1}\left(\begin{matrix} a, b, b_1+m_1, \cdots, b_p+m_p \\ b+1, b_1, \cdots, b_p \end{matrix}; 1\right) = \frac{\Gamma(b+1)\Gamma(1-a)}{\Gamma(1+b-a)} \frac{(b_1-b)_{m_1} \cdots (b_p-b)_{m_p}}{(b_1)_{m_1} \cdots (b_p)_{m_p}},$$

where $(a)_n = a(a+1)\cdots(a+n-1)$, $(a)_0 = 1$, and, as usual, it is assumed that no denominator parameter in the generalized hypergeometric series is a negative integer or zero. Karlsson [5] showed that (1) also holds when a is not a negative integer provided that the series converges, i.e. if $\operatorname{Re}(-a) > m_1 + \cdots + m_p - 1$, and he deduced from (1) that

$$(2) \qquad {}_{p+1}F_p\left(\begin{matrix} a, b_1+m_1, \cdots, b_p+m_p \\ b_1, \cdots, b_p \end{matrix}; 1\right) = 0, \qquad \operatorname{Re}(-a) > m_1 + \cdots + m_p,$$

$$(3) \qquad {}_{p+1}F_p\left(\begin{matrix} (m_1+\cdots+m_p), b_1+m_1, \cdots, b_p+m_p \\ b_1, \cdots, b_p \end{matrix}; 1\right) = \frac{(-1)^{m_1+\cdots+m_p}(m_1+\cdots+m_p)!}{(b_1)_{m_1} \cdots (b_p)_{m_p}}.$$

It turned out that the ${}_3F_2$ cases of (1) and (2) were precisely the formulas that the author needed in [3] to prove that the functions $C_n^{(\alpha,\beta)}(e^{i\theta})$ defined by the generating function

$$(1-te^{-i\theta})^{-\alpha}(1-te^{i\theta})^{-\beta} = \sum_{n=0}^{\infty} C_n^{(\alpha,\beta)}(e^{i\theta})t^n$$

satisfy the orthogonality relation

$$\int_0^{2\pi} C_m^{(\alpha,\beta)}(e^{i\theta}) C_n^{(\alpha,\beta)}(e^{i\theta})(1-e^{-2i\theta})^{\alpha}(1-e^{2i\theta})^{\beta}\, d\theta = 0, \qquad n \neq m,$$

when $\alpha, \beta, \alpha+\beta > -1$, and to evaluate this integral for $n = m$. My interest in these functions arose from Greiner's observation [4] that they yield spherical harmonics on the Heisenberg group, and the fact that $C_n^{(\alpha,\alpha)}(e^{i\theta}) = C_n^{\alpha}(\cos\theta)$, where $C_n^{\alpha}(x)$ is the ultraspherical polynomial of degree n and order α. Since analogues of (1) and (2) for ${}_3\phi_2$ basic hypergeometric series were also needed in [3] to prove the orthogonality of q-analogues of the functions $C_n^{(\alpha,\beta)}(e^{i\theta})$ which include the continuous q-ultraspherical polynomials [2] as special cases, the author was led to consider the q-analogues of (1), (2), (3) and the limit cases contained in this paper.

* Received by the editors March 31, 1980. This work was supported in part by the National Science Foundation under grant MCS 76-06635 A01.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

2. Summation formulas. For $|z|<1$ the ${}_{p+1}\phi_p$ basic hypergeometric series is defined by

$$ {}_{p+1}\phi_p\left(\begin{matrix}a_1,\cdots,a_{p+1}\\ b_1,\cdots,b_p\end{matrix};q,z\right)=\sum_{n=0}^{\infty}\frac{(a_1;q)_n\cdots(a_{p+1};q)_n}{(q;q)_n(b_1;q)_n\cdots(b_p;q)_n}z^n, $$

where $(a;q)_n=(1-a)(1-aq)\cdots(1-aq^{n-1})$, $(a;q)_0=1$ and, as elsewhere, it is assumed that $|q|<1$ and no denominator parameter is 1 or a negative integer power of q. The derivation of our summation formulas depends on the expansion formula

$$ \begin{aligned}{}_{p+1}\phi_p\left(\begin{matrix}a_1,\cdots,a_p,b_pq^m\\ b_1,\cdots,b_p\end{matrix};q,z\right)&=\sum_{n=0}^{m}\frac{(q^{-m};q)_n(a_1;q)_n\cdots(a_p;q)_n}{(q;q)_n(b_1;q)_n\cdots(b_p;q)_n}(-zq^m)^nq^{n(1-n)/2}\\ &\quad\times{}_p\phi_{p-1}\left(\begin{matrix}a_1q^n,\cdots,a_pq^n\\ b_1q^n,\cdots,b_{p-1}q^n\end{matrix};q,zq^{m-n}\right),\qquad |z|<1;\end{aligned} \tag{4} $$

this is easily proved by using [7, (3.3.2.7)],

$$ {}_2\phi_1\left(\begin{matrix}a,q^{-n}\\ b\end{matrix};q,q\right)=\frac{(b/a;q)_n}{(b;q)_n}a^n, \tag{5} $$

with $a=q^{-m}$, $b=b_p$, in the left side of (4) and changing the order of summation. Formula (4) is a q-analogue of an expansion formula employed by Minton [6, (4)].

When $p=2$, formulas (4), (5) and the q-analogue of Gauss' formula [7, (3.3.2.5)] give

$$ \begin{aligned}{}_3\phi_2\left(\begin{matrix}a,b,b_1q^m\\ bq,b_1\end{matrix};q,a^{-1}q^{1-m}\right)&=\frac{(q;q)_\infty(bq/a;q)_\infty}{(bq;q)_\infty(q;a)_\infty}{}_2\phi_1\left(\begin{matrix}b,q^{-m}\\ b_1\end{matrix};q,q\right)\\ &=\frac{(q;q)_\infty(bq/a;q)_\infty(b_1/b;q)_m}{(bq;q)_\infty(q/a;q)_\infty(b_1;q)_m}b^m,\end{aligned} \tag{6} $$

provided $|a^{-1}q^{1-m}|<1$, where $(a;q)_\infty=\prod_{n=0}^{\infty}(1-aq^n)$. By induction it follows from (4) and (6) that if $m_1,\cdots,m_p$ are nonnegative integers and $|a^{-1}q^{1-m_1-\cdots-m_p}|<1$, then

$$ \begin{aligned}&{}_{p+2}\phi_{p+1}\left(\begin{matrix}a,b,b_1q^{m_1},\cdots,b_pq^{m_p}\\ bq,b_1,\cdots,b_p\end{matrix};q,a^{-1}q^{1-(m_1+\cdots+m_p)}\right)\\ &\quad=\frac{(q;q)_\infty(bq/a;q)_\infty(b_1/b;q)_{m_1}\cdots(b_p/b;q)_{m_p}}{(bq;q)_\infty(q/a;q)_\infty(b_1;q)_{m_1}\cdots(b_p;q)_{m_p}}b^{m_1+\cdots+m_p},\end{aligned} \tag{7} $$

which is the desired q-analogue of (1). Formula (1) can be obtained from (7) by replacing $a, b, b_1,\cdots,b_p$ by $q^a, q^b, q^{b_1},\cdots,q^{b_p}$ and letting $q\to 1$.

Setting $b_p=b$, $m_p=1$ and then replacing p by $p+1$ in (7) gives

$$ {}_{p+1}\phi_p\left(\begin{matrix}a,b_1q^{m_1},\cdots,b_pq^{m_p}\\ b_1,\cdots,b_p\end{matrix};q,a^{-1}q^{-(m_1+\cdots+m_p)}\right)=0,\qquad |a^{-1}q^{-(m_1+\cdots+m_p)}|<1, \tag{8} $$

while letting $b\to\infty$ in the case $a=q^{-(m_1+\cdots+m_p)}$ of (7) gives

$$ \begin{aligned}&{}_{p+1}\phi_p\left(\begin{matrix}q^{-(m_1+\cdots+m_p)},b_1q^{m_1},\cdots,b_pq^{m_p}\\ b_1,\cdots,b_p\end{matrix};q,1\right)\\ &\quad=\frac{(-1)^{m_1+\cdots+m_p}(q;q)_{m_1+\cdots+m_p}}{(b_1;q)_{m_1}\cdots(b_p;q)_{m_p}}q^{-(m_1+\cdots+m_p)(m_1+\cdots+m_p+1)/2},\end{aligned} \tag{9} $$

which are q-analogues of (2) and (3). Another q-analogue of (3) can be found by letting

$b \to 0$ in (7), to find that

$$
\begin{aligned}
&{}_{p+1}\phi_p\left(\begin{matrix} a, b_1q^{m_1}, \cdots, b_pq^{m_p} \\ b_1, \cdots, b_p \end{matrix}; q, a^{-1}q^{1-(m_1+\cdots+m_p)}\right) \\
&\quad = \frac{(-1)^{m_1+\cdots+m_p}(q;q)_\infty b_1^{m_1}\cdots b_p^{m_p}}{(q/a;q)_\infty (b_1;q)_{m_1}\cdots(b_p;q)_{m_p}} q^{m_1(m_1-1)/2+\cdots+m_p(m_p-1)/2}
\end{aligned} \tag{10}
$$

when $|a^{-1}q^{1-(m_1+\cdots+m_p)}|<1$.

We can also let $a \to \infty$ in (7), (8) and (10) to obtain the summation formulas

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \frac{(-1)^n (b;q)_n (b_1q^{m_1};q)_n \cdots (b_pq^{m_p};q)_n}{(q;q)_n (bq;q)_n (b_1;q)_n \cdots (b_p;q)_n} q^{n(n+1)/2-n(m_1+\cdots+m_p)} \\
&\quad = \frac{(q;q)_\infty (b_1/b;q)_{m_1}\cdots(b_p/b;q)_{m_p}}{(bq;q)_\infty (b_1;q)_{m_1}\cdots(b_p;q)_{m_p}} b^{m_1+\cdots+m_p},
\end{aligned} \tag{11}
$$

$$
\sum_{n=0}^{\infty} \frac{(-1)^n (b_1q^{m_1};q)_n \cdots (b_pq^{m_p};q)_n}{(q;q)_n (b_1;q)_n \cdots (b_p;q)_n} q^{n(n-1)/2-n(m_1+\cdots+m_p)} = 0, \tag{12}
$$

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \frac{(-1)^n (b_1q^{m_1};q)_n \cdots (b_pq^{m_p};q)_n}{(q;q)_n (b_1;q)_n \cdots (b_p;q)_n} q^{n(n+1)/2-n(m_1+\cdots+m_p)} \\
&\quad = \frac{(-1)^{m_1+\cdots+m_p}(q;q)_\infty b_1^{m_1}\cdots b_p^{m_p}}{(b_1;q)_{m_1}\cdots(b_p;q)_{m_p}} q^{m_1(m_1-1)/2+\cdots+m_p(m_p-1)/2}.
\end{aligned} \tag{13}
$$

In addition, if $a = q^{-n}$ and n is a nonnegative integer, then (7), (8) and (10) can be inverted to give

$$
\begin{aligned}
&{}_{p+2}\phi_{p+1}\left(\begin{matrix} q^{-n}, b, b_1q^{m_1}, \cdots, b_pq^{m_p} \\ bq, b_1, \cdots, b_p \end{matrix}; q, q\right) \\
&\quad = \frac{b^n (q;q)_n (b_1/b;q)_{m_1}\cdots(b_p/b;q)_{m_p}}{(bq;q)_n (b_1;q)_{m_1}\cdots(b_p;q)_{m_p}}, \qquad n \geqq m_1+\cdots+m_p,
\end{aligned} \tag{14}
$$

$$
{}_{p+1}\phi_p\left(\begin{matrix} q^{-n}, b_1q^{m_1}, \cdots, b_pq^{m_p} \\ b_1, \cdots, b_p \end{matrix}; q, q\right) = 0, \qquad n > m_1+\cdots+m_p, \tag{15}
$$

which are used in [3], and the following generalization of (9):

$$
{}_{p+1}\phi_p\left(\begin{matrix} q^{-n}, b_1q^{m_1}, \cdots, b_pq^{m_p} \\ b_1, \cdots, b_p \end{matrix}; q, 1\right) = \frac{(-1)^n (q;q)_n q^{-n(n+1)/2}}{(b_1;q)_{m_1}\cdots(b_p;q)_{m_p}}, \qquad n \geqq m_1+\cdots+m_p, \tag{16}
$$

which also follows by letting $b \to \infty$ in (14).

It would be of interest to see what partition theorems and other applications follow from these formulas. For applications of basic hypergeometric functions to partitions and number theory, see Andrews [1] and his references.

Addendum. Shortly after preprints of the above were circulated, the author received a letter from I. M. Gessel showing how Minton's formula (1) can be derived from the Lagrange interpolation formula, a letter from M. E.-H. Ismail generalizing the expansion formula (4) and pointing out that the hypergeometric limit case of (4) had been obtained by C. Fox in [Proc. Lond. Math. Soc. (2), 26 (1927), pp. 201–210], and a letter from L. Carlitz pointing out how finite differences can be used to prove (1) (this has also been observed for the terminating case by R. Askey several years ago) and, in

addition, to evaluate the series

$$(17)\qquad {}_{p+2}F_{p+1}\left(\begin{matrix}a, b, b_1+m_1, \cdots, b_p+m_p\\ b+m+1, b_1, \cdots, b_p\end{matrix}; 1\right)$$

for $m = 0, 1, \cdots$, as a sum of m terms. Less than two weeks after Carlitz's letter arrived, Mizan Rahman wrote to the author asking if there existed a transformation formula for the ${}_4F_3$ case of (17), since he needed one in his research on $9-j$ symbols. Here we shall show how simply (1) can be used to derive a transformation formula for the series (17) with m replaced by a complex parameter c, which gives Carlitz's sum when $c = m = 0, 1, \cdots$, and then give two q-analogues.

From the case $p = 0$ of (1), which is a special case of Gauss' formula, we have

$$\frac{(b)_j}{(b+c+1)_j} = \frac{\Gamma(b+c+1)}{(b+j)\Gamma(b)\Gamma(c+1)}\,{}_2F_1\left(\begin{matrix}-c, b+j\\ b+j+1\end{matrix}; 1\right),$$

and hence

$$\begin{aligned}&\frac{\Gamma(b)(c+1)}{\Gamma(b+c+1)}\,{}_{p+2}F_{p+1}\left(\begin{matrix}a, b, b_1+m_1, \cdots, b_p+m_p\\ b+c+1, b_1, \cdots, b_p\end{matrix}; 1\right)\\ &\quad=\sum_{j=0}^{\infty}\frac{(a)_j(b_1+m_1)_j\cdots(b_p+m_p)_j}{j!(b_1)_j\cdots(b_p)_j(b+j)}\sum_{k=0}^{\infty}\frac{(-c)_k(b+j)_k}{k!(b+j+1)_k}\\ &\quad=\sum_{k=0}^{\infty}\frac{(-c)_k}{k!(b+k)}\,{}_{p+2}F_{p+1}\left(\begin{matrix}a, b+k, b_1+m_1, \cdots, b_p+m_p\\ b+k+1, b_1, \cdots, b_p\end{matrix}; 1\right)\\ &\quad=\sum_{k=0}^{\infty}\frac{(-c)_k}{k!(b+k)}\,\frac{\Gamma(b+k+1)\Gamma(1-a)}{\Gamma(b+k+1-a)}\,\frac{(b_1-b-k)_{m_1}\cdots(b_p-b-k)_{m_p}}{(b_1)_{m_1}\cdots(b_p)_{m_p}}\end{aligned}$$

by (1), provided that the sums involved converge absolutely. From $(\alpha-k)_n = (\alpha)_n(1-\alpha)_k/(1-\alpha-n)_k$ and analytic continuation, it follows that if $\mathrm{Re}\,(c-a) > m_1+\cdots+m_p-1$, then

$$\begin{aligned}&{}_{p+2}F_{p+1}\left(\begin{matrix}a, b, b_1+m_1, \cdots, b_p+m_p\\ b+c+1, b_1, \cdots, b_p\end{matrix}; 1\right)\\ &\quad=\frac{\Gamma(b+c+1)\Gamma(1-a)}{\Gamma(b+1-a)\Gamma(c+1)}\,\frac{(b_1-b)_{m_1}\cdots(b_p-b)_{m_p}}{(b_1)_{m_1}\cdots(b_p)_{m_p}}\\ &\quad\times{}_{p+2}F_{p+1}\left(\begin{matrix}-c, b, 1+b-b_1, \cdots, 1+b-b_p\\ b+1-a, 1+b-b_1-m_1, \cdots, 1+b-b_p-m_p\end{matrix}; 1\right),\end{aligned}\qquad(18)$$

which is the desired formula. Here, as elsewhere, it is assumed that $m_1, \cdots, m_p$ are nonnegative integers. Analogously, it follows from (7) that if $|cq|<1$ and $|a^{-1}q^{1-(m_1+\cdots+m_p)}|<1$, then

$$\begin{aligned}&{}_{p+2}\phi_{p+1}\left(\begin{matrix}a, b, b_1q^{m_1}, \cdots, b_pq^{m_p}\\ bcq, b_1, \cdots, b_p\end{matrix}; q, a^{-1}q^{1-(m_1+\cdots+m_p)}\right)\\ &\quad=\frac{(bq/a;q)_\infty(cq;q)_\infty}{(bcq;q)_\infty(q/a;q)_\infty}\,\frac{(b_1/b;q)_{m_1}\cdots(b_p/b;q)_{m_p}}{(b_1;q)_{m_1}\cdots(b_p;q)_{m_p}}\,b^{m_1+\cdots+m_p}\\ &\quad\times{}_{p+2}\phi_{p+1}\left(\begin{matrix}c^{-1}, b, bq/b_1, \cdots, bq/b_p\\ bq/a, bq^{1-m_1}/b_1, \cdots, bq^{1-m_p}/b_p\end{matrix}; q, cq\right),\end{aligned}\qquad(19)$$

while if $|a^{-1}q^{m+1-(m_1+\cdots+m_p)}|<1$ and $m=0, 1, \cdots$, it follows from (7) and the $p=0$ case of (14) that

$$
\begin{aligned}
&{}_{p+2}\phi_{p+1}\left(\begin{matrix} a, b, b_1q^{m_1}, \cdots, b_pq^{m_p} \\ bq^{1+m}, b_1, \cdots, b_p \end{matrix}; q, a^{-1}q^{m+1-(m_1+\cdots+m_p)}\right) \\
&= \frac{(q;q)_\infty (bq/a;q)_\infty}{(bq;q)_\infty (q/a;q)_\infty} \frac{(bq;q)_m (b_1/b;q)_{m_1}\cdots(b_p/b;q)_{m_p}}{(q;q)_m (b_1;q)_{m_1}\cdots(b_p;q)_{m_p}} b^{m_1+\cdots+m_p-m} \\
&\quad \times {}_{p+2}\phi_{p+1}\left(\begin{matrix} q^{-m}, b, bq/b_1, \cdots, bq/b_p \\ bq/a, bq^{1-m_1}/b_1, \cdots, bq^{1-m_p}/b_p \end{matrix}; q, q\right).
\end{aligned} \tag{20}
$$

Of course, inversions (when the series terminate), limit cases, and (more complicated) generalizations of these formulas can be derived.

REFERENCES

[1] G. E. ANDREWS, *Applications of basic hypergeometric functions*, SIAM Rev., 16 (1974), pp. 441–484.

[2] R. ASKEY and M. E.-H. ISMAIL, *A generalization of ultraspherical polynomials*, to appear.

[3] G. GASPER, *Orthogonality of certain functions with respect to complex valued weights*, Canad. J. Math., to appear.

[4] P. GREINER, *Spherical harmonics on the Heisenberg group*, to appear.

[5] PER W. KARLSSON, *Hypergeometric functions with integral parameter differences*, J. Math. Phys., 12 (1971), pp. 270–271.

[6] B. M. MINTON, *Generalized hypergeometric functions of unit argument*, J. Math. Phys., 11 (1970), pp. 1375–1376.

[7] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0008 $01.00/0

THE RESOLVENT PROBLEM FOR THE STOKES EQUATIONS ON HALFSPACE IN L_p*

MARJORIE McCRACKEN†

Abstract. The resolvent problem for the Stokes equations on halfspace in R^3 is considered. Letting $H=\{(x_1, x_2, x_3)\in R^3 | x_3<0\}$ and given $\mathbf{f}\in L_p(H)$, we find $\mathbf{u}$, ∇p such that

$$\left.\begin{aligned}\lambda\mathbf{u}(x)-\nu\Delta\mathbf{u}(x)+\nabla p(x)&=\mathbf{f}(x),\\ \nabla\cdot u(c)&=0,\end{aligned}\right\} \quad x\in H,$$
$$\mathbf{u}|_{\partial H}=0.$$

We show that if $\lambda\not\leqq 0$ and if $\nu>0$ and $1<p<\infty$, the solution is unique and $\mathbf{u}\in W^{2,p}$ satisfies

$$|\lambda|\|\mathbf{u}\|_{L_p(H)}+\nu\|\Delta\mathbf{u}\|_{L_p(H)}+\|\nabla p\|_{L_p(H)}\leqq\|\mathbf{f}\|_{L_p(H)},$$

where c depends on p and $\arg\lambda$ only.

This enables us to prove that the nonstationary Stokes equations generate a bounded analytic semigroup on $L_p(H)$, $1<p<\infty$. That is, given $\mathbf{u}_0\in L_p(H)$, the problem

$$\left.\begin{aligned}\frac{\partial\mathbf{u}}{\partial t}(x,t)-\nu\Delta_x\mathbf{u}(x,t)+\nabla_x p(x,t)&=0,\\ \nabla\cdot\mathbf{u}(x,t)&=0,\end{aligned}\right\} \quad x\in H,$$
$$\mathbf{u}|_{\partial H}=0,$$
$$\mathbf{u}(x,0)=\mathbf{u}_0(x)$$

has a unique solution $\mathbf{u}$ satisfying the conditions that $\|\mathbf{u}\|_{L_p(H)}\leqq M\|u_0\|_{L_p(H)}$, that $\mathbf{u}$ is an analytic function of t, and other properties of analytic semigroups.

Introduction. In this paper we consider the resolvent problem for the Stokes equations on halfspace in R^3. Let $H=\{(x_1, x_2, x_3)\in R^3 | x_3<0\}$. Given $\mathbf{f}\in L_p(H)$, find $\mathbf{u}$, ∇p such that

$$\left.\begin{aligned}\lambda\mathbf{u}(x)-\nu\Delta\mathbf{u}(x)+\nabla p(x)&=\mathbf{f}(c),\\ \nabla\cdot\mathbf{u}(x)&=0,\end{aligned}\right\} \quad x\in H,$$
$$\mathbf{u}|_{\partial H}=0. \tag{0.1}$$

We shall show that if $\lambda\not\leqq 0$ and if $\nu>0$ and $1<0<\infty$, then the problem has a unique $W^{2,p}$ solution satisfying

$$|\lambda|\|\mathbf{u}\|_{W^{s,p}(H)}+\nu\|\Delta\mathbf{u}\|_{W^{s,p}(H)}+\|\nabla p\|_{W^{s,p}(H)}\leqq c\|\mathbf{f}\|_{W^{s,p}(H)},$$

where c depends on p and $\arg\lambda$ only.

This will enable us to prove that the nonstationary Stokes equations generate a bounded analytic semigroup on $J_p(H)$, the range of the Hodge projection in $L_p(H)$, $1<p<\infty$. That is, given $\mathbf{u}_0\in J_o(H)$, the problem

$$\left.\begin{aligned}\frac{\partial\mathbf{u}}{\partial t}(x,t)-\nu\Delta_x\mathbf{u}(x,t)+\nabla_x p(x,t)&=0,\\ \nabla\cdot\mathbf{u}(x,t)&=0,\end{aligned}\right\} \quad x\in H,$$
$$\mathbf{u}|_{\partial H}=0,$$
$$\mathbf{u}(x,0)=\mathbf{u}_0(x), \tag{0.2}$$

* Received by the editors June 5, 1979 and in revised form May 27, 1980.

† Department of Mathematics, Swain Hall East, Indiana University, Bloomington, Indiana 47401.

has a unique solution $\mathbf{u}$ satisfying the estimate $\|\mathbf{u}\|_{L_p(H)} \leqq M\|\mathbf{u}_0\|_{L_p(H)}$ and such that $\mathbf{u}$ is an analytic function of t and has the other properties of analytic semigroups.

The equations (0.2) have been studied in more general domains by V. A. Solonnikov [1], [2] who considered the problem in $L_P(H \times [0, T])$ and, more recently, by P. E. Sobolevsky [1], who used Solonnikov's results to show that (0.2) generate an analytic, but not necessarily bounded, semigroup on $L_P(\Omega)$, if Ω is a bounded domain. The resolvent problem (0.1) for $\lambda = 0$ has been studied by Ladyzhenskaya [1].

A few comments on this problem are in order. The result of this paper is certainly not surprising. It is predictable that the Stokes equations should behave like the heat equation; indeed, on a manifold without boundary, the Stokes problem and the heat problem are equivalent. What is surprising is that the result is so difficult to obtain and cannot easily be extended to more general domains except in the L_2 case.

The reasons for this are discussed in the concluding section. The fact that the Stokes equations in L_P are so difficult to analyze may have some effect on future work on the incompressible Navier-Stokes equations, which have traditionally been considered to be composed of the "good" heat part plus the "difficult" convection part. The work in this paper together with that of Ebin and Marsden [1], which shows the Euler equations in a very elegant and tractable form, suggest that the traditional view may be an oversimplification.

1. Statement of the problem. The resolvent problem for the Stokes equations in the halfspace $H = \{(x_1, x_2, x_3)\} R^3 | x^3 < 0\}$ is to find $\mathbf{u}$ and p satisfying the equations

$$\begin{aligned} \lambda \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } H, \end{aligned} \tag{1.1}$$

and such that

$$\mathbf{u}|_{\partial H} = 0,$$

where $\mathbf{f}$, λ, and ν are given. Here we take

$$\Delta \mathbf{u}(x) = \frac{\partial^2 \mathbf{u}}{\partial x_1^2} + \frac{\nabla^2 \mathbf{u}}{\partial x_2^2} + \frac{\partial^2 \mathbf{u}}{\partial x_3^2}, \qquad \nabla \cdot \mathbf{u}(x) = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3},$$

$$\nabla p(x) = \left(\frac{\partial p}{\partial x_1}, \frac{\partial p}{\partial x_2}, \frac{\partial p}{\partial x_3} \right).$$

The nonstationary, or evolutionary, problem is to find $\mathbf{u}$ and p satisfying the equations

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + \nabla_x p &= 0, \\ \nabla_x \cdot \mathbf{u} &= 0 \quad \text{in } H \end{aligned} \tag{1.2}$$

for positive t, and such that the conditions $\mathbf{u}|_{\partial H} = 0$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$ are also met. In this case, $\mathbf{u}_0$ and ν are given. The subscript x refers to differentiation with respect to space variables only.

We shall solve (1.1) explicitly and use the solution to solve (1.2). In order to do this, we shall put the equations in the form

$$\lambda \mathbf{u} - \nu \mathbb{P} \Delta \mathbf{u} = \mathbf{f}, \tag{1.3}$$

$$\frac{d\mathbf{u}}{dt} = \nu \mathbb{P} \Delta \mathbf{u}, \tag{1.4}$$

$\mathbf{u}|_{t=0} = \mathbf{u}_0$, respectively. Note that we have assumed that $\mathbb{P}\mathbf{f} = \mathbf{f}$. This can be done because the gradient part of $\mathbf{f}$ can be absorbed into the pressure.

The operator $\mathbb{P}$ above is the Hodge projection and $\mathbb{P}\Delta$ is a densely defined operator on the range of $\mathbb{P}$.

We first make some definitions and state a lemma.

DEFINITION 1.5. If $j = (j_1, j_2, j_3)$ is a 3-tuple of negative integers, then $D^j f(x) = (\partial^{j_1+j_2+j_3} f(x))/(\partial x_1^{j_1}\, \partial x_2^{j_2}\, \partial x_3^{j_3})$.

In the remaining definitions the subscript 0 denotes compact support.

DEFINITION 1.6. $W^{s,p}(H)$ is the closure of $C_0^\infty(\bar{H})$,

$$\|f\|_{W^{s,p}(H)} = \left(\sum_{j_1+j_2+j_3 \leqq s} \|D^j f\|^2_{L_p(H)} \right)^{1/2}.$$

DEFINITION 1.7. $\tilde{W}^{s,p}(H)$ is the closure of $\{f \in C_0^\infty(\bar{H}) | f(x_1, x_2, 0) = 0\}$.

DEFINITION 1.8. $J_p(H)$ is the closure of $\{f \in C_0^\infty(H) | \nabla \cdot f = 0\}$ in the L_p-norm.

LEMMA 1.9 (Hodge theorem in halfspace). *$J_p(H)$ is complemented in $L_p(H)$, $1 < p < \infty$. In fact, $L_p(H) = J_p(H) \oplus G_p(H)$, where $G_p(H)$ is the closure in $L_p(H)$ of $\{\nabla p | p \in C^\infty(H)$ and $\nabla p \in L_p(H)\}$. We will call the Hodge projection from $L_p(H) \to J_p(H)$ corresponding to this decomposition $\mathbb{P}$. In $L_2(H)$, the decomposition is orthogonal, so $\|\mathbb{P}\| = 1$.*

Although this result is not new, for convenience a proof of it is given in the appendix.

With the above definitions, we can put the Stokes equations in the form (1.3) and (1.4) by defining the operator $\nu\mathbb{P}\Delta$ on $J_p(H)$ to have the domain $\tilde{W}^{2,p}(H) \cap J_p(H)$.

We will compute the resolvent of the operator $\mathbb{P}\Delta$ and prove that, if $1 < p < \infty$, it satisfies the estimate

$$\|(\lambda - \nu\mathbb{P}\Delta)^{-1}\|_{J_p(H) \to J_p(H)} \leqq \frac{M_{p,\phi}}{|\lambda|}, \qquad \text{for all } \lambda \not\leqq 0 \quad \text{with } |\arg \lambda| < \phi < \pi. \tag{1.10}$$

Furthermore, in the case $p = 2$, the constant M is 1 if $\lambda > 0$. Hence, the operator $\nu\mathbb{P}\Delta$ generates a bounded, analytic semigroup on $L_p(H)$, $1 < p < \infty$, and a semigroup of contractions on $L_2(H)$ (see Friedman [1]).

2. Reduction of the problem by scaling. To prove that $\lambda - \nu\mathbb{P}\Delta$ is boundedly invertible is equivalent to proving that, for all $\mathbf{f} \in J_p(H)$, the equations

$$\begin{aligned} &\lambda \mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f}, \\ &\nabla \cdot \mathbf{u} = 0 \quad \text{in } H, \\ &\mathbf{u}(x_1, x_2, 0) = 0 \end{aligned} \tag{2.1}$$

can be uniquely solved for $\mathbf{u} \in \tilde{W}^{2,p}(H) \cap L_p(H)$ and that $\|\mathbf{u}\|_{L_p(H)} \leqq C\|\mathbf{f}\|_{L_p(H)}$, where C may depend on λ, ν. To see this, take the Hodge projection of (2.1).

However, because halfspace can be scaled, to prove (1.10) it is only necessary to prove (2.1) for $\nu = 1$ and for all $\lambda = e^{i\theta}$, where $|\theta| < \pi$. Estimate (1.8) then follows for all $\nu > 0$ and $\lambda \not\leqq 0$. Furthermore, the dependence of the L_p-norms of $\mathbf{u}$ and its derivatives on λ and ν can easily be found. First note that because $(\lambda - \nu\mathbb{P}\Delta)^{-1}$ depends analytically on λ, if $(e^{i\theta} - \mathbb{P}\Delta)^{-1}$ exists for all θ with $|\theta| < \pi$, then, given θ with $|\theta| < \theta_0 < \pi$, we will have $\|(e^{i\theta} - \mathbb{P}\Delta)^{-1}\| \leqq C_p(\theta_0)$. Letting K_α be the dilation by α,

that is, $(K_\alpha f)(x) = f(\alpha x)$, we get

LEMMA 2.2. *If* $(e^{i\theta} - \mathbb{P}\Delta)^{-1}$ *exists, then* $(re^{i\theta} - \nu\mathbb{P}\Delta)^{-1}$ *exists for all* $r, \nu > 0$ *and*

$$(re^{i\theta} - \nu\mathbb{P}\Delta)^{-1} = \frac{1}{r} K_{\sqrt{r/\nu}} (e^{i\theta} - \mathbb{P}\Delta)^{-1} K_{\sqrt{\nu/r}}. \tag{2.3}$$

Proof. Let $\mathbf{u} = (e^{i\theta} - \mathbb{P}\Delta)^{-1}\mathbf{f}$. Then there is a ∇p such that $e^{i\theta}\mathbf{u}(x) - \Delta\mathbf{u}(x) + \nabla p(x) = \mathbf{f}(x)$, for all $x \in H$. Let $\hat{u}(x) = (1/r)\mathbf{u}(\sqrt{r/\nu}x)$ and $\hat{p}(x) = \sqrt{\nu/r}p(\sqrt{r/\nu}x)$. Then $re^{i\theta}\hat{u} - \nu\Delta\hat{u} + \nabla\hat{p} = K_{\sqrt{r/\nu}}\mathbf{f}$. This shows that $re^{i\theta} - \nu\mathbb{P}\Delta$ is 1–1 and onto and that $\hat{u} = (1/r)K_{\sqrt{r/\nu}}(e^{i\theta} - \mathbb{P}\Delta)^{-1}\mathbf{f} = (re^{i\theta} - \nu\mathbb{P}\Delta)^{-1}K_{\sqrt{r/\nu}}\mathbf{f}$. So it follows that $(re^{i\theta} - \nu\mathbb{P}\Delta)^{-1} = (1/r)K_{\sqrt{r/\nu}}(e^{i\theta} - \mathbb{P}\Delta)^{-1}K_{\sqrt{\nu/r}}$.

Because of the closed graph theorem, $(\lambda - \Delta\mathbb{P}\Delta)^{-1}$ will, if it exists, be continuous as a map from $J_p(H) \to \tilde{W}^{2,p}(H) \cap J_p(H)$, where the latter is considered as a closed subspace of $\tilde{W}^{2,p}(H)$. This together with (2.3) yields

LEMMA 2.4. *Let* $|\arg\lambda| < \theta_0 < \pi$ *and assume the hypothesis of Lemma* 2.2 *to be true. Then*

$$\left\| \frac{\partial}{\partial x_j} (\lambda - \nu\mathbb{P}\Delta)^{-1}\mathbf{f} \right\|_{L_p(H)} \leqq \frac{M_{p,\theta_0}}{\sqrt{|\lambda|\nu}} \|\mathbf{f}\|_{L_p(H)},$$

$$\left\| \frac{\partial^2}{\partial x_i \partial x_j} (\lambda - \nu\mathbb{P}\Delta)^{-1}\mathbf{f} \right\|_{L_p(H)} \leqq \frac{M_{p,\theta_0}}{\nu} \|\mathbf{f}\|_{L_p(H)}.$$

The proof is a trivial calculation using (2.3).

3. The free space problem. We shall find the solution $\mathbf{u}, \nabla p$ to (2.1) in the form $\mathbf{u} = \mathbf{v} + \mathbf{w}$, $\nabla p = \nabla q + \nabla\psi$, where $\mathbf{v}, \nabla q$ are the solutions to the free space problem (3.1) and $\mathbf{w}, \nabla\psi$ are the solutions to an appropriate boundary value problem. The free space problem, which we consider below, is

$$\lambda\mathbf{v} - \Delta\mathbf{v} + \nabla q = \mathbf{f}, \qquad \nabla \cdot \mathbf{v} = 0. \tag{3.1}$$

We assume $\mathbf{f} \in L_p(R^3)$, $1 < p < \infty$, $\lambda \not\leqq 0$ and $|\lambda| = 1$. In order to fix our conventions about constants we will define the Fourier transform $\hat{f}(\alpha) = (1/(2\pi)^{3/2}) \int_{R^3} f(x)\, e^{-i\alpha\cdot x} dy$ and the convolution $(f * g)(x) = \int_{R^3} f(x-y)g(y)\, dy$. We seek the solution to (3.1) in the form

$$\begin{aligned} v_k(x) &= \sum_{j=1}^{3} (u_j^k * f_j)(x), \\ \frac{\partial q}{\partial x_k}(x) &= \sum_{j=1}^{3} \left(\frac{\partial p^k}{\partial x_j} * f_j \right)(x). \end{aligned} \tag{3.2}$$

Thus, the fundamental solution $\mathbf{u}^k, p^k$ will have to satisfy

$$\lambda\mathbf{u}^k - \Delta\mathbf{u}^k + \nabla p^k = \delta\mathbf{e}^k, \qquad \nabla \cdot \mathbf{u}^k = 0, \tag{3.3}$$

where δ is the Dirac δ function, $\mathbf{e}^k$ is a unit vector in the k direction, and $\mathbf{u}^k = (u_1^k, u_2^k, u_3^k)$. When we have found the fundamental solution to this, we will show that

$$\mathbf{v} \in J_p(R^3) \cap W^{2,p}(R^3), \qquad \nabla q \in G_p(R^3),$$

$$\|\mathbf{v}\|_{W^{2,p}(R^3)} \leqq C_p \|\mathbf{f}\|_{L_p(R^3)}, \qquad \|\nabla q\|_{L_p(R^3)} \leqq C_p \|\mathbf{f}\|_{L_p(R^3)},$$

and finally that $\mathbf{v}, \nabla q$ satisfy (3.1).

To find the fundamental solution, we take the Fourier transform of (3.3), obtaining

$$\begin{aligned} &\lambda\hat{u}_j^k(\alpha)+|\alpha|^2\hat{u}_j^k(\alpha)+i\alpha_j\hat{p}^k(\alpha)=\delta_j^k\frac{1}{(2\pi)^{3/2}},\\ &\sum_{j=1}^{3}\alpha_j\hat{u}_j^k(\alpha)=0, \end{aligned} \tag{3.4}$$

where δ_j^k is the Kronecker δ.

Multiplying the first equation by α_j and then summing over j yields $\hat{p}^k(\alpha)=-i\alpha_k/((2\pi)^{3/2}|\alpha|)^2$. Substituting this in the first equation, we find that

$$\begin{aligned} \hat{u}_j^k(\alpha)&=\frac{1}{(2\pi)^{3/2}}\left(\frac{1}{\lambda+|\alpha|^2}\right)\left\{\delta_j^k-\frac{\alpha_j\alpha_k}{\alpha^2}\right\},\\ \hat{p}^k(\alpha)&=\frac{-i\alpha_k}{(2\pi)^{3/2}|\alpha|^2}. \end{aligned} \tag{3.5}$$

From this we see that $p^k(x)=(1/4\pi)(x_k/|x|^3)$. To calculate $u_j^k(x)$, note first that the inverse transform of $1/((2\pi)^{3/2}(\lambda+|\alpha|^2))$ is $(1/4\pi|x|)e^{-\sqrt{\lambda}|x|}$, as long as $\lambda\not\leqq 0$.

If we let

$$\hat{G}(\alpha)=\frac{1}{(2\pi)^{3/2}}\frac{1}{|\alpha|^2}\left(\frac{1}{\lambda+\alpha^2}\right),$$

we see that

$$\begin{aligned} G&=\left(\frac{1}{4\pi|x|}e^{-\sqrt{\lambda}|x|}\right)*\left(-\frac{1}{4\pi|x|}\right)\\ &=\frac{1}{(4\pi)^2}\int_{R^3}\frac{1}{|x-y|}\frac{1}{|y|}e^{-\sqrt{\lambda}|y|}dy\\ &=\frac{1}{(16\pi)^2}\int_0^\infty\int_{S_2}\frac{1}{|x-rs|}\frac{1}{r}e^{-\sqrt{\lambda}r}r^2\,ds\,dr, \end{aligned}$$

where S_2 is the unit 2-sphere and s is a parametrization of it. Since this problem is rotation invariant, it is sufficient to consider the case where $\mathbf{x}=(0,0,|x|)$. Then we have

$$\begin{aligned} \int_{S_2}\frac{ds}{|x-rs|}&=\int_0^\pi\int_0^{2\pi}\frac{\sin\phi\,d\theta\,d\phi}{\sqrt{r^2+|x|^2-2r|x|\cos\phi}}\\ &=2\pi\int_{-1}^{1}\frac{du}{\sqrt{r^2+|x|^2-2r|x|u}}\\ &=\begin{cases}\dfrac{4\pi}{|x|}, & |x|\geqq r,\\[2ex] \dfrac{4\pi}{r}, & |x|\leqq r.\end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} G(x)&=\frac{1}{4\pi}\left\{\int_0^{|x|}\frac{1}{|x|}re^{-\sqrt{\lambda}|x|}dr+\int_{|x|}^\infty e^{-\sqrt{\lambda}r}dr\right\}\\ &=\frac{1}{4\pi\lambda|x|}\{1-e^{-\sqrt{\lambda}|x|}\}. \end{aligned}$$

Now

$$u_j^k(x)=\frac{\delta_j^k}{4\pi|x|}e^{-\sqrt{\lambda}|x|}+\frac{\partial^2 G}{\partial x_j\partial x_k}(x),$$

so we have derived

$$u_j^k(x)=\frac{-x_jx_k}{4\pi\lambda|x|^3}\left\{\lambda e^{-\sqrt{\lambda}|x|}+\frac{3}{|x|^2}(\sqrt{\lambda}|x|e^{-\sqrt{\lambda}|x|}+e^{-\sqrt{\lambda}|x|}-1)\right\}$$

$$+\frac{\delta_j^k}{4\pi\lambda|x|}\left\{\lambda e^{-\sqrt{\lambda}|x|}+\frac{1}{|x|^2}(\sqrt{\lambda}|x|e^{-\sqrt{\lambda}|x|}+e^{-\sqrt{\lambda}|x|}-1)\right\}, \tag{3.6}$$

$$p^k(x)=\frac{x_k}{4\pi|x|^3}.$$

We now turn to a discussion of the differentiability properties of $\mathbf{v}$, ∇q. We will use an L_p multiplier theorem from Stein [1].

THEOREM 3.7. (Multiplier theorem). *Suppose $\hat{m}(\alpha)$ is of class C^k in $R^n-\{0\}$, where $k>n/2$. Assume that for all $j=(j_1,j_2,j_3)$ with $j_1+j_2+j_3\leqq k$, we have $|D^j\hat{m}(\alpha)|\cdot|\alpha|^{j_1+j_2+j_3}\leqq B$, for all $\alpha\in k$. Then the operator $T_m(f)=m*f$ is a bounded operator on $L_p(R^n)$, $1<p<\infty$, and $\|T_m\|$ depends only on B, p.*

LEMMA 3.8. *$\hat{m}(\alpha)=\alpha_j\alpha_k/(2\pi)^{3/2}|\alpha|^2$ satisfies the hypotheses of Theorem 3.7 with $k=2$.*

Proof. This is a simple calculation, for $|\alpha_j\alpha_k/(2\pi)^{3/2}|\alpha|^2|\leqq 1/(2\pi)^{3/2}$. Also,

$$\frac{\partial}{\partial\alpha_l}\left(\frac{\alpha_j\alpha_k}{|\alpha|^2}\right)=\frac{1}{(2\pi)^{3/2}}\left\{\frac{-2\alpha_j\alpha_k\alpha_l}{|\alpha|^4}+\frac{(\delta_l^j\alpha_k+\delta_l^k\alpha_j)}{|\alpha|^2}\right\},$$

so that we have

$$\left||\alpha|\frac{\partial}{\partial\alpha_l}\left(\frac{\alpha_j\alpha_k}{(2\pi)^{3/2}|\alpha|^2}\right)\right|\leqq\frac{4}{(2\pi)^{3/2}}.$$

Similarly,

$$\frac{\partial^2}{\partial\alpha_s\partial\alpha_l}\left(\frac{\alpha_j\alpha_k}{(2\pi)^{3/2}|\alpha|^2}\right)$$

$$=\frac{1}{(2\pi)^{3/2}}\left\{\frac{8\alpha_j\alpha_k\alpha_l\alpha_s}{|\alpha|^6}-\frac{2(\delta_j^s\alpha_k\alpha_l+\delta_k^s\alpha_j\alpha_l+\delta_l^s\alpha_j\alpha_k+\delta_l^i\alpha_k\alpha_s+\delta_l^k\alpha_j\alpha_s)}{|\alpha|^4}\right.$$

$$\left.+\frac{(\delta_l^j\delta_s^k+\delta_l^k\delta_s^j)}{|\alpha|^2}\right\},$$

so,

$$\left||\alpha|^2\frac{\partial^2}{\partial\alpha_s\partial\alpha_s}\left(\frac{\alpha_j\alpha_k}{(2\pi)^{3/2}|\alpha|^2}\right)\right|\leqq\frac{20}{(2\pi)^{3/2}}.$$

LEMMA 3.9.

$$\hat{m}(\alpha)=\frac{1}{\lambda+|\alpha|^2}\quad or\quad\frac{\alpha_j}{\lambda+|\alpha|^2}\quad or\quad\frac{\alpha_j\alpha_k}{\lambda+|\alpha|^2}$$

satisfies the conditions of Theorem 3.7 with $k=2$ as long as $\lambda\not\leqq 0$.

Proof. This is a routine calculation and we shall write out one case only, the case of $\hat{m}(\alpha) = 1/(\lambda + |\alpha|^2)$. Let $\lambda = e^{i\theta}$, and note that $|\hat{m}(\alpha_{|\alpha|\to\infty})| \to 0$. Thus, to show that it is bounded, it is enough to show that the denominator cannot equal zero. But

$$\frac{1}{|\lambda + |\alpha|^2|} = \frac{1}{\sqrt{(\cos\theta + |\alpha|^2)^2 + \sin^2\theta}},$$

if $\lambda = e^{i\theta}$. The denominator cannot be zero unless $\theta = \pi \pmod{2\pi}$, which is the excluded case. Now, we also have

$$\frac{\partial}{\partial\alpha_l}\left(\frac{1}{\lambda + |\alpha|^2}\right) = \frac{-2\alpha_l}{(\lambda + |\alpha|^2)^2},$$

so

$$|\alpha|\left|\frac{\partial}{\partial\alpha_l}\left(\frac{1}{\lambda + |\alpha|^2}\right)\right| \leqq \frac{2|\alpha|^2}{(\lambda + |\alpha|^2)^2},$$

which is bounded by the above reasoning. Finally,

$$\frac{\partial^2}{\partial\alpha_s\partial\alpha_l}\left(\frac{1}{\lambda + |\alpha|^2}\right) = \frac{8\alpha_l\alpha_s}{(\lambda + |\alpha|^2)^3} - \frac{2\delta_s^l}{(\lambda + |\alpha|^2)^2},$$

so that

$$\left||\alpha|^2\frac{\partial^2}{\partial\alpha_s\alpha_l}\left(\frac{1}{\lambda + |\alpha|^2}\right)\right| \leqq \frac{8|\alpha|^3}{(\lambda + |\alpha|^2)^3} + \frac{2|\alpha|^2}{(\lambda + |\alpha|^2)^2},$$

which is also bounded.

We now easily see that $v_k(x) \in W^{2,p}(R^3)$ and $\partial q/\partial x_k \in L_p(R^3)$. In fact,

THEOREM 3.10. *If* $v_k(x) = \sum_{j=1}^3 (u_j^k * f_j)(x)$ *and* $h_k(x) = \sum_{j=1}^3 (\partial p^k/\partial x_j * f_j)(x)$, *where* $\mathbf{f} \in L_p(R^3)$, $1 < p < \infty$, *then* $\mathbf{v} \in W^{2,p}(R^3) \cap J_p(R^3)$ *and* $\mathbf{h} = \nabla q \in G_p(R^3)$. *Furthermore,* $\mathbf{v}$, q *satisfy* (3.1) *and* $\|\mathbf{v}\|_{W^{2,p}(R^3)} \leqq C_p\|\mathbf{f}\|_{L_p(R^3)}$ *and* $\|\nabla q\|_{L_p(R^3)} \leqq C_p\|\mathbf{f}\|_{L_p(R^3)}$.

Proof. The previous lemmas show that $\hat{u}_j^k$, $i\alpha_l\hat{u}_j^k$, $-\alpha_s\alpha_l\hat{u}_j^k$ and $i\alpha_j\hat{p}^k$ are all L_p-multipliers. Thus $\mathbf{v} \in W^{2,p}(R^3)$ and $\mathbf{h} \in L_p(R^3)$ with estimates given above. Since (3.4) hold, we also have $\mathbf{v}, \nabla\mathbf{v} \in J_p(R^3)$ and (3.1) hold (that is, (3.1) hold with h_j in place of $\alpha g/\alpha x_j$). Thus, to complete the proof, we need only show that $h = \nabla q \in G_p(R^3)$. Let

$$g(x) = \int_{c_1}^{x_1} h_1(t_1, c_2, c_3)\,dt_1 + \int_{c_2}^{x_2} h_2(x_1, t_2, c_3)\,dt_2 + \int_{c_3}^{x_3} h_3(x_1, x_2, t_3)\,dt_3,$$

where c_2, c_3 are chosen so that the first two integrals exist a.e. The function q is easily seen to be locally L_p, for we have that

$$\left|\int_{c_3}^{x_3} h_3(x_1, x_2, t_3)\,dt_3\right|^p \leqq |x_3 - c_3|^{p/q}\int_{c_3}^{x_3} |h_3(x_1, x_2, t_3)|^p\,dt_3,$$

where $(1/p) + (1/q) = 1$, so that

$$\int_{a_3}^{b_3}\int_{a_2}^{b_2}\int_{a_1}^{b_1}\left|\int_{c_3}^{x_3}(h_3(x_1, x_2, t_3)\,dt_3\right|^p dx_1\,dx_2\,dx_3 \leqq \|h\|_p^p\int_{a_3}^{b_3}|x_3 - c_3|^{p/q}\,dx_3$$
$$\leqq C_p\|h\|_p^p.$$

We have chosen c_2, c_3 so that the sections $h_1(x_1, t_2, c_3)$ and $h_2(t_1, c_2, c_3)$ are both in L_p;

$$\int_{a_3}^{b_3}\int_{a_2}^{b_2}\int_{a_1}^{b_1}\left|\int_{c_2}^{x_2}|h_2(x_1, t_2, c_3)|\,dt_2\right|^p$$

$$\leqq \int_{a_3}^{b_3}\int_{a_2}^{b_2}\int_{a_1}^{b_1}|x_2-c_2|^{p/q}\int_{-\infty}^{+\infty}|h_2(x_1, t_2, c_3)|^p\,dt_2\,dx_1\,dx_2\,dx_3$$

$$\leqq \int_{a_3}^{b_3}\int_{a_2}^{b_2}|x_2-c_2|^{p/q}\left(\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|h_2(x_1, t_2, c_3)|^p\,dt_2\,dx_1\right)dx_2\,dx_3$$

$$\leqq C_p\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|h_2(x_1, t_2, c_3)|^p\,dt_2\,dx_1.$$

A similar estimate shows that $\int_{c_1}^{x_1} h_1(t_1, c_2, c_3)\,dt_1$ is locally L_p, too. Now consider ∇q; we have $(\partial q/\partial x_3)(x) = h_3(x)$. Note that $\partial h_2/\partial x_3 = \partial h_3/\partial x_2$ in the distributional sense, so $(\partial q/\partial x_2)(x) = h_2(x)$. Similarly, $(\partial q/\partial x_1)(x) = h_1(x)$.

It should be remarked here that the techniques employed in this section yield a stronger result than Theorem 3.10. Suppose $\mathbf{f}\in W^{s,p}(R^3)$. Employing integration by parts and the previous results, we easily see that the following theorem holds.

THEOREM 3.11. *If* $\mathbf{v}$, ∇q *and* $\mathbf{f}$ *are as in Theorem* 3.10 *and if* $\mathbf{f}\in W^{s,p}(R^3)$, *then* $\mathbf{v}\in W^{s+2,p}(R^3)$ *and* $\nabla q\in W^{s,p}(R^3)$ *and we have the estimates*

$$\|\mathbf{v}\|_{W^{s+2,p}(R^3)} \leqq C_p\|\mathbf{f}\|_{W^{s,p}(R^3)},$$

$$\|\nabla q\|_{W^{s,p}(R^3)} \leqq C_p\|\mathbf{f}\|_{W^{s,p}(R^3)}.$$

Proof. Notice that $D^l v_k = \sum_{j=1}^3 D^l(u_j^k * f_j) = \sum_{j=1}^3 u_j^k * D^l f_j$ and similarly for $D^l\nabla q$. Then apply Theorem 3.10 using $D^l\mathbf{v}$, $\nabla D^l q$, and $D^l\mathbf{f}$ in place of $\mathbf{v}$, ∇q, and $\mathbf{f}$, respectively.

The Green's formulas on halfspace. Let $\mathbf{u}$, $\mathbf{v}$ be smooth, divergence-free vector fields and p, q be smooth functions. We compute the Green's formulas for the Stokes equations by letting

$$T_{ij}(\mathbf{u}, p) = -\delta_{ij}^i p + \left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right),$$

$$T'_{ij}(\mathbf{u}, p) = \delta_{ij}^i p + \left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right). \tag{4.1}$$

It follows that

$$\frac{\partial(T_{ij}(\mathbf{u}, p)v_i)}{\partial x_j} = \frac{1}{2}\left(\frac{\nabla u_i}{\partial x_j}+\frac{\nabla u_j}{\partial x_i}\right)\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}\right) + \left(\Delta u_i - \frac{\partial p}{\partial x_i}\right)v_i. \tag{4.2}$$

Using the divergence theorem, we find that

$$\int_H\left\{\left(\lambda v_i(y) - \Delta v_i(y) + \frac{\partial q}{\partial y_i}(y)\right)u_i(y) - v_i(y)\left(\lambda u_i(y) - \Delta u_i(y) - \frac{\partial p}{\partial y_i}(y)\right)\right\}dy$$

$$= \int_{y_3=0}[T'_{i3}(\mathbf{u}, p)v_i(y) - T_{i3}(\mathbf{v}, q)u_i(y)]\,dy_1\,dy_2, \tag{4.3}$$

as long as

$$\int_{S_r(0)}[T'_{ij}(\mathbf{u}, p)v_i(y)N_j(y) - T_{ij}(\mathbf{v}, q)u_i(y)N_j(y)]\,dy \to 0, \qquad r\to\infty,$$

where $S_r(0)$ is the intersection of a smooth surface in $\{x \in R^3 \,|\, |x| > r\}$ with H, and N is the exterior unit normal to $S_r(0)$.

Therefore, assuming that integration by parts is valid, and letting $\mathbf{u}^k$, p^k be the Green's functions and $\mathbf{v}$, q, $\mathbf{f}$ satisfy $\lambda \mathbf{v} - \Delta \mathbf{v} + \nabla q = \mathbf{f}$ in H, we have

$$\begin{aligned} &\int_H u_i^k(x-y) f_i(y)\, dy + \int_{y_3=0} T_{i3}(\mathbf{v}, q) u_i^k(x-y)\, dy \\ &\qquad - \int_{y_3=0} T'_{i3}(\mathbf{u}^k, p^k)(x-y) v_i(y)\, dy \\ &\qquad = \begin{cases} v_k(x), & x \in H, \\ 0, & x \notin H, \end{cases} \end{aligned} \tag{4.4}$$

where all differentiation is done with respect to y. We now develop a corresponding formula for q. Since

$$\begin{aligned} \Delta_x T'_{ij}(\mathbf{u}^k(x-y)) &= \Delta_x \left\{ \delta_j^i p^k(x-y) + \frac{\partial u_i^k}{\partial y_j}(x-y) + \frac{\partial u_j^k}{\partial y_i}(x-y) \right\} \\ &= \delta_j^i \Delta_x p^k(x-y) - \left(\frac{\partial}{\partial x_j} \Delta_x u_i^k(x-y) + \frac{\partial}{\partial x_i} \Delta_x u_j^k(x-y) \right) \\ &= \left(\frac{\partial}{\partial x_j} \Delta_x u_i^k(x-y) + \frac{\partial}{\partial x_i} \Delta_x u_j^k(x-y) \right) \quad \text{for } x \neq y, \end{aligned}$$

we see that

$$\frac{\partial}{\partial x_j} \Delta_x u_i^k(x-y) = \lambda \frac{\partial u_i^k}{\partial x_j}(x-y) - \frac{\partial^2 p^k}{\partial x_j\, \partial x_i}(x-y) \quad \text{if } x \neq y.$$

Thus,

$$\Delta_x T'_{ij}(\mathbf{u}^k(x-y)) = -\left(2 \frac{\partial^2 p^k}{\partial x_j\, \partial x_i}(x-y) \right) + \lambda \left(\frac{\partial u_i^k}{\partial x_j}(x-y) + \frac{\partial u_j^k}{\partial x_i}(x-y) \right) \quad \text{if } x \neq y.$$

Hence, we have

$$-\lambda T'_{ij}(u^k(x-y)) + \Delta_x T'_{ij}(u^k(x-y)) = -2 \frac{\partial^2 p^k}{\partial x_i\, \partial x_j}(x-y) - \lambda \delta_j^i p^k(x-y) \quad \text{if } x \neq y.$$

This, together with (4.4) and the fact that $\lambda \mathbf{v} - \Delta \mathbf{v} + \nabla q = \mathbf{f}$, yields

$$\begin{aligned} q(x) = &\int_H p^i(x-y) f_i(y)\, dy + \int_{y_3=0} p^i(x-y) T_{i3}(\mathbf{v}, q)\, dy \\ &+ 2 \int_{y_3=0} \frac{\partial p^i}{\partial x_3}(x-y) v_i(y)\, dy - \frac{\lambda}{4\pi} \int_{y_3=0} \frac{v_3(y)}{|x-y|}\, dy + C. \end{aligned} \tag{4.5}$$

This leads us to define the potentials of a double layer.

DEFINITION 4.6. By the potentials of a double layer with density $\boldsymbol{\phi}$, we shall mean the integrals

$$\begin{aligned} w_k(x, \boldsymbol{\phi}) &= -\int_{y_3=0} T'_{i3}(\mathbf{u}^k(x-y)) \phi_i(y)\, dy, \\ \psi(x, \boldsymbol{\phi}) &= 2 \int_{y_3=0} \frac{\partial p^i}{\partial x_3}(x-y) \phi_i(y)\, dy - \frac{\lambda}{4\pi} \int_{y_3=0} \frac{\phi_3(y)}{|x-y|}\, dy. \end{aligned}$$

For densities ϕ for which we can justify differentiation under the integral sign, we will have

$$\lambda \mathbf{w} - \Delta \mathbf{w} + \nabla \psi = 0, \qquad \nabla \cdot \mathbf{w} = 0 \tag{4.7}$$

in both the upper and lower halfspaces. Before discussing the behavior of $\mathbf{w}$, we calculate $T'_{i3}(\mathbf{u}^k(x-y))$, and find that

$$\begin{aligned}
&T'_{ij}(\mathbf{u}^k|x-y|) \\
&= \frac{-(x_j-y_j)(x_k-y_k)(x_i-y_i)}{2\pi|x-y|^5}\Big\{\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|}+6e^{-\sqrt{\lambda}|x-y|} \\
&\qquad + \frac{15(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|}+e^{-\sqrt{\lambda}|x-y|}-1)}{|x-y|^2}\Big\} \\
&+ \frac{\delta^i_k(x_j-y_j)+\delta^k_j(x_i-y_i)}{4\pi|x-y|^3}\Big\{\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} \\
&\qquad + 3e^{-\sqrt{\lambda}|x-y|}+\frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|}+e^{-\sqrt{\lambda}|x-y|}-1)}{|x-y|^2}\Big\} \\
&+ \frac{\delta^i_j(x_k-y_k)}{4\pi|x-y|^3}\Big\{1+2e^{-\sqrt{\lambda}|x-y|}+\frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|}+e^{-\sqrt{\lambda}|x-y|}-1)}{\lambda|x-y|^2}\Big\}.
\end{aligned} \tag{4.8}$$

The first term has a singularity of the order of $|x-y|^2$ as $x \to y$; the other two terms are bounded as $x \to y$. As $y \to \infty$ for fixed x, the terms have orders $|y|^{-4}$, $|y|^{-4}$, and $|y|^{-2}$, respectively. Hence, we get the following theorem.

THEOREM 4.9. *Let $\phi \in L_p(R^3)$. Then the integral defining $w_k(x, \phi)$ converges absolutely and locally uniformly in each of the three regions $H^+ = \{x \in R^3 | x_3 > 0\}$, $\partial H = \{x \in R^3 | x_3 = 0\}$, and H. Hence, $w_k(x, \phi$ is continuous in each of these regions.*

Proof. For fixed $x \in H^+$ or H^+ or H, the integrand is L_1, because $T_{i3} \in L_q$, where $(1/p)+(1/q)=1$. On the other hand, if $x_3 = 0$, then the first term of $T'_{i3} = 0$; hence $T'_{i3}|_{x_3=0}$ is again in L_q. Therefore, the integral defining $w_k(x, \phi)$ converges absolutely in each of the three regions. To show that the integral converges locally uniformly in x in H and H^+, consider x in any bounded subset of H or H^+. Clearly, the integral of $|T'_{i3}(\mathbf{u}^k(x-y))\phi_i(y)|$ over the exterior of a large ball goes to zero uniformly for x in the bounded set. On the interior of the ball $T'_{i3}(\mathbf{u}^k(x-y))$ is continuous if $x \in H$ or H^+. By almost the same reasoning $w_k(x, \phi)$ is continuous on ∂H. The only difference is that the integrand is not continuous at $x = y$. It is, however, bounded, so if $x_1, x_2 \in \Omega$, a bounded set, then

$$\begin{aligned}
&\int_\Omega |T'_{i3}(\mathbf{u}^k(x_1-y)) - T'_{i3}(\mathbf{u}^k(x_2-y))| \cdot |\phi(y)|\, dy \\
&\leqq \Big(\int_\Omega |T'_{i3}(\mathbf{u}^k(x_1-y))| - T'_{i3}(\mathbf{u}^k(x_2-y))|^q\, dy\Big)^{1/q} \|\phi\|_p^p.
\end{aligned}$$

$T'_{i3}(\mathbf{u}^k(z))$ is continuous as a function of z, if $z \neq 0$. Hence, given A, $\varepsilon > 0$ there exists $\delta > 0$ such that if $\varepsilon < |z_i| < A$ and $|z_1 - z_2| < \delta$, then $|T'_{i3}(\mathbf{u}^k(z_1)) - T'_{i3}(\mathbf{u}^k(z_2))|^q < \varepsilon$. Let A be the diameter of Ω and let ε be small enough so that the ball $B_{2\varepsilon}(x_1)$ of radius 2ε

about x_1 is in Ω. Let $|x_2-x_1|<\min\{\varepsilon, \delta\}$. Then,

$$\begin{aligned}\int_\Omega & |T'_{i3}(\mathbf{u}_k(x_1-y))-T'_{i3}(\mathbf{u}^k(x_2-y))|^q\,dy \\ &= \int_{\Omega-B_{2\varepsilon}} |T'_{i3}(\mathbf{u}^k(x_1-y))-T'_{i3}(\mathbf{u}^k(x_2-y))|^q\,dy \\ &\quad + \int_{B_{2\varepsilon}} |T_{i3}(\mathbf{u}^k(x_1-y))-T_{i3}(\mathbf{u}^k(x_2-y))|^q\,dy \\ &\leqq \varepsilon C_1 + \pi(2\varepsilon)^2 C_2,\end{aligned}$$

where C_1 is the volume of Ω and C_2 is twice the max of $|T'_{i3}(\mathbf{u}^k(z))|^q$. Hence, $w_k(x, \boldsymbol{\phi})$ is continuous at x_1 for all $x_1 \in \Omega$.

Of course, $w_k(x, \boldsymbol{\phi})$ has the usual sort of jump across the x_1, x_2-plane. We prove this below.

THEOREM 4.10. *Let* $\mathbf{c}=(c_1, c_2, c_3)$ *be any constant vector. Then*

$$w_k(x, \mathbf{c}) = \begin{cases} c_k, & x \in H, \\ 0, & x \in H^+, \\ c_k/2, & x \in \partial H, \end{cases}$$

where we define $w_k(x, \mathbf{c})$ *by integrating over spheres centered at* $(x_1, x_2, 0)$.

Proof. If we can justify integration by parts, letting $\mathbf{v}=\mathbf{c}$ and $q=0$ and using (4.4), we get

$$\lambda \int_H u_i^k(x-y)c_i\,dy + w_k(x, \mathbf{c}) = \begin{cases} c_k, & x \in H, \\ 0, & x \in H^+. \end{cases}$$

Note that $\int_H u_i^k(x-y)\,dy = 0$, where integration is over spheres centered at $(x_1, x_2, 0)$, since it can easily be seen that the integral of $u_i^k(z)$ over the intersection of H with any sphere centered at the origin is zero. Therefore, when integration by parts has been justified, we will have the first two parts of the theorem. To justify integration by parts, consider $\int_{\nabla B_r(x_1, x_2, 0)\cap H} T'_{ij}(\mathbf{u}^k(x-y))c_i N_j(y)\,dy$, where $\mathbf{N}$ is the exterior unit normal to the sphere. The first two terms in the integrand go to zero as $1/r^4$ as $r \to \infty$. Since the region of integration is two-dimensional, their intergral goes to zero as $r \to \infty$. Consider the final term. We divide it into $\int_{\partial B_r(x_1, x_2, 0)\cap H}[(x_k-y_k)/4\pi|x-y|^3]N_i(y)c_i\,dy$ plus another term to which the previous argument applies. By calculating all the possibilities, we see that the integral above goes to zero as $r \to \infty$. To prove that $w_k(x, c)=c_k/2$ if $x_3=0$, let H_ε be H with a ball of radius ε about x removed. From (4.2) we construct the analogue to (4.4) for this region. In exactly the same way we can justify integration by parts for $\boldsymbol{\phi}(y)=\mathbf{c}$ for this region. Thus, since $x \notin H_\varepsilon$, we have $-\int_{\partial H_\varepsilon} T'_{ij}(\mathbf{u}^k(x-y))c_i N_j(y)\,dy = 0$. Note that $\partial H_\varepsilon = (\partial H \cap \partial H_\varepsilon) \cup S_\varepsilon$, where S_ε is half of the sphere of radius ε centered at x. Taking limits as $\varepsilon \to 0$, we see that

$$-\int_{\partial H} T_{i3}(\mathbf{u}^k(x-y))c_i\,dy = \lim_{\varepsilon \to 0} \int_{S\varepsilon} T'_{ij}(\mathbf{u}^k(x-y))c_i N_j(y)\,dy = \frac{c_k}{2},$$

as can be seen by direct calculation.

We now use Theorem 4.10 to calculate the jump in $w(x, \boldsymbol{\phi})$ across ∂H.

THEOREM 4.11. *Let* $\phi \in L_p(R^2) \cap C(R^2)$. *Let* $x^0 \in R^2$. *Then*

$$\lim_{\substack{x \to X \\ x \in H}} w_k(x, \phi) = \frac{\phi_k(x^0)}{2} + w_k(x^0, \phi),$$

$$\lim_{\substack{x \to x^0 \\ x \in H^+}} w_k(x, \mathbf{f}) = \frac{-\phi_k(x_0)}{2} + w_k(x^0, \phi).$$

Proof. Consider the function $w_k(x, \phi) + \int_{\partial H} T'_{i3}(\mathbf{u}^k(x-y))\phi_i(x^0)\, dy$, where $x \in H$. We shall show that this function is continuous at x^0. Now,

$$w_k(x, \phi) + \int_{\partial H} T'_{i3}(\mathbf{u}^k(x-y))\phi_i(x^0)\, dy - w_k(x^0, \phi)$$

$$- \int_{\partial H} T_{i3}(\mathbf{u}^k(x^0-y))\phi_i(x^0)\, dy$$

$$= \int_{\partial H} [T'_{i3}(\mathbf{u}^k(x-y)) - T'_{i3}(\mathbf{u}^k(x^0-y))](\phi_i(x^0)\phi_i(y))\, dy.$$

Choose $\delta > 0$ so that if $|x^0 - y| < \delta$, then $|\phi_i(x^0) - \phi_i(y)| < \varepsilon$. Then

$$\int_{B_\delta(x^0)} |T'_{i3}(\mathbf{u}^k(x-y)) - T'_{i3}(\mathbf{u}^k(x^0-y))|(\phi_i(x^0) - \phi_i(y))\, dy$$

$$\leqq \varepsilon \int_{B_\delta(x^0)} |T'_{i3}(\mathbf{u}^k(x-y)) - T'_{i3}(\mathbf{u}^k(x^0-y))|\, dy$$

$$\leqq \varepsilon \left(\int_{B_\delta(x^0)} |T'_{i3}(\mathbf{u}^k(x-y))|\, dy + C \right),$$

since $T'_{i3}(\mathbf{u}^k(x^0-y))$ is bounded because $x^0 \in \partial H$. Since $y \in \partial H$, $x \in \partial H$, $T'_{i3}(\mathbf{u}^k(x-y))$ has a singularity as $x \to y$ like $|x_3|/|x-y|^3$. A computation now shows that $\int_{B_\delta(x_0)} |T'_{i3}(\mathbf{u}^k(x-y))|\, dy$ is bounded, independent of δ and x.

Now consider

$$\int_{\partial H - B_\delta(x^0)} \{T'_{i3}(\mathbf{u}^k(x-y)) - T'_{i3}(\mathbf{u}^k(x^0-y))\}(\phi_i(x^0) - \phi_i(y))\, dy.$$

The argument used in Theorem 4.9 shows that half of it, namely, $-\int_{\partial H - B_\delta(x^0)} T'_{i3}(\mathbf{u}^k(x-y)\phi_i(y)\, dy$, is continuous at $x = x^0$. On the other hand, $\int_{\partial H - B_\delta(x^0)} T'_{i3}(\mathbf{u}(x-y))\phi_i(x^0)\, dy$ is also continuous at $x = x^0$. The argument of Theorem 4.9 applies to all the terms in the integrand except the one which goes to zero only as $|y|^{-2}$ as $y \to \infty$. That term is

$$\frac{\delta^i_3(x_k - y_k)}{4\pi|x-y|^3}\phi_i(x^0) = \frac{(x_k - y_k)\phi_3(x^0)}{4\pi|x-y|^3}.$$

Therefore, consider

$$\int_{\partial H \cap B_\delta(x^0)} \frac{(x_k - y_k)\phi_3(x^0)}{4\pi|x-y|^3}\, dy = 0 \quad \text{if } k = 1, 2.$$

On the other hand, if $k=3$, then we have

$$\left|\int_{\partial H\cap B_\delta(x^0)} \frac{x_3\phi_3(x^0)}{(x_1-y_1)^2+(x_2-y_2)^2+x_3^2}\,dy_1\,dy_2\right|$$
$$\leqq |x_3||\phi_3(x^0)|\int_{\delta/2}^{\infty}\frac{r\,dr}{(r^2+x_3^2)^{3/2}}$$
$$=|\phi_3(x^0)|\int_{\delta/2|x_3|}^{\infty}\frac{u\,du}{(u^2+1)^{3/2}},$$

as long as x is close enough to x^0. Since this goes to zero as $x\to x^0$ because then $x_3\to 0$, we have established that $\int_{\partial H\cap B_\delta(x^0)} T'_{i3}(\mathbf{u}(x-y))\phi_i(x^0)\,dy$ is continuous at $x=x_0$, and hence that $w_k(x,\mathbf{f})+\int_{\partial H} T'_{i3}(\mathbf{u}^k(x-y))\phi_i(x^0)\,dy$ is continuous at $x=x_0$. Hence,

$$\lim_{\substack{x\to x^0\\ x\in H}} w_k(x,\boldsymbol{\phi}) = w_k(x^0,\boldsymbol{\phi})+\int_{\partial H} T'_{i3}(\mathbf{u}^k(x^0-y))\phi_i(x^0)\,dy$$
$$-\lim_{\substack{x\to x^0\\ x\in H}}\int_{\partial H} T'_{i3}(\mathbf{u}^k(x-y))\phi_i(x^0)\,dy$$
$$= w_k(x^0,\boldsymbol{\phi})-\frac{\phi_k(x_0)}{2}+\phi_k(x^0)$$

by Theorem 4.10. This shows that

$$\lim_{\substack{x\to x^0\\ x\in H}} w_k(x,\boldsymbol{\phi}) = \frac{\phi_k(x^0)}{2}+w_k(x^0,\boldsymbol{\phi}).$$

The second part of the theorem is proved exactly the same way.

5. The integral equation. In § 3 we solved the free space resolvent problem for the Stokes equations; i.e., given $\mathbf{f}$ find $\mathbf{v}$, q so that $\lambda\mathbf{v}-\Delta\mathbf{v}+\nabla q=\mathbf{f}$, $\nabla\cdot\mathbf{v}=0$. In this section and the next, we shall seek a solution to the boundary value problem $\lambda\mathbf{w}-\Delta\mathbf{w}+\nabla\psi=0$, $\nabla\cdot\mathbf{w}=0$ in H and $\mathbf{w}|_{\partial H}=-\mathbf{v}|_{\partial H}$. The solution to (2.1) will then be $\mathbf{u}=\mathbf{v}+\mathbf{w}$, $p=q+\psi$. In this section, we solve the boundary value problem, and in the next section we derive estimates for the solution.

We seek a solution to

$$\left.\begin{aligned}\lambda\mathbf{w}-\Delta\mathbf{w}+\nabla\psi&=0\\ \nabla\cdot\mathbf{w}&=0\end{aligned}\right\}\ \text{in } H,\qquad |\mathbf{w}|_{\partial H}=-\mathbf{v}|_{\partial H}, \tag{5.1}$$

in the form $\mathbf{w}=\mathbf{w}(x,\boldsymbol{\phi})$, $\psi=\psi(x,\mathbf{f})$, the potentials of a double layer with density $\boldsymbol{\phi}$. Because of (4.7), we are in fact seeking $\boldsymbol{\phi}$ so that $\lim_{x\to x^0, x\in H}\mathbf{w}(x,\boldsymbol{\phi})=-\lim_{x\to x^0,x\in H}\mathbf{v}(x)$. Letting $\mathbf{a}(x)$ be a mapping from $R^2\to R^3$, we shall solve

$$\mathbf{a}(x)=\frac{\boldsymbol{\phi}(x)}{2}+\mathbf{w}(x,\boldsymbol{\phi})\quad\text{where } x\in R^2. \tag{5.2}$$

In this section, we will show that, given $\mathbf{a}\in L_p(R^2)$, then the solution $\boldsymbol{\phi}$ is unique, $\boldsymbol{\phi}\in L_p(R^2)$ and the map from $\mathbf{a}$ and $\boldsymbol{\phi}$ is bounded. Using (4.8) and the fact that

$x_3 = y_3 = 0$, we see that

$$\frac{\phi_k(x)}{2} + w_k(x, \boldsymbol{\phi})$$

$$= \frac{\phi_k(x)}{2} - \frac{1}{4\pi}\int\int_{R^2} \left\{ \delta_3^k \frac{(x_i - y_i)}{|x-y|^3}\left(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} + 3e^{-\sqrt{\lambda}|x-y|} + \frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} - 1)}{\lambda|x-y|^2}\right)\right.$$

$$\left. + \delta_3^i \frac{(x_k - y_k)\phi_i(y)}{|x-y|^3}\left(1 + 2e^{-\sqrt{\lambda}|x-y|} + \frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} + e^{-\sqrt{\lambda}|x-y|} - 1)}{\lambda|x-y|^2}\right)\right\} \phi_i(y)\, dy_1\, dy_2.$$

That is, regarding $\mathbf{w}(\cdot, \boldsymbol{\phi})$ as an operator on $\boldsymbol{\phi}$, we have

$$(\tfrac{1}{2} + \mathbf{w})\boldsymbol{\phi} = \begin{bmatrix} \frac{1}{2} & 0 & \underline{\bar{X}}_1 \\ 0 & \frac{1}{2} & \underline{\bar{X}}_2 \\ \underline{\bar{Y}}_1 & \underline{\bar{Y}}_2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix}, \tag{5.3}$$

where

$$(\underline{\bar{X}}_j \phi_i)(x)$$

$$= -\frac{1}{4\pi}\int\int_{R^2} \frac{(x_j - y_j)\phi_i(y)}{|x-y|^3}\left(1 + 2e^{-\sqrt{\lambda}|x-y|} + \frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} + e^{-\sqrt{\lambda}|x-y|} - 1)}{\lambda|x-y|^2}\right) dy_1\, dy_2$$

and

$$(\underline{\bar{Y}}_j \phi_i)(x)$$

$$= -\frac{1}{4\pi}\int\int_{r^2} \frac{(x_j - y_j)\phi_i(y)}{|x-y|^3}\left(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} + 3e^{-\sqrt{\lambda}|x-y|} + \frac{6(\sqrt{\lambda}|x-y|e^{-\sqrt{\lambda}|x-y|} + e^{-\sqrt{\lambda}|x-y|} - 1)}{\lambda|x-y|^2}\right) dy.$$

Hence,

$$(\tfrac{1}{2} + \mathbf{w})\boldsymbol{\phi} = \begin{bmatrix} \frac{1}{2} & 0 & 2\pi\hat{\underline{\bar{X}}}_1 \\ 0 & \frac{1}{2} & 2\pi\hat{\underline{\bar{X}}}_2 \\ 2\pi\hat{\underline{\bar{Y}}}_1 & 2\pi\hat{\underline{\bar{Y}}}_2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{bmatrix}. \tag{5.4}$$

First of all, we compute $\hat{\underline{\bar{X}}}_j$, $\hat{\underline{\bar{Y}}}_j$.

$$\hat{\underline{\bar{Y}}}_1(\alpha) = \frac{-1}{8\pi^2\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\alpha \cdot z}\frac{z_1}{|z|^2}\left\{\lambda^{3/2}|z|^3 e^{-\sqrt{\lambda}|z|} + 3\lambda|z|^2 e^{-\sqrt{\lambda}|z|} + 6\sqrt{\lambda}|z|\, e^{-\sqrt{\lambda}|z|} + 6e^{-\sqrt{\lambda}|z|} - 6\right\} dz_1\, dz_2$$

$$= \frac{\lambda}{24\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{z_1}{|z|}e^{-\sqrt{\lambda}|z|-i\alpha\cdot z}\,dz_1\,dz_2$$

$$+\frac{i\alpha_1}{24\pi^2\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{e^{-i\alpha\cdot z}}{|z|^3}\Big\{\lambda^{3/2}|z|^3e^{-\sqrt{\lambda}|z|}+3\lambda|z|^2e^{-\sqrt{\lambda}|z|}$$

$$+6\sqrt{\lambda}|z|\,e^{-\sqrt{\lambda}|z|}+6e^{-\sqrt{\lambda}|z|}-6\Big\}\,dz_1\,dz_2$$

$$= \frac{\lambda}{24\pi^2}\int_0^{2\pi}\frac{\cos\theta\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}$$

$$+\frac{i\alpha_1}{24\pi^2\lambda}\int_0^{2\pi}d\theta\int_0^{\infty}\frac{1}{r^2}(\lambda^{3/2}r^3\,e^{-\sqrt{\lambda}r}+3\lambda r^2\,e^{-\sqrt{\lambda}r}+6\sqrt{\lambda}\,re^{-\sqrt{\lambda}r}$$

$$+6e^{-\sqrt{\lambda}r}-6)\,e^{-i\alpha_1 r\cos\theta-i\alpha_2 r\sin\theta}\,dr$$

$$= \frac{\lambda}{24\pi^2}\int_0^{2\pi}\frac{\cos\theta\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}+\frac{i\alpha_1\sqrt{\lambda}}{24\pi^2}\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}$$

$$+\frac{i\alpha_1}{8\pi^2}\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}-\frac{i\alpha_1}{4\pi^2}\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}$$

$$-\frac{i\alpha_1}{4\pi^2\lambda}\int_0^{2\pi}(i\alpha_1\cos\theta+i\alpha_2\sin\theta)\,d\theta\cdot\int_0^{\infty}e^{i\alpha_1\cos\theta-i\alpha_2\sin\theta}\left(\frac{e^{-\sqrt{\lambda}r}-1}{r}\right)dr$$

$$= \frac{\lambda}{24\pi^2}\int_0^{2\pi}\frac{\cos\theta\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}+\frac{i\alpha_1}{24\pi^2}\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}$$

$$-\frac{i\alpha_1}{8\pi^2}\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}-\frac{i\alpha_1}{4\pi^2\sqrt{\lambda}}\int_0^{2\pi}\frac{i\alpha_1\cos\theta+i\alpha_2\sin\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}\,d\theta$$

$$-\frac{i\alpha_1}{4\pi^2\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{(i\alpha_1z_1+i\alpha_2z_2)}{|z|^3}(e^{-\sqrt{\lambda}|z|}-1)\,e^{-1\alpha\cdot z}\,dz.$$

The first four integrals can easily be calculated using residues. The results are as follows:

$$\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}=\frac{2\pi}{\sqrt{\lambda+|\alpha|^2}},$$

$$\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}=\frac{2\pi\sqrt{\lambda}}{(\lambda+|\alpha|^2)^{3/2}},$$

$$\int_0^{2\pi}\frac{\cos\theta\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)^2}=\frac{-2\pi i\alpha}{(\lambda+|\alpha|^2)^{3/2}},$$

$$\int_0^{2\pi}\frac{(i\alpha_1\cos\theta+i\alpha_2\sin\theta)\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}=\frac{2\pi(\sqrt{\lambda+|\alpha|^2}-\sqrt{\lambda})}{\sqrt{\lambda+|\alpha|^2}}.$$

We turn now to the last integral, which is

$$\frac{-i\alpha_1}{4\pi^2\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{(i\alpha_1 z_1+i\alpha_2 z_2)}{|z|^3}(e^{-\sqrt{\lambda}|z|}-1)\,e^{-i\alpha\cdot z}\,dz.$$

Note that

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{z_j}{|z|^3}e^{-i\alpha\cdot z}\,dz_1\,dz_2=\frac{-i\alpha_j}{|\alpha|},$$

since it is the Fourier transform of the Riesz kernel. Now consider

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{z_i}{|z|^3}e^{-\sqrt{\lambda}|z|}\,e^{-i\alpha\cdot z}\,dz_1\,dz_2\\
&\quad=-\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{z_1}{|z|^2}\sqrt{\lambda}\,e^{-\sqrt{\lambda}|z|-i\alpha\cdot z}\,dz-i\alpha_1\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{e^{-\sqrt{\lambda}}|z|}{|z|}e^{-i\alpha\cdot z}\,dz\\
&\quad-\sqrt{\lambda}\int_0^{2\pi}\frac{\cos\theta\,d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}-i\alpha_1\int_0^{2\pi}\frac{d\theta}{(\sqrt{\lambda}+i\alpha_1\cos\theta+i\alpha_2\sin\theta)}\\
&\quad=\frac{2\pi i\sqrt{\lambda}\alpha_1(\sqrt{\lambda+|\alpha|^2}-\sqrt{\lambda})}{|\alpha|^2\sqrt{\lambda+|\alpha|^2}}-\frac{2\pi i\alpha_2}{\sqrt{\lambda+|\alpha|^2}}.
\end{aligned}$$

Hence,

$$\frac{i\alpha_1}{4\pi^2\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{(i\alpha_1 z_1+i\alpha_2 z_2)}{|z|^3}e^{-\sqrt{\lambda}|z|}\,e^{-i\alpha\cdot z}\,dz=\frac{-i\alpha_1(\sqrt{\lambda+|\alpha|^2}-\sqrt{\lambda})}{2\pi\sqrt{\lambda}\sqrt{\lambda+|\alpha|^2}}+\frac{i\alpha_1|\alpha|^2}{2\pi\lambda\sqrt{\lambda+|\alpha|^2}}.$$

We are now in a position to compute $\hat{\underline{\hat{Y}}}_j(\alpha)$. An exactly similar calculation yields $\hat{\underline{\hat{X}}}_j(\alpha)$. The results are:

$$\begin{aligned}
\hat{\underline{\hat{Y}}}_j(\alpha)&=\frac{-i\alpha_j}{4\pi\lambda\sqrt{\lambda+|\alpha|^2}}(\lambda+2|\alpha|^2-2|\alpha|\sqrt{\lambda+|\alpha|^2})\\
&=\frac{-i\lambda\alpha_j}{4\pi\sqrt{\lambda+|\alpha|^2}(\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2}},\\
\hat{\underline{\hat{X}}}_j(\alpha)&=\frac{-i\alpha_j}{2\pi\lambda\sqrt{\lambda+|\alpha|^2}}(\lambda+|\alpha|^2-|\alpha|\sqrt{\lambda+|\alpha|^2})+\frac{i\alpha_j}{4\pi|\alpha|}\qquad(5.5)\\
&=\frac{-i\alpha_j\sqrt{\lambda+|\alpha|^2}}{2\pi(\lambda+|\alpha|^2+|\alpha|\sqrt{\lambda+|\alpha|^2})}+\frac{i\alpha_j}{4\pi|\alpha|}.
\end{aligned}$$

We now calculate the determinant of the operator defined by (5.4). It is equal to

$$\frac{1}{8}-2\pi^2(\hat{\underline{\hat{X}}}_1\hat{\underline{\hat{Y}}}_1+\hat{\underline{\hat{X}}}_2\hat{\underline{\hat{Y}}}_2)=\frac{1}{8}+\frac{|\alpha|^2(\lambda+2|\alpha|^2)}{2\lambda^2}-\frac{|\alpha|(\lambda^2+8\lambda|\alpha|^2+8|\alpha|^4)}{8\lambda^2\sqrt{\lambda+|\alpha|^2}}.$$

We will show that this determinant cannot vanish unless $\lambda\leqq 0$.

LEMMA 5.6. *Let*

$$\hat{D}=\frac{1}{8}+\frac{|\alpha|^2(\lambda+2|\alpha|^2)}{2\lambda^2}-\frac{|\alpha|(\lambda^2+8\lambda|\alpha|^2+8|\alpha|^4)}{8\lambda^2\sqrt{\lambda+|\alpha|^2}}.$$

Then if $\lambda\leqq 0$, *we have* $\hat{D}\neq 0$.

Proof. Assume $\lambda \neq 0$; then $\hat{D} = 0$ if and only if

$$\lambda^2 + 4\lambda|\alpha|^2 + 8|\alpha|^4 = \frac{|\alpha|(\lambda^2 + 8\lambda|\alpha|^2 + 8|\alpha|^4)}{\sqrt{\lambda + |\alpha|^2}}.$$

Since this will not be true for $|\alpha| = 0$ (unless $\lambda = 0$), we may divide both sides by $|\alpha|^4$, obtaining $z^2 + 4z + 8 = (z^2 + 8z + 8)/\sqrt{z+1}$, where $z = \lambda/|\alpha|^2$. Consider the equation $\sqrt{z+1} = (z^2 + 8z + 8)/(z^2 + 4z + 8)$. Note that the real part of the left-hand side is always positive for $z \not\leqq 0$. Bearing this in mind, we square both sides of the equation. Thus, we see that $z + 1 = (z^2 + 8z + 8)^2/(z^2 + 4z + 8)^2$ if and only if $\pm\sqrt{z+1} = (z^2 + 8z + 8)/(z^2 + 4z + 8)$. Now $z + 1 = (z^2 + 8z + 8)^2/(z^2 + 4z + 8)^2$ if and only if $z^3 + 8z^2 + 24z + 16 = 0$. By differentiating, we see that $z^3 + 8z^2 + 24z + 16$ is an increasing function for z real; hence, it has one real, negative root and two complex roots. Let the complex roots be z_0 and $\bar{z}_0$. We will show that the complex roots are not roots of $\sqrt{z+1} = (z^2 + 8z + 8)/(z^2 + 4z + 8)$. Note that it is enough to do this for one of them. So we will show that $\mathrm{Re}\,[(z_0^2 + 8z_0 + 8)/(z_0^2 + 4z_0 + 8)] < 0$. Using the cubic formula, we find that

$$\mathrm{Re}\,(z_0) = -\tfrac{1}{3}[(17 + \sqrt{17^2 + 8})^{1/3} - (\sqrt{17^2 + 8} - 17)^{1/3} + 8]$$

and

$$\mathrm{Im}\,(z_0) = (1/\sqrt{3})[(17 + \sqrt{17^2 + 8})^{1/3} + (\sqrt{17^2 + 8} - 17)^{1/3}].$$

Note that $17.233 < \sqrt{17^2 + 8} < 17.234$ and $1.732 < \sqrt{3} < 1.7321$. Hence, we find that $3.246 < (17 + \sqrt{17^2 + 8})^{1/3} < 3.248$ and $0.6153 < (\sqrt{17^2 + 8} - 17)^{1/3} < 0.6163$. We can now estimate z_0, z_0^2; we obtain

$$-3.5443 < \mathrm{Re}\,(z_0) < -3.5432, \qquad 2.2293 < \mathrm{Im}\,(z_0) < 2.2312,$$

$$7.5760 < \mathrm{Re}\,(z_0^2) < 7.5923, \qquad -15.816 < \mathrm{Im}\,(z_0^2) < -15.7977.$$

Estimates on $z_0^2 + 8z_0 + 8$, $z_0^2 + 4z_0 + 8$ follow:

$$-12.7784 < \mathrm{Re}\,(z_0^2 + 8z_0 + 8) < -12.7533, \qquad 1.3988 < \mathrm{Re}\,(z_0^2 + 4z_0 + 8) < 1.4195,$$

$$2.0184 < \mathrm{Im}\,(z_0^2 + 8z_0 + 8) < 2.0519, \qquad -6.8988 < \mathrm{Im}\,(z_0^2 + 4z_0 + 8) < -6.8729.$$

The sign of $\mathrm{Re}\,[(z_0^2 + 8z_0 + 8)/(z_0^2 + 4z_0 + 8)]$ is equal to the sign of $\mathrm{Re}\,[(z_0^2 + 8z_0 + 8)(\bar{z}_0^2 + 4\bar{z}_0 + 8)]$, which we can now easily see to be negative.

We now consider the solution $\boldsymbol{\phi}$ to (5.2) with $\mathbf{a} \in L_p(R^2)$. Letting $\hat{\Sigma}_\lambda$ be the matrix in (5.4), we see that for $\lambda \not\leqq 0$ there is a unique solution with $\hat{\boldsymbol{\phi}} = \hat{\Sigma}_\lambda^{-1}\hat{a}$ because the determinant of $\hat{\Sigma}_\lambda$, $\hat{D}$, does not vanish on R^2. It is also the case that all the components of $\hat{\Sigma}_\lambda^{-1}$ satisfy the conditions of Theorem 3.7 with $k = 2$, so that the map from $\mathbf{a} \to \boldsymbol{\phi}$ is bounded on $L_p(R^2)$, $1 < p < \infty$. To prove this, note that if $\hat{m}_1(\alpha)$ and $\hat{m}_2(\alpha)$ satisfy the conditions of Theorem 3.7, then clearly so do $\hat{m}_1 + \hat{m}_2$ and $\hat{m}_1\hat{m}_2$. Hence, it is enough to show that $\hat{X}_j$, $\hat{Y}_j$, and $1/\hat{D}$ satisfy the conditions of the multiplier theorem.

THEOREM 5.7. *$\hat{X}_j$, $\hat{Y}_j$, and $1/\hat{D}$ satisfy the conditions of Theorem 3.7 with $k = 2$.*

Proof. Note first that $\sqrt{\lambda + |\alpha|^2}$, $\lambda + 2|\alpha|^2 + 2|\alpha|\sqrt{\lambda + |\alpha|^2}$, and $\lambda + |\alpha|^2 + |\alpha|\sqrt{\lambda + |\alpha|^2}$ cannot vanish unless $\lambda \leqq 0$. This is clear for $\sqrt{\lambda + |\alpha|^2}$. If $\lambda + 2\alpha|^2 + 2|\alpha|\sqrt{\lambda + |\alpha|^2} = 0$, then $(\lambda + 2|\alpha|^2)^2 = 4|\alpha|^2(\lambda + |\alpha|^2)$, so that $\lambda^2 = 0$; similarly, if $\lambda + |\alpha|^2 + |\alpha|\sqrt{\lambda + |\alpha|^2} = 0$, we must have $\lambda(\lambda + |\alpha|^2) = 0$. Note that for this reason

$$\hat{\underline{Y}}_j(\alpha) = \frac{-i\lambda\alpha_j}{4\pi\sqrt{\lambda + |\alpha|^2}(\lambda + 2|\alpha|^2 + 2|\alpha|\sqrt{\lambda + |\alpha|^2})}$$

is continuous and $|\hat{\underline{Y}}_j(\alpha)| = O(1/|\alpha|^2)$ for large $|\alpha|$. Hence, $\hat{\underline{Y}}_j(\alpha)$ is bounded.

We will only consider one term of $\hat{\underline{X}}_j(\alpha)$, namely $(-i\alpha_j\sqrt{\lambda+|\alpha|^2})/(2\pi(\lambda+|\alpha|^2+|\alpha|\sqrt{\lambda+|\alpha|^2}))$, which is also continuous and bounded by $1/4\pi$, so it is uniformly bounded. The other term, $i\alpha_j/4\pi|\alpha|$, was shown to be a multiplier in § 3. Now,

$$\begin{aligned}\frac{\partial\hat{\underline{Y}}_j(\alpha)}{\partial\alpha_k}&=\frac{-i\lambda\delta_j^k}{4\pi\sqrt{\lambda+|\alpha|^2}(\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2})}\\&\quad-\frac{i\lambda\alpha_j}{16\pi^2}\left(\frac{1}{(\lambda+|\alpha|^2)(\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2})^2}\right)\\&\quad\cdot\left(\frac{\alpha_k}{\sqrt{\lambda+|\alpha|^2}}(\lambda+2|\alpha|+2|\alpha|\sqrt{\lambda+|\alpha|^2})\right)\\&\quad+\sqrt{\lambda+|\alpha|^2}\left(4\alpha_k+\frac{2\alpha_k}{|\alpha|}\sqrt{\lambda+|\alpha|^2}+\frac{2\alpha_k|\alpha|}{\sqrt{\lambda+|\alpha|^2}}\right),\end{aligned}$$

so

$$\begin{aligned}\left||\alpha|\frac{\nabla\hat{\underline{Y}}_j(\alpha)}{\partial\alpha_k}\right|&\leqq\left|\frac{|\alpha|}{4\pi\sqrt{\lambda+|\alpha|^2}(\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2})}\right|\\&\quad+\frac{|\lambda|}{16\pi^2}\frac{|\lambda|^3}{|(\lambda+|\alpha|^2)|^{3/2}}\frac{1}{|\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2}|}\\&\quad+\frac{|\lambda|}{16\pi^2}\frac{|\alpha|^2}{|\sqrt{\lambda+|\alpha|^2}|}\frac{1}{|\lambda+2|\alpha|^2+2|\alpha|\sqrt{\lambda+|\alpha|^2}|^2}\\&\quad\cdot\left(4|\alpha|+2|\sqrt{\lambda+|\alpha|^2}|+\frac{2|\alpha|^2}{|\sqrt{\lambda+|\alpha|^2}|}\right)\\&=O\left(\frac{1}{|\alpha|^2}\right),\end{aligned}$$

as $|\alpha|\to 0$, and s_0 is bounded. Another differentiation shows that $||\alpha|^2(\partial^2\hat{\underline{Y}}_j/\partial x_k\partial x_l)|$ is also bounded. A similar calculation applies to $\hat{\underline{X}}_j(\alpha)$. Now consider $1/\hat{D}$. $\hat{D}$ is a linear combination of $\hat{\underline{X}}_j$ and $\hat{\underline{Y}}_j$ so that $\hat{D}$ satisfies the conditions of Theorem 3.7. Furthermore, since by Lemma 5.6 $\hat{D}$ does not vanish and since $\hat{D}$ is continuous and $\hat{D}\to\frac{1}{8}$ as $|\alpha|\to\infty$, we have that $\hat{D}$ is bounded below. The fact that $\hat{D}$ satisfies the conditions of Theorem 3.7 and that $1/\hat{D}$ is bounded implies that $1/\hat{D}$ satisfies the conditions of the multiplier theorem, too. In this case ($k=2$), we have that

$$\begin{aligned}\left||\alpha|\frac{\partial}{\partial\alpha_j}\left(\frac{1}{\hat{D}}\right)\right|&=\left||\alpha|\left(\frac{\frac{-\partial\hat{D}}{\partial\alpha_j}}{\hat{D}^2}\right)\right|\\&\leqq\left(\max_\alpha\left|\frac{1}{\hat{D}}\right|\right)^2\max_\alpha\left||\alpha|\frac{\partial\hat{D}}{\partial\alpha_j}\right|\leqq C\end{aligned}$$

and

$$\begin{aligned}\left||\alpha|^2\frac{\partial^2}{\partial\alpha_k\,\partial\alpha_j}\left(\frac{1}{\hat{D}}\right)\right|&=|\alpha|^2\left|\frac{-\hat{D}^2\frac{\partial^2\hat{D}}{\partial\alpha_j\,\partial\alpha_k}+2\hat{D}\frac{\partial\hat{D}}{\partial\alpha_k}\frac{\partial\hat{D}}{\partial\alpha_j}}{\hat{D}^4}\right|\\&\leqq\frac{1}{|\hat{D}^2|}|\alpha|^2\left|\frac{\hbar^2\hat{D}}{\partial\alpha_j\,\partial\alpha_k}\right|+\frac{2}{|\hat{D}^3|}|\alpha|\left|\frac{\partial\hat{D}}{\partial\alpha_k}\right||\alpha|\left|\frac{\partial\hat{D}}{\partial\alpha_j}\right|\\&\leqq C.\end{aligned}$$

It easily follows from Theorem 5.7 (in the same way that Theorems 3.10 and 3.11 were proved) that:

THEOREM 5.8. *Given* $\mathbf{a} \in L_p(R^2)$, $1 < p < \infty$, *and* $\lambda \not\leqq 0$, *there is a unique solution* $\boldsymbol{\phi} \in L_p(R^2)$ *to* (5.2) *and* $\|\boldsymbol{\phi}\|_{L_p} \leqq C(\lambda)\|\boldsymbol{\alpha}\|_{L_p}$. *Furthermore, if* $\mathbf{a} \in W^{s,p}(R^2)$, *then* $\boldsymbol{\phi} \in W^{s,p}(R^2)$ *and* $\|\boldsymbol{\phi}\|_{W^{s,p}} \leqq C(\lambda)\|\mathbf{a}\|_{W^{s,p}}$.

6. Estimates for the solution of the boundary value problem. In this section, we will consider the solution to (5.1) and the differentiability properties of that solution. In particular, given $\lambda \not\leqq 0$ with $|\lambda| = 1$, $\mathbf{f} \in L_p(H)$, for $1 < p < \infty$, and letting $\mathbf{v}, q$ satisfy (3.1), we shall find the solution to (5.1) in the form $\mathbf{w} = \mathbf{w}(x, \boldsymbol{\phi})$, $\psi = \psi(x, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is found as in § 5. We shall also prove the estimate

$$\|\mathbf{w}\|_{W^{s+2,p}(H)} + \|\nabla\psi\|_{W^{s,p}(H)} \leqq C_{\sigma_0}\|\mathbf{f}\|_{W^{s,p}(H)} \quad \text{for } |\arg\lambda| < \sigma_0 < \pi. \tag{6.1}$$

We proceed as follows. First, we consider the solution $\mathbf{w}$, ψ to (5.2) with $\mathbf{a} = -\mathbf{v}|_{\partial H}$ and show that this satisfies (6.1), and then we show that it is the solution to (5.1). Recall that

$$\nabla\psi = 2\nabla \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\partial p^i}{\partial x_3}(x-y)\phi_i(y)\,dy_1\,dy_2 - \frac{\lambda}{4\pi}\nabla \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\phi_3(y)}{|x-y|}\,dy_1\,dy_2,$$

where $\boldsymbol{\phi}$ solves (5.2) with $\mathbf{a}$ as above. We consider first the term

$$2\nabla \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\partial p^i}{\partial x_3}(x-y)\phi_i(y)\,dy_1\,dy_2.$$

The jth component is

$$2\frac{\partial^2}{\partial x_j\,\partial x_i}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{x_3}{4\pi|x-y|^2}\phi_i(y)\,dy_1\,dy_2.$$

THEOREM 6.2. *Let*

$$h_i(x) = -\frac{1}{4\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{x_3}{|x-y|^3}\phi_i(y)\,dy_1\,dy_2,$$

where $\boldsymbol{\phi}$ *is as defined above. Then all second partials of* $h_i \in L_p(H)$, *and in particular, we will have* $\|(\partial^2 h_i/\partial x_j\,\partial x_i)\|_{L_p(H)} \leqq C(\sigma_0)\|\mathbf{f}\|_{L_p(H)}$.

Proof. First of all, notice that $\phi_i(y) \in W^{2-1/p,p}(\partial H)$. This is because $\mathbf{f} \in L_p(H)$, so $\mathbf{v} \in W^{2,p}(H)$ and so $-\mathbf{v}|_{\partial H} \in W^{2-1/p,p}(\partial H)$ by standard trace theorems. Since L_p multipliers satisfying Theorem 3.7 are in fact $W^{s,p}$ multipliers, we then have $\boldsymbol{\phi} \in W^{2-1/p,p}(\partial H)$ and $\|\boldsymbol{\phi}\|_{W^{2-1/p,p}(\partial H)} \leqq C(\sigma_0)\|\mathbf{f}\|_{L_p(H)}$. Now, the argument used in Theorem 4.9 shows that h_i is continuous on H and that differentiation under the integral sign is justified (because the integral converges locally uniformly in x), so that $\Delta h_j \equiv 0$ on H. Now suppose that $\mathbf{f} \in C_0^\infty(H)$. Since $\mathbf{f}$ will then be $W^{s,r}$ for all s, r, we will have $\boldsymbol{\phi} \in W^{s,r}$ for all s and r satisfying $1 < r < \infty$. Hence, $\boldsymbol{\phi}(x)$ will be continuous and decreasing to zero as $|x| \to \infty$. We can then conclude, using arguments similar to those in Theorems 4.10 and 4.11, that $\lim_{x \to x^0 \in \partial H, x \in H} h_i(x) = \phi_i(x^0)/2$. Indeed, we calculate that if $x_3 < 0$, then

$$-\frac{1}{4\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{x_3 c\,dx}{|x-y|^3} = \frac{c}{2}.$$

Next let $x^0 \in \partial H$, $x \in H$, and consider

$$h_i(x) + \frac{\phi_i(x^0)}{2} = -\frac{x_3}{4\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\phi_i(y) - \phi_i(x^0)}{|x-y|^3}\,dy_1\,dy_2.$$

Choose $\delta>0$ so that if $|y-x^0|<\delta$, then $|\phi_i(y)-\phi_i(x^0)|<\varepsilon$ and then let $\sqrt{(x_1x_1^0)^2+(x_2-x_2^0)^2}<\delta/2$. We see that

$$\left|h_i(x)+\frac{\phi_i(x^0)}{2}\right| \leqq \frac{|x_3|}{4\pi}\int\int_{B_\delta(x^0)} \frac{|\phi_i(y)-\phi_i(x^0)|}{|x-y|^3}\,dy_1\,dy_2$$

$$+\frac{|x_3|}{4\pi}\int\int_{\partial H-B_\delta(x^0)} \frac{|\phi_i(x^0)|}{|x-y|^3}\,dy_1\,dy_2$$

$$\leqq c\varepsilon+\max_y|\phi(y)|c\frac{|x_3|}{\delta}\to 0 \quad \text{as} \quad (x_1,x_2,x_3)\to(x_1^0,x_2^0,0)$$

This argument can also be used to show $h_i(x)\to_{|x|\to\infty}0$. First, notice that because $\boldsymbol{\phi}\in L_p(\partial H)$, we have

$$|h_i(x)|\leqq\frac{c\|\boldsymbol{\phi}\|_{L_p}}{|x_3|^{3(q-1)}}\to_{|x_3|\to\infty}0.$$

Now choose r such that if $|y|>r$, then $|\phi_i(y)|<\varepsilon$. Choose x so that $\sqrt{x_1^2+x_2^2}\geqq r+1/\varepsilon$. Then we have

$$|h_i(x)|\leqq\frac{|x_3|}{4\pi}\int_{B_{1/\varepsilon}(x_1,x_2)}\frac{|\phi_i(y)|}{|x-y|^3}\,dy_1\,dy_2+\frac{|x_3|}{4\pi}\int_{\partial H-B_{1/\varepsilon}(x_1,x_2)}\frac{|\phi_i(y)|\,dy_1\,dy_2}{|x-y|^3}$$

$$\leqq c\varepsilon+\max_{y\in\partial H}|\phi_i(y)|c\varepsilon.$$

Hence, $|h_i(x)|$ will be small if either $|x_3|$ is large or $\sqrt{x_1^2+x_2^2}$ is large. Thus, $h_i(x)\to_{|x|\to\infty}0$.

Thus, we see that $h_i$satisfies $\Delta h_i=0$ on H, $h_i|_{\partial H}=\phi_i/2$ and $|h_i(x)|\to_{|x|\to\infty}0$. Given ϕ_i, this makes h_i unique. However, if $\boldsymbol{\phi}\in W^{s,r}(\partial H)$ for all s and r with $1<r<\infty$, then there is a solution g to the problem

$$\Delta g=0 \quad \text{on } H, \qquad g|_{\partial H}=\boldsymbol{\phi}, \tag{6.3}$$

satisfying

$$g(x)\to_{|x|\to\infty}0$$

and

$$\left\|\frac{\partial^2 g}{\partial x_i\,\partial x_j}\right\|_{L_p(h)}\leqq C_p\|\boldsymbol{\phi}\|_{W^{2-1/p,p}(\partial H)}, \qquad 1\leqq i,j\leqq 3.$$

Hence, we have, for a dense set of $\mathbf{f}\in L_p(H)$, that

$$\left\|\frac{\partial^2 h_i}{\partial x_i\,\partial x_j}\right\|_{L_p(H)}\leqq C_r\|\phi_i\|_{W^{2-1/p,p}(\partial H)}\leqq C_p\|\mathbf{f}\|_{L_p(H)}.$$

Thus, the estimate is true for all $\mathbf{f}\in L_p(H)$.

THEOREM 6.4. *Under the same hypotheses as Theorem* 6.2, *we have that the functions*

$$\frac{\lambda}{4\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{(x_j-y_j)}{|x-y|^3}\phi_3(y)\,dy_1\,dy_2$$

are in $L_p(H)$, $1\leqq j\leqq 3$, *and*

$$\left\|\frac{\lambda}{4\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{(x_j-y_j)}{|x-y|^3}\phi_3(y)\,dy_1\,dy_2\right\|\leqq C(\sigma_0)\|\mathbf{f}\|_{L_p(H)}.$$

Proof. For simplicity, we shall ignore multiplicative constants throughout this proof. Let $a_k(x_1, x_2) = -v_k(x_1, x_2, 0)$. Then, using (5.4) and letting $r^2 = \alpha_1^2 + \alpha_2^2$, we have

$$4\hat{D}\hat{\phi}_3(\alpha_1, \alpha_2) = \frac{i\alpha_1\hat{a}_1 + i\alpha_2\hat{a}_2}{\sqrt{1+r^2}}(1+2r^2-2r\sqrt{1+r^2}) + \hat{a}_3,$$

where $\mathbf{a} = -\mathbf{v}|_{\partial H}$. We then use (3.5) and compute that

$$\hat{\phi}_3(\alpha_1, \alpha_2) = \sum_{k=1}^{3} \int_{-\infty}^{+\infty} \hat{n}_k(\alpha_1, \alpha_2, \alpha_3)\hat{f}_k(\alpha_1, \alpha_2, \alpha_3)\, d\alpha_3,$$

where

$$\begin{aligned}
\hat{n}_k(\alpha_1, \alpha_2, \alpha_3) &= \frac{1}{4\hat{D}} \frac{1}{1+|\alpha|^2}\left\{\frac{i\alpha_k\alpha_3^2}{|\alpha|^2} \frac{1}{\sqrt{1+r^2}(1+2r^2+2r\sqrt{1+r^2})} - \frac{\alpha_k\alpha_3}{|\alpha|^2}\right\}, \quad k = 1, 2, \\
\hat{n}_3(\alpha_1, \alpha_2, \alpha_3) &= \frac{1}{4\hat{D}} \frac{1}{1+|\alpha|^2}\left\{\frac{-i\alpha_3 r^2}{|\alpha|^2} \frac{1}{\sqrt{1+r^2}(1+2r^3+2r\sqrt{1+r^2})} + \frac{r^2}{|\alpha|^2}\right\}.
\end{aligned} \tag{6.5}$$

Note that $\hat{n}_k(\alpha_1, \alpha_2, \alpha_3) = \hat{m}_k(\alpha_1, \alpha_2, \alpha_3)(\alpha_k\alpha_3/|\alpha|^2)$, $k = 1, 2$, and $\hat{n}_3(\alpha_1, \alpha_2, \alpha_3) = \hat{m}_3(\alpha_1, \alpha_2, \alpha_3)r^2/|\alpha|^2$, where $\hat{m}_k(\alpha_1, \alpha_2, \alpha_3)$ is an L_p multiplier. If we let

$$g_j(x) = \frac{\lambda}{4\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{(x_j - y_j)}{|x-y|^3} \phi_3(y)\, dy_1\, dy_2,$$

we now see that

$$\begin{aligned}
\hat{g}_j(\alpha_1, \alpha_2, \alpha_3) = \frac{\alpha_j}{|\alpha|^2} \int_{-\infty}^{+\infty} \Bigg[&\frac{\alpha_1 z}{r^2+z^2}\hat{m}_1(\alpha_1, \alpha_2, z)\hat{f}_1(\alpha_1, \alpha_2, z) \\
&+ \frac{\alpha_2 z}{r^2-z^2}\hat{m}_2(\alpha_1, \alpha_2, z)\hat{f}_2(\alpha_1, \alpha_2, z) \\
&+ \frac{r^2}{r^2+z^2}\hat{m}_3(\alpha_1, \alpha_2, z)\hat{f}_3(\alpha_1, \alpha_2, z)\Bigg]\, dz.
\end{aligned}$$

Now we define operators K_{jk} on $L_p(R^3)$ via

$$\begin{aligned}
\widehat{K_{jk}f}(\alpha_1, \alpha_2, \alpha_3) &= \frac{\alpha_j}{|\alpha|^2} \int_{-\infty}^{+\infty} \frac{\alpha_k z}{r^2+z^2}\hat{f}(\alpha_1, \alpha_2, z)\, dz, \qquad k = 1, 2, \\
\widehat{K_{j3}f}(\alpha_1, \alpha_2, \alpha_3) &= \frac{\alpha_j}{|\alpha|^2} \int_{-\infty}^{+\infty} \frac{r^2}{r^2+z^2}\hat{f}(\alpha_1, \alpha_2, z)\, dz.
\end{aligned} \tag{6.6}$$

Clearly, we will be done if we can show that K_{jk} is a bounded operator on $L_p(R^3)$, $1 < p < \infty$. We will do this by proving that K_{jk} is bounded on $L_2(R^3)$ and that it and its dual are both of weak type 1–1. A trivial calculation shows that each K_{jk} is bounded on $L_2(R^3)$. To give one example:

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |K_{31}f(\alpha_1, \alpha_2, \alpha_3)|^2\, d\alpha_3\, d\alpha_2\, d\alpha_1 \\
&\quad \leqq \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\alpha_3^2}{|\alpha|^4}\left(\int_{-\infty}^{+\infty} \frac{\alpha_1^2 z^2}{(r^2+z^2)^2}\, dz \int_{-\infty}^{+\infty} |\hat{f}(\alpha_1, \alpha_2, w)|^2\, dw\right) d\alpha_3\, d\alpha_2\, d\alpha_1 \\
&\quad = c \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\alpha_1^2\alpha_3^2}{r|\alpha|^4} \int_{-\infty}^{+\infty} |\hat{f}(\alpha_1, \alpha_2, w)|^2\, dw\, d\alpha_3\, d\alpha_2\, d\alpha_1 \\
&\quad = c \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\alpha_1^2}{r^2} |\hat{f}(\alpha_1, \alpha_2, w)|^2\, dw\, d\alpha_2\, d\alpha_1 \\
&\quad \leqq c\|\hat{f}\|_{L_2} = c\|f\|_{L_2}.
\end{aligned}$$

Letting

$$\hat{L}_{jk}(\alpha_1, \alpha_2, \alpha_3, z) = \frac{\alpha_j}{|\alpha|^2} \frac{\alpha_k z}{(r^2+z^2)}, \qquad k = 1, 2,$$

and

$$\hat{L}_{j3}(\alpha_1, \alpha_2, \alpha_3, z) = \frac{\alpha_j}{|\alpha|^2} \frac{r^2}{r^2+z^2},$$

we see that

$$K_{jk}f(x_1, x_2, x_3) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L_{jk}(x_1-y_1, x_2-y_2, x_3-y_3, -z) f(y_1, y_2, y_3)\, dy_1\, dy_2\, dz.$$

We now compute L_{jk} and find that (mod multiplicative constants)

$$L_{jk}(x_1, x_2, x_3, z) = \frac{\partial^2}{\partial x_j\, \partial x_k}\left(\frac{\operatorname{sgn}(z)}{(x_1^2+x_2^2+(|x_3|+|z|)^2)^{1/2}}\right), \qquad j, k = 1, 2,$$

$$L_{3k}(x_1, x_2, x_3, z) = \frac{\partial^2}{\partial z\, \partial x_k}\left(\frac{\operatorname{sgn}(x_3)}{(x_1^2+x_2^2+(|x_3|+|z|)^2)^{1/2}}\right), \qquad k = 1, 2, \tag{6.7}$$

$$L_{j3}(x_1, x_2, x_3, z) = \frac{\partial}{\partial x_j}\left(\frac{|x_3|+|z|}{x_1^2+x_2^2+(|x_3|+|z|)^2)^{3/2}}\right).$$

At this point, we see that $L_{jk}(x_1, x_2, x_3, z) = A_{jk}(x_1, x_2, |x_3|+|z|)$, and so

$$K_{jk}f(x_1, x_2, x_3) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} A_{jk}(x_1-y_1, x_2-y_2, |x_3|+|y_3|) f(y_1, y_2, y_3)\, dy. \tag{6.8}$$

Since this operator is clearly self-adjoint, it is sufficient to prove that K_{jk} is of weak type 1–1. The following is true:

LEMMA 6.9. *Let* $Tf(x) = \int_{R^n} K(x, y) f(y)\, dy$ *where* K *is* C' *except when* $x = 0$ *or* $y = 0$ *and*

a) $\|Tf\|_{L_2(R^n)} \leqq C\|f\|_{L_2(R^n)}$,

b) $|\nabla_x K(x, y)| \leqq \dfrac{c}{|x-y|^{n+1}}$ *and* $|\nabla_y K(x, y)| \leqq \dfrac{c}{|x-y|^{n+1}}$.

Then T *is of weak type* $1-1$.

The proof of this lemma is an easy consequence of Stein [1, p. 29, proof of 2.2].

Since it is clear that

$$|\nabla A(s_1-y_1 x_2-y_2, |x_3|+|y_3|)| \leqq \frac{c}{((x_1-y_1)^2+(x_2-y_2)^2+(|x_3|+|y_3|)^2)^2}$$

$$\leqq \frac{c}{|x-y|^4},$$

the theorem follows.

We are now ready to consider the final estimates.

THEOREM 6.10. *Let* ϕ *be defined as in Theorem* 6.2 *and*

$$(\nabla\psi)_j = 2\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{\partial^2 p^i}{\partial x_j\, \partial x_3}(x-y)\phi_i(y)\, dy_1\, dy_2 + \frac{\lambda}{4\pi}\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{(x_j-y_j)}{|x-y|^2}\phi_3(y)\, dy_1\, dy_2,$$

Then $\nabla\psi \in G_p(H)$.

Proof. Because of Theorems 6.2 and 6.4 we already know that $\nabla\psi \in L_p(H)$. The proof that it is in $G_p(H)$ is the same as the proof in Theorem 3.10.

THEOREM 6.11. *If* ϕ *is as above, then* $\mathbf{w}(x, \phi)$ *and* $\nabla\psi(x, \phi)$ *satisfy*

$$\left.\begin{aligned} \lambda\mathbf{w} - \Delta\mathbf{w} + \nabla\psi = 0 \\ \nabla \cdot \mathbf{w} = 0 \end{aligned}\right\} \quad \text{in } H.$$

Furthermore, $\mathbf{w} \in W^{2,p}(H)$ *and* $\mathbf{w}$, ψ *satisfy* (6.1).

Proof. The first part of the theorem is true because differentiation under the integral sign is justified by the argument used in Theorem 4.9. By exactly the same argument as in Theorem 6.2, we see that $\mathbf{w}(x, \phi) \to_{|x|\to\infty} 0$ for smooth enough $\mathbf{f}$. In this case, $\mathbf{w}$ is a solution to

$$\lambda\mathbf{w} - \Delta\mathbf{w} = \nabla\psi \quad \text{on } H, \quad \text{and} \quad \mathbf{w}|_{\partial H} = -\mathbf{v}|_{\partial H}, \tag{6.12}$$

which vanishes at infinity.

Such solutions are unique. Since $\lambda - \Delta$ is invertible on halfspace, there is a solution $\mathbf{g}$ to (6.12) which vanishes at infinity. Furthermore,

$$\|\mathbf{g}\|_{W^{2,p}(H)} \leqq C(\sigma_0)(\|\nabla\psi\|_{L_p(H)} + \|\mathbf{v}\|_{W^{2-1/p,p}(\partial H)}) \leqq C(\sigma_0)\|\mathbf{f}\|_{L_p(H)}.$$

Hence, $\mathbf{w} = \mathbf{g}$ and satisfies (6.1) for smooth $\mathbf{f}$. By continuity in $L_p(H)$, $\mathbf{w}$ satisfies estimate 6.1 for all $\mathbf{f} \in L_p(H)$.

Now we know that $\mathbf{w}(x, \phi)$, $\psi(x, \phi)$ satisfy (5.1). Estimate (6.1) follows immediately using Theorems 3.11 and 5.8.

7. The inverse of $\lambda - \mathbb{P}\Delta$. Let $\mathbf{f} \in L_p(H)$, $\lambda \not\leqq 0$ and $|\lambda| = 1$ be given, and let $\mathbf{v}$, q and $\mathbf{w}$, ψ be defined as in previous sections. Let $\mathbf{u} = \mathbf{v} + \mathbf{w}$, $p = q + \psi$; then we have

$$\left.\begin{aligned} \lambda\mathbf{u} - \Delta\mathbf{u} + \nabla p = \mathbf{f}, \\ \nabla \cdot \mathbf{u} = 0, \\ \mathbf{u}|_{\partial H} = 0, \end{aligned}\right. \tag{7.1}$$

and $\mathbf{u} \in J_p(H) \cap \tilde{W}^{2,p}(H)$, $\nabla p \in L_p(H)$ and

$$\|\mathbf{u}\|_{W^{s+2,p}(H)} + \|\nabla p\|_{W^{s,p}(H)} \leqq C(\sigma_0)\|\mathbf{f}\|_{W^{s,p}(H)} \quad \text{if} \quad |\arg\lambda| < \sigma_0 < \pi. \tag{7.2}$$

This is exactly what we proved in previous sections. We first deal with $p = 2$.

LEMMA 7.3. *Let* $p = 2$. *Then* $\mathbb{P}\Delta$ *is a self-adjoint, nonpositive operator. For* $\lambda \not\leqq 0$, $\lambda - \mathbb{P}\Delta$: $\tilde{W}^{2,2}(H) \cap J_2(H)$ *is onto. Hence,* $\lambda - \mathbb{P}\Delta$ *is invertible and* $\mathbb{P}\Delta$ *generates a semigroup of contractions on* $J_2(H)$.

Proof. Let $\mathbf{u} \in C_0^\infty(H)$ with $\nabla \cdot \mathbf{u} = 0$. Then

$$\int_H \mathbf{u}\mathbb{P}\,\Delta\mathbf{u} = \int_H \mathbf{u}(\Delta\mathbf{u} + \nabla p) = \int_H \mathbf{u}\,\Delta\mathbf{u} = -\int_H |\Delta\mathbf{u}|^2 \leqq 0.$$

Taking limits in the $W^{2,2}(H)$-norm, we see that $\mathbb{P}\Delta$ is a nonpositive operator. Hence, $1 - \mathbb{P}\Delta$ is symmetric and onto, so it is self-adjoint and $\mathbb{P}\Delta$ is also self-adjoint.

The result that $\lambda - \mathbb{P}\Delta$ is invertible now follows in L_p because $\lambda - \mathbb{P}\Delta$: $\tilde{W}^{2,p}(H) \cap J_p(H) \to J_p(H)$ is onto, and because (7.1) and the fact that the operator is 1–1 if $p = 2$ imply that the inverse exists and is bounded.

8. Conclusions. We will now state concisely what we have proven. We use § 7 combined with § 1 and 2.

THEOREM 8.1. *Let* $\mathbf{f} \in J_p(H)$, $1 < p < \infty$. *Let* $\nu > 0$ *and let* $\lambda \not\geqq 0$ *and let* $|\arg \lambda| < \theta_0 < \pi$. *Then there exists a unique* $\mathbf{u} \in \mathring{W}^{2,p}(H) \cap J_p(H)$ *and a* $\nabla p \in G_p(H)$ *such that*

$$\left.\begin{aligned} \lambda \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= \mathbf{f} \\ \nabla \cdot \mathbf{u} &= 0 \\ \mathbf{u}|_{\partial H} &= 0 \end{aligned}\right\} \quad \text{in } H. \tag{8.2}$$

The following estimates are satisfied:

$$|\lambda| \|\mathbf{u}\|_{W^{s,p}(H)} + \sqrt{|\lambda|\nu} \|\nabla \mathbf{u}\|_{W^{s,p}(H)} + \nu \|\Delta \mathbf{u}\|_{W^{s,p}(H)} + \|\nabla p\|_{W^{s,p}(H)} \leqq C(\theta_0) \|\mathbf{f}\|_{W^{s,p}(H)}. \tag{8.3}$$

The consequence now follows that $\nu \mathbb{P}\Delta$ generates a bounded, analytic semigroup on $J_p(H)$. Estimates on the semigroup also follow.

THEOREM 8.4. *Let* $\mathbf{u}_0 \in J_p(H)$, $1 < p < \infty$. *Let* $e^{t\nu\mathbb{P}\Delta(\mathbf{u}_0)}$ *be the solution to the problem*

$$\left.\begin{aligned} \frac{\partial \mathbf{u}}{\partial t}(x, t) &= \nu \mathbb{P}\Delta \mathbf{u}(x, t) \\ \mathbf{u}(x, 0) &= \mathbf{u}_0(x) \end{aligned}\right\} \quad \text{for } x \in H, \quad t > 0. \tag{8.5}$$

Then

$$\frac{d^n e^{t\nu\mathbb{P}\Delta}(\mathbf{u}_0)}{dt^n} \in W^{2s,p}(H)$$

for all n, s *if* $\operatorname{Re}(t) \not\geqq 0$. *Furthermore,*

$$\left\| \frac{d^n e^{t\nu\mathbb{P}\Delta}(\mathbf{u}_0)}{dt^n} \right\|_{W^{2s,p}(H)} \leqq \frac{c(\theta_0, p, n, s) \nu^{l-i}}{|t|^{n+2-l}} \|\mathbf{u}_0\|_{W^{2l,p}(H)} \quad \text{as} \quad t \to 0,$$

if $|\arg t| < \theta_0 < \pi/2$ *and* $\mathbf{u}_0 \in \mathscr{D}((\nu\mathbb{P}\Delta)^l$.

Proof. As is well known, analytic semigroups can be continued into the complex plane. In fact, this theorem is a consequence of the fact that $\nu\mathbb{P}\Delta$ generates an analytic semigroup. (For the relevant properties, see Friedman [1]). Note that

$$\frac{d^n e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0}{dt^n} = \nu^n (\mathbb{P}\Delta)^n e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0.$$

Now,

$$(\mathbb{P}\Delta)^s (\mathbb{P}\Delta)^n e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0 = (\mathbb{P}\Delta\, e^{(t\nu/(n+s-l))\mathbb{P}\Delta})^{n+s-l} (\mathbb{P}\Delta)^l \mathbf{u}_0.$$

Because any continuous semigroup maps the underlying Banach space into the domain of the generator, we see that $d^n(e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0)/dt^n \in \mathscr{D}((\mathbb{P}\Delta)^s)$. Also, for any $\mathbf{u} \in \mathscr{D}((\mathbb{P}\Delta)^j)$, we have $\|\mathbf{u}\|_{W^{2j,p}(H)} \leqq c\|(1 - \mathbb{P}\Delta)^j \mathbf{u}\|_{L_p(H)}$. Letting $\mathbf{u} = d^n (e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0)/dt^n$, we see that

$$\begin{aligned} \left\| \frac{d^n e^{t\nu\mathbb{P}\Delta}\mathbf{u}_0}{dt^n} \right\|_{W^{2s,p}(H)} &\leqq c \left\| (1 - \mathbb{P}\Delta)^2 \left(\frac{d^n e^{t\nu\mathbb{P}}\Delta\mathbf{u}_0}{dt^n} \right) \right\|_{L_p(H)} \\ &\leqq \nu^n c \sum_{j=0}^{s} \left\| (\mathbb{P}\Delta\, e^{(t\nu/(n+j-l))\mathbb{P}\Delta})^{n+j-l} (\mathbb{P}\Delta)^l \mathbf{u}_0 \right\|_{L_p(H)} \\ &\leqq c \sum_{j=0}^{s} \frac{\nu^n}{|\nu t|^{n+j-l}} \|\mathbf{u}\|_{W^{2l,p}(H)} \\ &\leqq \frac{c\nu^{l-j}}{|t|^{n+s-l}} \|\mathbf{u}_0\|_{W^{2l,p}(H)} \quad \text{as} \quad t \to 0. \end{aligned}$$

We now consider the inhomogeneous problem for the Stokes equations using standard techniques. Many theorems are possible, depending upon the sense in which we define a solution to the inhomogeneous problem. We present one theorem here. The inhomogeneous problem is

$$\begin{aligned} &\frac{\partial \mathbf{u}}{\partial t}(x,t) - \nu \mathbb{P}\Delta \mathbf{u}(x,t) = \mathbf{f}(x,t) \quad \text{for} \quad x \in H, \\ &\mathbf{u}(x,0) = \mathbf{u}_0(x). \end{aligned} \tag{8.6}$$

The theorem below is a corollary of Kato [1, Chapt. IX, Thm. 1.27].

THEOREM 8.7. *Let $1<p<\infty$. Let $d^j\mathbf{f}(t)/dt^j \in \mathcal{D}((\nu\mathbb{P}\Delta)^{n-j})$ for some $n \leqq 0$ and all $j<n$, and let the map $t \to \mathbf{f}(t)$ from $\{t \in \mathbb{C} \mid t=0 \text{ or } \operatorname{Re}(t)>0\}$ be $C^{(n-j)}$ if $j<n$ and be locally Hölder continuous with exponent $\beta>0$ if $j=n$. Then for any $\mathbf{u}_0 \in \mathcal{D}((\nu\mathbb{P}\Delta)^m)$, $m \geqq 0$, $\mathbf{u}(t) = e^{t\nu\mathbb{P}\Delta}(\mathbf{u}_0) + \int_0^t e^{(t-2)\nu\mathbb{P}\Delta}(f(s))\,ds$ is a solution to (8.6). Furthermore, if $j \leqq n+1$ we have*

$$\frac{d^j\mathbf{u}(t)}{dt^j} = (\nu\mathbb{P}\Delta)^j\mathbf{u}(t) + \sum_{l=0}^{j-1} (\nu\mathbb{P}\Delta)^{j-l-1}\left(\frac{d^l\mathbf{f}(t)}{dt^l}\right).$$

We have the estimate

$$\begin{aligned} \left\|\frac{d^j\mathbf{u}(t)}{dt^j}\right\|_{W^{2(n+1-j),p}(H)} &\leqq \frac{C(\phi,j,n)\nu^{j-m}}{|t|^{n+1-m}}\|\mathbf{u}_0\|_{W^{2m,p}(H)} \\ &\cdot C(\phi,p,n,j)\Bigg(\sum_{l=0}^{j} \nu^{j-l-1} \max_{0\leqq|s|\leqq t}\left\|\frac{d^l\mathbf{f}(s)}{ds}\right\|_{W^{2(n-l),p}(H)} \\ &+ |t|^{\beta} \max_{|s_1|,|s_2|\leqq t} \frac{\|\mathbf{f}(s_1)-\mathbf{f}(s_2)\|}{|s_1-s_2|^{\beta}} \quad W^{2n,p}(H)\Bigg), \end{aligned}$$

if $|\arg t| < \theta_0 < \pi/2$.

Proof. See Kato [1, p. 491] for the proof that

$$\frac{d^j\mathbf{u}(t)}{dt^j} = (\nu\mathbb{P}\Delta)^j\mathbf{u}(t) + \sum_{l=0}^{j=1} (\nu\mathbb{P}\Delta)^{j-l-1}\left(\frac{d^l\mathbf{f}(t)}{dt^l}\right).$$

To prove the estimate, note that

$$\left\|(\nu\mathbb{P}\Delta)^{j-1-l}\frac{d^l\mathbf{f}(t)}{dt^l}\right\|_{W^{2(n+1-j),p}(H)} \leqq C\left\|\frac{d^l\mathbf{f}(t)}{dt^l}\right\|_{W^{2(n-l),p}(H)}.$$

We have that

$$(\nu\mathbb{P}\Delta)^j\mathbf{u}(t) = (\nu\mathbb{P}\Delta)^j \int_0^t e^{(t-s)\nu\mathbb{P}\Delta}(\mathbf{f}(s))\,ds + (\nu\mathbb{P}\Delta)^J e^{T\nu\mathbb{P}\Delta}(\mathbf{u}_0).$$

Note that

$$\|(\nu\mathbb{P}\Delta)^j e^{t\nu\mathbb{P}\Delta}(\mathbf{u}_0)\|_{W^{2(n+1-j),p}(H)} \leqq \frac{C(p,n,j,t)}{|t|^{n+1-m}}\|\mathbf{u}_0\|_{W^{2m,p}(H)}.$$

Now consider

$$\left\|(\nu\mathbb{P}\Delta)^j \int_0^t e^{(t-s)\nu\mathbb{P}\Delta}\mathbf{f}(s)\,ds\right\|_{W^{2(n+1-j),p}(H)}$$
$$\leqq C\left\|(1-\nu\mathbb{P}\Delta)^{n+1-j}(\nu\mathbb{P}\Delta)^j \int_0^t e^{(t-s)\nu\mathbb{P}\Delta}\mathbf{f}(s)\,ds\right\|_{L_p(H)}$$
$$\leqq C \sum_{l\leqq n+1}\left\|(\nu\mathbb{P}\Delta)^l \int_0^t e^{(t-s)\nu\mathbb{P}\Delta}\mathbf{f}(s)\,ds\right\|_{L_p(H)}.$$

Let $l < n+1$, and consider

$$\left\|(\nu\mathbb{P}\Delta)^l \int_0^t e^{(t-s)\nu\mathbb{P}\Delta}\mathbf{f}(s)\,ds\right\|_{L_p(H)} = \left\|\int_0^t e^{(t-s)\nu\mathbb{P}\Delta}((\nu\mathbb{P}\Delta)^l\mathbf{f}(s))\,ds\right\|_{L_p(H)}$$
$$\leqq C \max_{0\leqq|s|\leqq|t|}\|\mathbf{f}(s)\|_{W^{2l,p}(H)}.$$

Finally, consider

$$\left\|(\nu\mathbb{P}\Delta)^{n+1}\int_0^t e^{(t-s)\nu\mathbb{P}\Delta}\mathbf{f}(s)\,ds\right\|_{L_p(H)}$$
$$= \left\|(\nu\mathbb{P}\Delta)^{n+1}\int_0^t e^{(t-s)\nu\mathbb{P}\Delta}(\mathbf{f}(s)-\mathbf{f}(t))\,ds\right\|_{L_p(H)} + \left\|(\nu\mathbb{P}\Delta)^{n+1}\int_0^t e^{(t-s)\nu\mathbb{P}\Delta}(\mathbf{f}(t))\,ds\right\|_{L_p(H)}$$

The second term is equal to $\|(\nu\mathbb{P}\Delta)^n(1-e^{t\nu\mathbb{P}\Delta})\mathbf{f}(t)\|_{L_p(H)} \leqq C\|\mathbf{f}\|_{W^{2n,r}(H)}$ (see Kato [1, p. 491]. We have

$$\left\|(\nu\mathbb{P}\Delta)^{n+1}\int_0^t e^{(t-s)\nu\mathbb{P}\Delta}(\mathbf{f}(s)-\mathbf{f}(t))\,ds\right\|_{L_p(H)}$$
$$= \left\|\int_0^t (\nu\mathbb{P}\Delta)\,e^{(t-s)\nu\mathbb{P}\Delta}(\nu\mathbb{P}\Delta)^n(\mathbf{f}(t)-\mathbf{f}(s))\,ds\right\|_{L_p(H)}$$
$$\leqq \int_0^t \frac{c}{|t-s|}\|(\nu\mathbb{P}\Delta)^n(\mathbf{f}(t)-\mathbf{f}(s))\|_{L_p(H)}\,ds$$
$$\leqq \int_0^t \frac{c}{|t-s|}\|\mathbf{f}(t)-\mathbf{f}(s)\|_{W^{2n,p}(H)}\,ds$$
$$\leqq C \max_{|s_1|,|s_2|\leqq t}\frac{\|\mathbf{f}(s_1)-\mathbf{f}(s_2)\|_{W^{2n,p}(H)}}{|s_1-s_2|^\beta}\int_0^t \frac{1}{|t-s|^{1-\beta}}\,ds$$
$$\leqq C|t|^\beta \max_{|s_1|,|s_2|\leqq t}\frac{\|\mathbf{f}(s_1)-\mathbf{f}(s_2)\|_{W^{2n,p}(H)}}{|s_1-s_2|^\beta}.$$

The theorem follows, and the dependence of the estimates on ν is obvious.

In conclusion, we wish to discuss the extension of our result to more general domains in R^3. We conjecture that the Stokes equations generate a bounded analytic semigroup in $J_p(\Omega)$ if Ω is a bounded open set in R^3 with sufficiently smooth boundary. However, the results in halfspace cannot easily be extended. This is because of the boundary value problem corresponding to the one discussed in §§ 2 and 3. The

dependence of solutions to the integral equation on λ is an extremely difficult problem in an arbitrary domain. This is in marked contrast to the ease with which the extension can be proved in $J_2(\Omega)$ because there the operator $\nu\mathbb{P}\Delta$ is a nonpositive self-adjoint operator. We do not at this time wish to make any conjecture about the semigroup properties of the Stokes equations on general unbounded domains. Analysis of that problem will undoubtedly prove difficult (see, for example Heywood [1]–[3] and Ma [1]).

Appendix: The Hodge theorem in halfspace. In the appendix, we give a brief sketch of the proof of Lemma 1.7 of § 1. The Hodge theorem follows from the lemma below.

LEMMA A.1. *Let* $\mathbf{f} \in C_0^\infty(H)$. *Then there is a unique decomposition* $\mathbf{f} = \bar{\mathbf{X}} + \nabla\phi$ *with* $\nabla \cdot \bar{\mathbf{X}} = 0$, $\bar{X}_3(x_1, x_2, 0) = 0$ *and if* $1 < p < \infty$, *then*

$$\|\bar{\mathbf{X}}\|_{W^{s,p}(H)} + \|\nabla\phi\|_{W^{s,p}(H)} \leqq C\|\mathbf{f}\|_{W^{s,p}(H)}. \tag{A.2}$$

Proof. We see that we must solve the problem $\Delta\phi = \nabla \cdot \mathbf{f}$, $(\partial\phi/\partial x_3)(x_1, x_2, 0) = f_3(x_1, x_2, 0)$. Letting $x^* = (x_1, x_2, -x_3)$ we find that a solution is given by

$$\phi(x) = -\frac{1}{4\pi}\int_H \left(\frac{1}{|x-y|} + \frac{1}{|x^*-y|}\right)\nabla \cdot \mathbf{f}(y)\, dy$$

because $\Delta(-1/(4\pi|x|)) = \delta(x)$ and

$$\frac{\partial}{\partial x_3}\left(\frac{1}{|x-y|} + \frac{1}{|x^*-y|}\right)\Big|_{\partial H} = 0.$$

Since

$$\nabla_y \cdot \left(-\frac{1}{4\pi}\left(\frac{1}{|x-y|} + \frac{1}{|x^*-y|}\right)\mathbf{f}(y)\right)$$

$$= \frac{1}{4\pi}\left(\frac{1}{|x-y|} + \frac{1}{|x^*-y|}\right)(\nabla \cdot \mathbf{f}(y)) + \mathbf{f}(y) \cdot \nabla_y\left(-\frac{1}{4\pi}\left(\frac{1}{|x-y|} + \frac{1}{|x^*-y|}\right)\right),$$

and $\mathbf{f}|_{\partial H} = 0$, we have

$$\begin{aligned}\phi(x) &= \frac{1}{4\pi}\int_H \left(\frac{x_j - y_j}{|x-y|^3} + \frac{x_j^* - y_j}{|x^*-y|^3}\right) f_j(y)\, dy \\ &= \frac{1}{4\pi}\int_{R^3} \left\{\frac{x_1 - y_1}{|x-y|^3}(f_1(y) + f_1(y^*)) + \frac{x_2 - y_2}{|x-y|^3}(f_2(y) + f_2(y^*))\right. \\ &\qquad \left. + \frac{x_3 - y_3}{|x-y|^3}(f_3(y) - f_3(y^*))\right\} dy_1\, dy_2\, dy_3,\end{aligned} \tag{A.3}$$

if we extend $\mathbf{f}$ to be zero on H^+. Estimate (A.2) now follows in the same way as the estimates in § 3. That is, one takes the Fourier transform of $\phi(x)$ and, using the multiplier theorem, easily sees that $\nabla\phi$ belongs to $W^{s,p}(R^3)$ for all s, p because $\mathbf{f}$ does. Uniqueness follows because if $\mathbf{f} = \bar{\mathbf{X}}_1 + \nabla\phi_1 = \bar{\mathbf{X}}_2 + \nabla\phi_2$, then $\bar{\mathbf{X}} = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 = \nabla(\phi_1 - \phi_2) = \nabla\phi$ and $\nabla \cdot \bar{\mathbf{X}} = 0$ and $\bar{\mathbf{X}}(x_1, x_2, 0) = 0$. Hence,

$$\int_H \bar{\mathbf{X}} \cdot \bar{\mathbf{X}} = \int_H \bar{\mathbf{X}} \cdot \nabla\phi = \int_{\partial H} \bar{\mathbf{X}}_3\phi - \int_H (\nabla \cdot \bar{\mathbf{X}})\phi = 0.$$

To prove Lemma 1.9, let $J_p(H)$ and $G_p(H)$ be as defined in § 1. First note that if $\bar{\mathbf{X}} \in C_0^\infty(H)$ such that $\nabla \cdot \bar{\mathbf{X}} = 0$ and $\phi \in C^\infty(H)$ such that $\nabla\phi \in L_p(H)$, then $\int_H \bar{\mathbf{X}} \cdot \nabla\phi =$

0. Hence, $J_p(H) \cap G_p(H) = \{0\}$. To see that $J_p(H) \oplus G_p(H) = L_p(H)$, let $\mathbf{f} \in L_p(H)$ and $\mathbf{f}_n \to \mathbf{f}$ in L_p where $f_n \in C_0^\infty(H)$. Then $\mathbf{f}_n = \mathbf{X}_n + \nabla\phi_n$ as in Lemma A.1. By (A.2), $\lim \bar{\mathbf{X}}_n = \bar{\mathbf{X}}$ and $\lim \nabla\phi_n = \psi$ exist in $L_p(H)$ and we have $\bar{\mathbf{X}} + \psi = \mathbf{f}$. Finally, note that $\bar{\mathbf{X}} \in J_p(H)$ and $\psi \in G_p(H)$ by definition.

Acknowledgments. The author would particularly like to thank J. Marsden for his help. In addition, she thanks P. Chernoff, A. Chorin, T. Kato, F. Weissler, and E. Stein for interesting conversations. She also thanks C. Brown, P. Mathlick and C. Wright for typing the manuscript.

REFERENCES

A. FRIEDMAN, [1] *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, 1969.

T. KATO, [1] *Perturbation Theory for Linear Operators*, Springer Verlag, New York, 1966.

O. A. LADYZHENSKAYA, [1] *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, 1969.

P. E. SOBOLEVSKY, [1] *Investigation of the Navier-Stokes equations by methods of the theory of parabolic equations in Banach spaces*, Dokl. Akad. Nauk SSSR, 156, pp. 745–748.

V. A. SOLONNIKOV, [1] *On a priori extimates for certain boundary value problems*, Dokl. Akad. Nauk SSSR, 138, (1961), pp. 781–784.

———, [2] *Estimates of solutions of non stationary linearized systems of Navier Stokes equations*, Trudy Mat. Inst. Steklov, 70, (1964), pp. 213–317.

E. M. STEIN, [1] *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

J. G. HEYWOOD, [1] *On nonstationary Stokes flow past an obstacle*, Indiana Univ. Math. J., 24 (1974).

———, [2] *On some paradoxes concerning two dimensional Stokes flow past an obstacle*, Indiana Univ. Math. J., 24 (1974).

———, [3] *On uniqueness questions in the theory of viscous flow*, 1975, preprint.

C. MA, [1] *On Square Summability and Uniqueness Questions Concerning Nonstationary Stokes Flow in an Exterior Domain*, Thesis, University of British Columbia, 1975.

D. EBIN AND J. MARSDEN, [1] *Groups of diffeomorphisms and the notion of an incompressible fluid*, Ann. Math., 92 (1970), pp. 102–163.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0009 $01.00/0

EXISTENCE AND UNIQUENESS OF A CLASSICAL SOLUTION OF AN INITIAL BOUNDARY VALUE PROBLEM OF THE THEORY OF SHALLOW WATERS*

BUI AN TON†

Abstract. The existence of a unique solution in the spaces of functions with Hölder continuous derivatives of an initial boundary value problem in the theory of shallow waters is established. The solution is local in time. The method of successive approximations and the Lagrangian coordinates is used.

The purpose of this paper is to establish the existence of a unique solution, local in time, in the spaces of functions with Hölder continuous derivatives of an initial boundary value problem in the theory of shallow waters.

Let u be the velocity of the fluid, and gh be its geopotential. The motion of the fluid is described by the initial boundary value problem

$$\begin{aligned} &h\left(\frac{\partial u}{\partial t}+u\cdot\nabla u\right)-\nabla(h\nabla u)+h\ \text{grad}\ h=0 \quad \text{on } (0,T)\times\Omega,\\ &u(x,t)=0 \quad \text{on } (0,T)\times\partial\Omega, \qquad u(x,0)=u_0(x) \quad \text{on } \Omega, \end{aligned} \tag{0.1}$$

where Ω is a bounded open subset of R^3 with a smooth boundary $\partial\Omega$.

Conservation of mass is expressed by the initial value problem

$$\begin{aligned} &\frac{\partial h}{\partial t}+u\cdot\text{grad}\ h+h\ \text{div}(u)=0 \quad \text{on } (0,T)\times\Omega,\\ &h(x,t)>0 \quad \text{on } (0,T)\times\Omega, \qquad h(x,0)=h_0(x)>0 \quad \text{on } \Omega. \end{aligned} \tag{0.2}$$

The usual Coriolis term in (0.2) has been omitted since it does not affect the nature of the equation. The system (0.1)–(0.2) is a coupled "parabolic-hyperbolic" nonlinear system of partial differential equations. It is of a type arising frequently in the theory of water waves and compressible fluids. Systems of the type (0.1)–(0.2) have been called "incompletely parabolic" by Belov and Yanenko [1], Gustafsson and Sundstrom [3].

The existence of a weak local solution of (0.1)–(0.2) in Sobolev spaces has been established by the writer [8] using the method of successive approximations and a compensated compactness argument. In this paper, we shall show the existence of a unique solution $\{u, h\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$ of (0.1)–(0.2), using Lagrangian coordinates and the method of successive approximations. The result obtained seems new.

In 1962, Nash [5] initiated the use of Lagrangian coordinates together with the method of successive approximations in the study of compressible fluids. The Nash approach was also found independently by Itaya [4] and used later on with some modifications by Tani [7]. In this paper, we use an approach different from that of Nash; it is related to the one introduced by Solonnikov [6] in the study of a free boundary problem of water waves.

The notations, the main result of the paper and a detailed outline of the proof of the theorem are given in § 1. The transformation relating Eulerian to Lagrangian coordinates is studied in § 2. The existence of a solution of a linear parabolic initial

* Received by the editors April 15, 1980.

† Department of Mathematics, University of British Columbia, Vancouver, B.C., Canada, V6K 2R7.

boundary value problem and an estimate for its solution are given in § 3. The construction of a sequence of successive approximations is carried out in § 4. The proof of the existence and uniqueness theorem is given in § 5.

1. Let Ω be a bounded open subset of R^3 with a boundary $\partial\Omega$ of class $C^{2+\alpha}$, $0<\alpha<1$ and let $(0, T)$ be a finite time interval of the real line. The generic point of Ω is $x=(x_1, x_2, x_3)$. Set $D_j=\partial/\partial x_j$, $Q_T=\Omega\times(0, T)$.

We denote

$$H_x^{\alpha}(u; Q_T)=\sup_{Q_T}\{|u(x, t)-u(y, t)|\,|x-y|^{-\alpha}\},$$

$$H_t^{\alpha}(u; Q_T)=\sup_{Q_T}\{|u(x, t)-u(x, s)|\,|t-s|^{-\alpha}\}.$$

We define the following norms:

$$\|u\|_{C^{\alpha,\alpha/2}(Q_T)}=\|u\|_{C(Q_T)}+H_x^{\alpha}(u; Q_T)+H_t^{\alpha/2}(u; Q_T), \tag{1.1}$$

$$\|u\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}=\|u\|_{C(Q_T)}+\sum_{j=1}^{3}\|D_j u\|_{C^{\alpha,\alpha/2}(Q_T)}+H_t^{(1+\alpha)/2}(u; Q_T), \tag{1.2}$$

and

$$\begin{aligned}\|u\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}&=\|u\|_{C^1(Q_T)}+H_x^{\alpha}\left(\frac{\partial u}{\partial t}; Q_T\right)+H_t^{\alpha/2}\left(\frac{\partial u}{\partial t}; Q_T\right)\\&\quad+\sum_{j,k=1}^{3}\|D_jD_k u\|_{C^{\alpha,\alpha/2}(Q_T)}+\sum_{j=1}^{3}H_t^{(1+\alpha)/2}(D_j u; Q_T).\end{aligned} \tag{1.3}$$

It is not difficult to check that $C^{s+\alpha,(s+\alpha)/2}(Q_T)$, $0<s<2$, is an algebra.

The main result of the paper is the following theorem.

THEOREM 1.1. *Let h_0 be a scalar function in $C^{1+\alpha}(\Omega)$, $0<\alpha<1$ with $h_0\geqq c>0$ on Ω. Let u_0 be a vector function in $C^{2+\alpha}(\Omega)$ with $u_0=0$ on $\partial\Omega$. Then there exist:*

(1) *a nonempty interval $(0, T^*)$,*

(2) *a unique $\{u, h\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$, a solution of (0.1)–(0.2)*

Before going into the details we shall first give a detailed outline of the steps involved.

Step 1. It will be carried out in § 2. The basic transformation

$$x=\xi+\int_0^t w(\xi, s)\,ds=X(\xi, t) \tag{1.4}$$

relating Eulerian coordinates $x=(x_1, x_2, x_3)$ to Lagrangian coordinates $\xi=(\xi_1, \xi_2, \xi_3)$ will be studied. Let

$$a^{jk}(\xi, t)=\delta_{jk}+\int_0^t\frac{\partial w_k}{\partial\xi_j}(\xi, s)\,ds, \qquad 1\leqq j, k\leqq 3.$$

Conditions on w and on T so that the matrix $A(\xi, t)=((a^{jk}(\xi, t)))$ has an inverse are given. Some simple estimates on A and on its inverse are established.

Step 2. Let

$$(T+T^{\beta})\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\leqq\delta, \qquad 0<\beta\leqq\frac{1-\alpha}{2},$$

with δ small. In § 3, we study the linear parabolic initial boundary value problem

$$\begin{aligned} &\frac{\partial v}{\partial t}-\nabla_w^2(v)=f(\xi,t) \quad \text{on } Q_T, \\ &v(\xi,t)=0 \quad \text{on } (0,T)\times\partial\Omega, \qquad v(\xi,0)=v_0(\xi) \quad \text{on } \Omega. \end{aligned} \tag{1.5}$$

∇_w is the operator $A^{-1}\nabla=\sum_{j=1}^{3} a_{jk}(\xi,t)\,\partial/\partial\xi$, $1\leqq k\leqq 3$ where a_{jk} are the entries of the inverse of the matrix A of Step 1. It will be shown that:

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\leqq K\{\delta\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|f\|_{C^{\alpha,\alpha/2}(Q_t)}+\|v_0\|_{C^{2+\alpha}(\Omega)}\}. \tag{1.6}$$

K is a constant independent of δ, t, v_0, f, v, w.

Step 3. It will be carried out in § 4. We construct $\{v^n,\rho^n\}$ by the equations

$$\rho^n(\xi,t)=\rho_0(\xi)\exp\left(-\int_0^t(\nabla_{n-1}\cdot v^{n-1})(\xi,s)\right)ds, \qquad v^0=v_0, \tag{1.7}$$

and by the initial boundary value problems

$$\begin{aligned} &\frac{\partial v^n}{\partial t}-\nabla_{n-1}^2 v^n+\nabla_{n-1}\rho^n-\frac{\{\nabla_{n-1}\rho^n\cdot\nabla_{n-1}v^{n-1}\}}{\rho^n}=0 \quad \text{on } Q_T, \\ &v^n(\xi,t)=0 \quad \text{on } (0,T)\times\partial\Omega, \qquad v^n(\xi,0)=v_0(\xi) \quad \text{on } \Omega, \end{aligned} \tag{1.8}$$

where $\nabla_k=(A_{v_k})^{-1}\nabla$.

With (1.6), the estimates of Step 1 and noting that $v^n(\xi,0)=v_0(\xi)$ on Ω, one can show that there exist:

(i) a nonempty interval $(0,T^*)$ independent of n,

(ii) a constant K independent of n such that

$$\|\rho^n\|_{C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})}+\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})}\leqq K.$$

Moreover,

$$\{v^n,\rho^n\}\to\{v,\rho\} \quad \text{in } C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*}) \quad \text{as } n\to+\infty.$$

Step 4. With $\{v,\rho\}$ of Step 3 we go back to Eulerian coordinates via the transformation (1.4) and the theorem is proved.

2. Let w be a vector function in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$, $w=0$ on $\partial\Omega\times(0,T)$ and consider the one-parameter family of transformations

$$x=\xi+\int_0^t w(\xi,s)\,ds=X(\xi,t), \tag{2.1}$$

of Ω into Ω_t. Set

$$a^{jk}(\xi,t)=\delta_{jk}+\int_0^t\frac{\partial w_k(\xi,s)}{\partial\xi_j}\,ds, \qquad 1\leqq j,k\leqq 3 \tag{2.2}$$

where δ_{jk} is the Kronecker delta function.

The matrix $A(\xi,t)=((a^{jk}(\xi,t)))$ is the Jacobian of the transformation X connecting Lagrangian coordinates ξ to Eulerian coordinates x. In this section, we shall study A.

It is known that without any further condition on w, $\det(A(\xi,t))\neq 0$ only for small t. We express that restriction by assuming that:

$$(T+T^\beta)\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\leqq\delta\leqq\frac{1}{8}, \qquad 0<\beta<\frac{1-\alpha}{2}. \tag{2.3}$$

PROPOSITION 2.1. *Suppose that* (2.3) *is verified. Then*

$$\tfrac{1}{2} \leqq \det\{A(\xi, t)\} \leqq \tfrac{3}{2} \quad \text{for } 0 \leqq t \leqq T.$$

Proof. We have

$$\delta_{kj} - t \max_{j,k} \left\| \frac{\partial w_k}{\partial \xi_j} \right\|_{C(Q_t)} \leqq a^{kj}(\xi, t) \leqq \delta_{kj} + t \max_{j,k} \left\| \frac{\partial w_k}{\partial \xi_j} \right\|_{C(Q_t)}.$$

With (2.3) we get

$$\tfrac{1}{2} \leqq 1 - 3\delta - 6\delta^2 - 6\delta^3 \leqq \det\{A(\xi, t)\} \leqq 1 + 3\delta + 6\delta^2 + 6\delta^3 \leqq \tfrac{3}{2}.$$

PROPOSITION 2.2. *Let $a(\xi, t)$ be in $C^{\alpha,\alpha/2}(Q_T)$. Then*

$$\left\| \int_0^t a(\xi, s)\, ds \right\|_{C^{\alpha,\alpha/2}(Q_T)} \leqq \int_0^T \|a(\cdot, s)\|_{C^{\alpha}(\Omega)}\, ds + T^{1-\alpha/2} \|a\|_{C(Q_T)}.$$

Let $\varepsilon > 0$; then there exists $c(\varepsilon) > 0$ such that

$$\left\| \int_0^t a(\xi, s)\, ds \right\|_{C^{\alpha,\alpha/2}(Q_t)} \leqq \varepsilon \|a\|_{C(Q_t)} + c(\varepsilon) \int_0^t \|a(\cdot, s)\|_{C^{\alpha}(\Omega)}\, ds.$$

Proof. The above simple proposition which is very useful has been proved by Solonnikov [6, Lemma 3, pp. 1331–1332].

PROPOSITION 2.3. *Let $b(\xi, t)$ be in $C^{1+\alpha,(1+\alpha)/2}(Q_T)$. Then*

$$\left\| \int_0^t b(\xi, s)\, ds \right\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq \int_0^T \|b(\cdot, s)\|_{C^{1+\alpha}(\Omega)}\, ds + T^{(1-\alpha)/2} \|b\|_{C(Q_T)} + T^{1-\alpha/2} \sum_{j=1}^{3} \left\| \frac{\partial b}{\partial \xi_j} \right\|_{C(Q_T)}.$$

Let $\varepsilon > 0$ be given. Then there exists $c(\varepsilon) > 0$ such that

$$\left\| \int_0^t b(\xi, s)\, ds \right\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq \varepsilon \|b\|_{C^{1+\alpha,(L+\alpha)/2}(Q_T)} + c(\varepsilon) \int_0^t \|b(\cdot, s)\|_{C^{1+\alpha}(\Omega)}\, ds.$$

Proof. Cf. Solonnikov [6, Lemma 3, pp. 1331–1332].

LEMMA 2.1. *Let w be in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ and satisfy* (2.3). *Let $a^{jk}(\xi, t)$ be given by* (2.2) *and $A = A(\xi, t) = ((a^{jk}(\xi, t)))$. Then*

1) $$\|A - I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq 2\delta,$$

2) $$\|A - I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq c_1,$$

3) $$\|A^{-1} - I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq c_2 \|A - I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq 2c_2\delta,$$

where c_1, c_2 are independent of δ and of w.

Proof. This is Lemma 4 of Solonnikov [6, p. 1332].

LEMMA 2.2. *Let v, w be two vector functions in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ satisfying the condition* (2.3). *Let A_v, A_w be the Jacobian of the transformation* (2.1) *corresponding to v and to w, respectively. Then for any $\varepsilon > 0$ there exists $c(\varepsilon) > 0$ such that*

$$\|A_v - A_w\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq \varepsilon \|v - w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} + c(\varepsilon) \int_0^T \|v - w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\, ds.$$

Proof. Cf. [6, Lemma 6, p. 1333].

3. We shall carry out Step 2. The main result of the section is the following theorem.

THEOREM 3.1. *Let w be a vector function in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ and suppose that*

$$(T+T^{\beta})\|w\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} \leqq \delta < \frac{1}{8}, \qquad 0<\beta<\frac{(1-\alpha)}{2},$$

with δ small. Let f and v_0 be in $C^{\alpha,\alpha/2}(Q_T)$ and in $C^{2+\alpha}(\Omega)$ respectively with $v_0=0$ on $\partial\Omega$. Then there exists v in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$, a solution of the initial boundary value problem

$$(3.1) \qquad \frac{\partial v}{\partial t}-\nabla_w^2 v = f \quad \text{on } Q_T, \quad v=0 \quad \text{on } (0,T)\times\partial\Omega, \quad v(\xi,0)=v_0(\xi) \quad \text{on } \Omega.$$

Moreover

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} \leqq K\{\delta\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|f\|_{C^{\alpha,\alpha/2}(Q_T)}+\|v_0\|_{C^{2+\alpha}(\Omega)}\}$$

for $0<t<T$.

K is a constant independent of t, δ, w, v_0, f, v. ∇_w is the operator

$$A^{-1}\nabla = \sum_{j=1}^{3} a_{jk}(\xi,t)\frac{\partial}{\partial\xi_j}, \qquad 1\leqq k\leqq 3,$$

and $a_{jk}(\xi,t)$ are the entries of the inverse of $A=((a^{jk}(\xi,t)))$ with a^{jk} given by (2.2).

First we shall verify that (3.1) is uniformly parabolic in the sense of Petrowsky.

LEMMA 3.1. *Let w and ∇_w be as in Theorem* 3.1. *Then*

$$(3.2) \qquad \frac{\partial v}{\partial t}-\nabla_w^2 v = 0 \quad \text{on } Q_T$$

is uniformly parabolic in the sense of Petrowsky.

Proof.

1) From the definition of ∇_w, we have

$$\nabla_w v(\xi,t) = \sum_{j=1}^{3} a_{jk}(\xi,t)\frac{\partial v}{\partial\xi_j}.$$

Thus,

$$\nabla_w^2(v_p) = \sum_{s,j,k=1}^{3} a_{sp}a_{jk}\frac{\partial^2 v_k}{\partial\xi_s\,\partial\xi_j}+a_{sp}\frac{\partial a_{jk}}{\partial\xi_s}\frac{\partial v_p}{\partial\xi_1}, \qquad 1\leqq p\leqq 3.$$

So,

$$\nabla_w^2(v_p) = \sum_{s,j,k=1}^{3} \mathscr{A}_{pk}^{sj}\frac{\partial^2 v_k}{\partial\xi_s\,\partial\xi_j}+a_{sp}\frac{\partial a_{jk}}{\partial\xi_s}\frac{\partial v_p}{\partial\xi_j},$$

with $\mathscr{A}_{pk}^{sj}=a_{sp}a_{jk}$.

2) Equation (3.2) is uniformly parabolic in the sense of Petrowsky on Q_T if there exists $\delta_1>0$ such that

$$\max_j \sup_{|\mu|=1} \operatorname{Re}\lambda_j(\xi,t;\mu)\leqq -\delta_1$$

for all (ξ,t) in Q_T where λ_j are the roots of

$$\det((\mathscr{A}_{pk}^{sj}(i\mu_s)(i\mu_j)-\lambda\delta_{pk}))=0.$$

A simple computation gives

$$\det((\mathscr{A}^{sj}_{pk}(i\mu_s)(i\mu_j)-\lambda\delta_{pk}))=(\lambda+a_{sp}a_{jp}\mu_s\mu_j)^2(\lambda+2a_{sp}a_{jp}\mu_s\mu_j).$$

The roots are

$$\lambda_1=\lambda_2=-a_{sp}a_{jp}\mu_s\mu_j, \qquad \lambda_3=-2a_{sp}a_{jp}\mu_s\mu_j.$$

We have

$$a_{sp}a_{jp}\mu_s\mu_j \geqq \sum_{k=1}^{3}\mu_k^2\Big\{\sum_{r=1}^{3}a_{rp}^2-\sum_{r\neq p}|a_{rp}|\,|a_{rs}|\Big\}.$$

Applying Lemma 2.2, we obtain

$$a_{sp}a_{jp}\mu_s\mu_j \geqq |\mu|^2\{1-6(2c_2\delta)-3(2c_2\delta)^3\}.$$

Hence,

$$\max_j \sup_{|\mu|=1}\operatorname{Re}\lambda_j(\xi,t;\mu)\leqq -2\{1-6(2c_2\delta)-3(2c_2\delta)^3\}\leqq -\delta,$$

if δ is sufficiently small, which we shall always assume.

Proof of Theorem 3.1.

1) Since (3.2) is uniformly parabolic, the existence of a solution v in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ of the initial boundary value problem (3.1) is known. We now establish the estimate of the theorem. We have

$$\begin{aligned}\frac{\partial v}{\partial t}-\nabla^2 v &= f+(\nabla_w^2-\nabla^2)v = f+(\nabla_w-\nabla)(\nabla_w+\nabla)v\\ &= f+(A^{-1}-I)\nabla\{(A+I)\nabla v\}\\ &= f+(A^{-1}-I)\nabla(A+I)\cdot\nabla v+(A^{-1}-I)(A+I)\nabla^2 v = f+g.\end{aligned}$$

Consider the initial boundary value problem

$$\frac{\partial v}{\partial t}-\nabla^2 v=f+g \quad \text{on } Q_T, \quad v=0 \quad \text{on } (0,T)\times\partial\Omega, \quad v(\xi,0)=v_0(\xi) \quad \text{on } \Omega.$$

It follows from a well-known result of the theory of linear parabolic equations that

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\leqq K\{\|f\|_{C^{\alpha,\alpha/2}(Q_T)}+\|g\|_{C^{\alpha,\alpha/2}(Q_T)}+\|v_0\|_{C^{2+\alpha}(\Omega)}\}. \tag{3.3}$$

K is independent of t, v_0, w, v, f and g. It depends on meas (Q_T), e.g., cf. [2, Theorem 3, p. 782] (with some trivial changes in the initial conditions).

2) We now compute $\|g\|_{C^{\alpha,\alpha/2}(Q_T)}$. Applying Lemma 2.1 we obtain

$$\begin{aligned}&\|g\|_{C^{\alpha,\alpha/2}(Q_T)}\\ &\leqq\|(A^{-1}-I)(A^{-1}-I)\nabla^2 v\|_{C^{\alpha,\alpha/2}(Q_T)}+\|(A^{-1}-I)\{\nabla(A^{-1}+I)\}\nabla v\|_{C^{\alpha,\alpha/2}(Q_T)}\\ &\leqq\|A^{-1}-I\|_{C^{\alpha,\alpha/2}(Q_T)}\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\{\|A^{-1}+I\|_{C^{\alpha,\alpha/2}(Q_T)}+\|\nabla(A^{-1}+I)\|_{C^{\alpha,\alpha/2}(Q_T)}\}\\ &\leqq 2c_2\delta\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\{c_3+\|A^{-1}+I\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)}\}.\end{aligned}$$

So,

$$\|g\|_{C^{\alpha,\alpha/2}(Q_T)}\leqq c_4\delta\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}, \tag{3.4}$$

where c_4 is a constant independent of δ, t, v, w.

It follows from (3.3)–(3.4) that

$$\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}\leqq K\{\delta\|v\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)}+\|f\|_{C^{\alpha,\alpha/2}(Q_T)}+\|v_0\|_{C^{2+\alpha}(\Omega)}\}.$$

The theorem is proved.

4. We shall carry out in this section Step 3 of the proof of Theorem 1.1. First let us introduce some notation. We denote

$$X^n(\xi, t) = \xi + \int_0^t v^n(\xi, s)\, ds, \qquad n = 0, 1, 2 \cdots$$

and

$$A_n(\xi, t) = \left(\left(\delta_{jk} + \int_0^t \frac{\partial v_k^n(\xi, s)}{\partial \xi_j}\, ds\right)\right).$$

A_n^{-1} is the inverse of A_n whenever it exists. Set

$$\nabla_n = A_n^{-1}\nabla = \sum_{j=1}^{3} a_{jk}^n(\xi, t)\frac{\partial}{\partial \xi_j}, \qquad 1 \leqq k \leqq 3,$$

where $a_{jk}^n(\xi, t)$ are the entries of A_n^{-1}.

Consider the equations

$$\rho^n(\xi, t) = \rho_0(\xi) \exp\left\{-\int_0^t (\nabla_{n-1} \cdot v^{n-1})(\xi, s)\, ds\right\}, \tag{4.1}$$

and the initial boundary value problems

$$\begin{aligned} &\frac{\partial v^n}{\partial t} - \nabla_{n-1}^2 v^n + \nabla_{n-1}\rho^n - \left\{\frac{\nabla_{n-1}\rho^n \cdot \nabla_{n-1} v^{n-1}}{\rho^n}\right\} = 0 \quad \text{on } Q_T, \\ &v^n(\xi, t) = 0 \quad \text{on } (0, T)\times\partial\Omega, \qquad v^n(\xi, 0) = v_0(\xi) \quad \text{on } \Omega. \end{aligned} \tag{4.2}$$

The main result of the section is the following theorem.

THEOREM 4.1. *Let v_0 be a vector function in $C^{2+\alpha}(\Omega)$ with $v_0 = 0$ on $\partial\Omega$ and let ρ_0 be a scalar function in $C^{1+\alpha}(\Omega)$ with $\rho_0 \geqq c > 0$ on Ω. Then there exist:*

i) *a nonempty interval $(0, T^*)$ independent of n,*

ii) *$\{v^n, \rho^n\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$, a solution of (4.1)–(4.2).*

Moreover, $\{v^n, \rho^n\} \to \{v, \rho\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$.

First we have

LEMMA 4.1. *Let v^0 be a vector function in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ with $v^0(\xi, 0) = v_0(\xi)$ and*

$$(T + T^\beta)\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_T)} \leqq \delta < \frac{1}{8}, \qquad 0 < \beta < \frac{1-\alpha}{2}.$$

Let $\{v_0, \rho_o\}$ be as in Theorem 4.1. Then

1) *$\rho^1(\xi, t)$ defined by (4.1) is in $C^{1+\alpha,(1+\alpha)/2}(Q_t)$ with*

$$\begin{aligned} \|\rho^1\|_{C^{1+\alpha,(1+\alpha)/2}(Q_T)} \leqq K_1\|\rho_0\|_{C^{1+\alpha}(\Omega)}(1 + \|v_0\|_{C^{2+\alpha}(\Omega)})^2 \\ \cdot \exp\{Kt(1 + \|v_0\|_{C^{2+\alpha}(\Omega)})^2\}; \end{aligned} \tag{4.3}$$

2) *there exists $v^1(\xi, t)$, a solution of (4.2) with*

$$\begin{aligned} \|Kv^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \leqq K_2\left(\|\rho_0\|_{C^{1+\alpha}(\Omega)} + \|v_0\|_{C^{2+\alpha}(\Omega)} + 1\right)^3 \\ \cdot \exp\{K_2 t(1 + \|v_0\|_{C^{2+\alpha}(\Omega)})^2\}, \quad \text{for } 0 < t < T. \end{aligned} \tag{4.4}$$

K_1, K_2 are independent of v^0, t.

Proof.

1) It follows from Proposition 2.1 and from our hypotheses on v^0 that A_0^{-1}, the inverse of the matrix

$$\left(\left(\delta_{jk}+\int_0^t \frac{\partial v_k^0(\xi, s)}{\partial \xi_j}\, ds\right)\right)$$

exists. So,

$$\nabla_0 = \sum_{j=1}^{3} a_{jk}^0(\xi, t)\frac{\partial}{\partial \xi_j}, \qquad k = 1, 2, 3$$

has a meaning and

$$\rho^1(\xi, t) = \rho_0(\xi) \exp -\left\{\int_0^t (\nabla_0 \cdot v^0)(\xi, s)\, ds\right\}$$

is defined. We have

$$\|\rho^1\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \leqq \|\rho_0\|_{C^{1+\alpha}(\Omega)} \left\| \exp\left(-\int_0^2 (\nabla_0 \cdot v^0)(\xi, s)\, ds\right)\right\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}.$$

From Proposition 2.3, we get

$$\begin{aligned}
&\left\|\int_0^s (\nabla_0 \cdot v^0)(\xi, s)\, ds\right\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \\
&\quad \leqq \int_0^t \|(\nabla_0 \cdot v^0)(\xi, s)\, ds\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} + t^{(1-\alpha)/2}\|(\nabla_0 \cdot v^0)\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \\
&\quad \leqq (t + t^{(1-\alpha)/2}\|(\nabla_0 \cdot v^0)\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \\
&\quad \leqq CT(1 + \|\nabla v^0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}.
\end{aligned}$$

C is independent of v^0, v_0, ρ_0, t. We have applied Lemma 2.1 to $(\nabla_0 \cdot v^0)$. It is now easy to check that

$$\begin{aligned}
\|\rho^1\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} &\leqq K\|\rho_0\|_{C^{1+\alpha}(\Omega)}(1 + \|\nabla v^0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}) \\
&\quad \cdot \exp\{Kt(1 + \|\nabla v^0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)})\}.
\end{aligned} \tag{4.5}$$

2) Since $v^0(\xi, 0) = v_0(\xi)$ on Ω, we have

$$\nabla v^0 = \nabla v_0 + \nabla v^0 - \nabla v_0.$$

So,

$$\begin{aligned}
\|\nabla v^0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} &\leqq \|\nabla v_0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} + \|\nabla(v^0 - v_0)\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \\
&\leqq \|\nabla v_0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} + t^{(1+\alpha)/2}\|v^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}.
\end{aligned}$$

With our hypothesis on v^0, we get

$$\|\nabla v^0\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \leqq \|v_0\|_{C^{2+\alpha}(\Omega)} + \delta. \tag{4.6}$$

From (4.5)–(4.6) we obtain

$$\begin{aligned}
\|\rho^1\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} &\leqq K\|\rho_0\|_{C^{1+\alpha}(\Omega)}\{1 + \delta + \|v_0\|_{C^{2+\alpha}(\Omega)}\} \\
&\quad \cdot \exp\{Kt(1 + \delta + \|v_0\|_{C^{2+\alpha}(\Omega)})\}, \qquad 0 < t < T.
\end{aligned} \tag{4.7}$$

K is independent of t, v^0, ρ_0, v_0.

3) For $v^1(\xi, t)$ we apply Theorem 3.1. Then,

$$\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \leqq K_2\Big\{\delta\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} + \|v_0\|_{C^{2+\alpha}(\Omega)} + \|\nabla_0\rho^1\|_{C^{\alpha,\alpha/2}(Q_t)} + \left\|\frac{(\nabla_0\rho^1 \cdot \nabla_0 v^0)}{\rho^1}\right\|_{C^{\alpha,\alpha/2}(Q_t)}\Big\}. \tag{4.8}$$

Applying Lemma 2.1 and (4.7) we obtain

$$\|\nabla_0\rho^1\|_{C^{\alpha,\alpha/2}(Q_t)} \leqq C\|\rho_0\|_{C^{1+\alpha}(\Omega)}(1+\|v_0\|_{C^{2+\alpha}(\Omega)}) \exp\{Ct(1+\|v_0\|_{C^{2+\alpha}(\Omega)}\}. \tag{4.9}$$

C is independent of t, δ, v^0, v_0.

Similarly,

$$\begin{aligned}\left\|\frac{(\nabla_0\rho^1 \cdot \nabla_0 v^0)}{\rho^1}\right\|_{C^{\alpha,\alpha/2}(Q_t)} &\leqq \|\nabla_0\rho^1\|_{C^{\alpha,\alpha/2}(Q_t)}\|\nabla_0 v^0\|_{C^{\alpha,\alpha/2}(Q_t)}\left\|\frac{1}{\rho^1}\right\|_{C^{\alpha,\alpha/2}(Q_t)} \\ &\leqq K\|\rho_0\|^2_{C^{1+\alpha}(\Omega)}(1+\|v_0\|_{C^{2+\alpha}(\Omega)})^3 \\ &\quad \cdot \exp\{Kt(1+\|v_0\|_{C^{2+\alpha}(\Omega)})^2\} \quad \text{for } 0<t<T.\end{aligned} \tag{4.10}$$

Combining (4.8)–(4.10) we get the estimate of the lemma

LEMMA 4.2. *Let v_0, ρ_0 be as in Theorem* 4.1 *and v^0 be as in Lemma* 4.1. *Then there exist:*

1) *a nonempty interval $(0, T^*)$ independent of n,*

2) *$\{v^n, \rho^n\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$, a solution of the system* (4.1)–(4.2) *with:*

i) $\|\rho^n\|_{C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})} + \|v^n\|_{C^{2+\alpha,(1+\alpha)/2}(Q_{T^*})} \leqq M$,

ii) $(T^* + (T^*)^\beta\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})} \leqq \delta, \qquad 0<\beta<\dfrac{1-\alpha}{2}$.

M is independent of n.

Proof.

1) From Lemma 4.1 we have $\{v^1, \rho^1\}$ *in* $C^{2+\alpha,(2+\alpha)/2}(Q_T)$.

$$\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \leqq K_2(1+\|\rho_0\|_{C^{1+\alpha}(\Omega)} + \|v_0\|_{C^{2+\alpha}(\Omega)})^3 \exp(K_2 t\{1+\|v_0\|_{C^{2+\alpha}(\Omega)}\}^2).$$

K_2 is independent of t. Thus, there exists a nonempty interval $(0, T^*)$ such that

$$(T^* + (T^*)^\beta)K_2(1+\|\rho_0\|_{C^{1+\alpha}(\Omega)} + \|v_0\|_{C^{2+\alpha}(\Omega)})^3 \exp(K_2 T^*\{1+\|v_0\|_{C^{2+\alpha}(\Omega)}\}^2) \leqq \delta,$$

with $0<\beta<(1-\alpha)/2$. Hence,

$$(T^* + (T^*)^\beta)\|v^1\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})} \leqq \delta, \qquad 0<\beta<\frac{(1-\alpha)}{2}.$$

2) With $0<t<T^*$, we have $\{v^2, \rho^2\}$ and hence, the right-hand sides of (4.3)–(4.4) are independent of v^0; we obtain

$$\|\rho^2\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)} \leqq K_1\|\rho_0\|_{C^{1+\alpha}(\Omega)}(1+\|v_0\|_{C^{2+\alpha}(\Omega)})^2 \exp\{Kt(1+\|v_0\|_{C^{2+\alpha}(\Omega)})^2\},$$

and

$$\|v^2\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \leqq K_2(1+\|\rho_0\|_{C^{1+\alpha}(\Omega)} + \|v_0\|_{C^{2+\alpha}(\Omega)})^3 \exp(K_2 t\{1+\|v_0\|_{C^{2+\alpha}(\Omega)}\}^3).$$

Therefore,

$$(T^*+(T^*)^\beta)\|v^2\|_{C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})}\leqq\delta, \qquad 0<\beta<\frac{1-\alpha}{2},$$

with the same T^* as before.

We may repeat the same argument again and by induction, we get the lemma.

Proof of Theorem 4.1. In view of Lemma 4.2, it remains to show that

$$\{v^n,\rho^n\}\to\{v,\rho\} \quad \text{in } C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})\times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*}).$$

1) We have

$$\begin{aligned}&\frac{\partial}{\partial t}(v^n-v^{n-1})-\nabla_{n-1}^2v^n+\nabla_{n-2}^2v^{n-1}+\nabla_{n-1}\rho^n-\nabla_{n-2}\rho^{n-1}\\ &+\frac{(\nabla_{n-2}\rho^{n-1}\cdot\nabla_{n-2}v^{n-2})}{\rho^{n-1}}-\frac{(\nabla_{n-1}\rho^n\cdot\nabla_{n-1}v^{n-1})}{\rho^n}=0 \quad \text{on } Q_{T^*},\\ &v^n-v^{n-1}=0 \quad \text{on } (0,T^*)=\partial\Omega, \qquad (v^n-v^{n-1})(\xi,0)=0 \quad \text{on } \Omega.\end{aligned} \tag{4.11}$$

Set $w^n=v^n-v^{n-1}$ and we have, by an elementary computation,

$$-\nabla_{n-1}^2v^n+\nabla_{n-2}^2v^{n-1}=-\nabla_{n-2}^2w^n-(\nabla_{n-1}-\nabla_{n-2})(\nabla_{n-1}+\nabla_{n-2})v^n. \tag{4.12}$$

Similarly,

$$\nabla_{n-1}\rho^n-\nabla_{n-2}\rho^{n-1}=\nabla_{n-1}(\rho^n-\rho^{n-1})+(\nabla_{n-1}-\nabla_{n-2})\rho^{n-1}. \tag{4.13}$$

A lengthy but completely elementary computation yields

$$\begin{aligned}&-\frac{(\nabla_{n-1}\rho^n\cdot\nabla_{n-1}v^{n-1})}{\rho^n}+\frac{(\nabla_{n-2}\rho^{n-1}\cdot\nabla_{n-2}v^{n-2})}{\rho^{n-1}}\\ &=\frac{\{(\nabla_{n-1}-\nabla_{n-2})\rho^{n-1}\cdot\nabla_{n-2}v^{n-2}\}}{\rho^{n-1}}-\frac{\{\nabla_{n-1}(\rho^n-\rho^{n-1})\cdot\nabla_{n-1}v^{n-1}\}}{\rho^n}\\ &+\frac{(\rho^n-\rho^{n-1})\nabla_{n-1}\rho^{n-1}\cdot\nabla_{n-1}v^{n-1}}{\rho^n\rho^{n-1}}+\nabla_{n-1}\rho^{n-1}\cdot\frac{(\nabla_{n-2}-\nabla_{n-1})v^{n-2}}{\rho^{n-1}}\\ &+\nabla_{n-1}\rho^{n-1}\frac{\{\nabla_{n-1}(v^{n-2}-v^{n-1})\}}{\rho^{n-1}}.\end{aligned} \tag{4.14}$$

We now apply Lemmas 2.1–2.2 and the estimates of Lemma 4.2. From (4.12) we get

$$\begin{aligned}&\|(\nabla_{n-1}-\nabla_{n-2})(\nabla_{n-1}+\nabla_{n-2})v^n\|_{C^{\alpha,\alpha/2}(Q_t)}\\ &\leqq\|\nabla_{n-1}-\nabla_{n-2}\|_{C^{\alpha,\alpha/2}(Q_t)}\|v^n\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\ &\leqq\mu\|v^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\Big\{\varepsilon\|v^{n-1}-v^{n-2}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\\ &\qquad+c(\varepsilon)\int_0^t\|v^{n-1}-v^{n-2}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}$$

μ, $c(\varepsilon)$ are independent of n, t. Hence,

$$\begin{aligned}&\|(\nabla_{n-1}-\nabla_{n-2})(\nabla_{n-1}+\nabla_{n-2})v^n\|_{C^{\alpha,\alpha/2}(Q_t)}\\ &\leqq K\Big\{\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned} \tag{4.15}$$

With (4.13) we get, by applying Lemmas 2.1–2.2 and the estimates of Lemma 4.2,

$$\begin{aligned}&\|\nabla_{n-1}\rho^{n-1}-\nabla_{n-2}\rho^{n-2}\|_{C^{\alpha,\alpha/2}(Q_t)}\\&\leqq K\Big\{\|\rho^n-\rho^{n-1}\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\quad+\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}\tag{4.16}$$

Applying the same lemmas to (4.14), we obtain

$$\begin{aligned}&\Big\|\frac{(\nabla_{n-2}\rho^{n-1}\cdot\nabla_{n-2}v^{n-2})}{\rho^{n-1}}-(\nabla_{n-1}\rho^n\cdot\rho_{n-1}v^{n-1})\rho^n\Big\|_{C^{\alpha,\alpha/2}(Q_t)}\\&\leqq K\Big\{\|\rho^n-\rho^{n-1}\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\quad+\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}\tag{4.17}$$

It now follows from Theorem 3.1 and from (4.11)–(4.17) that

$$\begin{aligned}\|w^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}&\leqq K\Big\{\delta\|w^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+\|\rho^n-\rho^{n-1}\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\quad+\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}\tag{4.18}$$

The different constants K are all independent of n, t, δ, ε.

2) We also have

$$\rho^n-\rho^{n-1}=\rho_0\Big\{\exp\Big(-\int_0^t\nabla_{n-1}\cdot v^{n-1}\,ds\Big)-\exp\Big(-\int_0^t\nabla_{n-2}\cdot v^{n-2}\,ds\Big)\Big\}.$$

So,

$$\begin{aligned}&\|\rho^n-\rho^{n-1}\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\leqq\Big\|\int_0^t(\nabla_{n-1}\cdot v^{n-1}-\nabla_{n-2}\cdot v^{n-2})\,ds\Big\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\quad\cdot\exp\Big(\Big\|\int_0^t\nabla_{n-1}\cdot v^{n-1}\,ds\Big\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}+\Big\|\int_0^t\nabla_{n-2}\cdot v^{n-2}\,ds\Big\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\Big).\end{aligned}$$

An elementary computation as in the first part yields

$$\begin{aligned}&\|\rho^n-\rho^{n-1}\|_{C^{1+\alpha,(1+\alpha)/2}(Q_t)}\\&\leqq K\Big\{\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}\tag{4.19}$$

3) Combining (4.18)–(4.19) we get

$$\begin{aligned}&(1-K\delta)\|w^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}\\&\leqq K\Big\{\varepsilon\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}+c(\varepsilon)\int_0^t\|w^{n-1}\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds\Big\}.\end{aligned}\tag{4.20}$$

The different K are all independent of n, ε, t, δ.

Take ε so that $K\varepsilon \leqq 1-K\delta$, then summing from $n=1$ to N, we obtain

$$\begin{aligned} A_N(t) &= (1-K\delta)\sum_{n=1}^{N} \|w^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \\ &\leqq (1-K\delta)\|w^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)} \\ &\quad + c(\varepsilon)\int_0^t \|w^0\|_{C^{2+\alpha,(2+\alpha)/2}(Q_s)}\,ds + c(\varepsilon)\int_0^t A_N(s)\,ds. \end{aligned} \tag{4.21}$$

It follows that $A_\infty(t)$ is bounded and hence, the series $\sum_{n=1}^{\infty} \|w^n\|_{C^{2+\alpha,(2+\alpha)/2}(Q_t)}$ converges for $0<t<T^*$. Therefore $v^n \to v$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})$.

From (4.19), we get $\rho^n - \rho$ in $C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$. The theorem is proved.

5. We shall carry out the proof of Theorem 1.1 in this section. First we have

THEOREM 5.1. *Let $\{v_0, \rho_o\}$ be as in Theorem* 4.1. *Then there exist*:

1) *a nonempty interval* $(0, T^*)$,

2) $\{v, \rho\}$ *in* $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$ *such that*

$$\rho(\xi, t) = \rho_0(\xi)\exp\left\{-\int_0^t (\nabla_v \cdot v)(\xi, s)\,ds\right\}, \tag{5.1}$$

and

$$\begin{aligned} &\frac{\partial v}{\partial t} - \nabla_v^2(v) + \nabla_v \rho - \frac{(\nabla_v \rho \cdot \nabla_v v)}{\rho} = 0 \quad \text{on } Q_{T^*}, \\ &v(\xi, t) = 0 \quad \text{on } (0, T^*)\times\partial\Omega, \qquad v(\xi, 0) = v_0(\xi) \quad \text{on } \Omega. \end{aligned} \tag{5.2}$$

∇_v *is the operator* $A^{-1}\nabla = \sum_{j=1}^{3} a_{jk}(\xi, t)\partial/\partial\xi_j$, $1<k<3$ *where* $a_{jk}(\xi, t)$ *are the inverse of the matrix* $A = ((\delta_{jk} + \int_0^t (\partial v_k(\xi, s))/\partial\xi_j)\,ds))$.

Moreover, $\{v, \rho\}$ *is unique.*

Proof.

1) Let v^n and ρ^n be as in Theorem 4.1. We have $\{v^n, \rho^n\} \to \{v, \rho\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*})$. From (4.1)–(4.2) and the above properties, it is easy to see that $\{v, \rho\}$ is a solution of (5.1)–(5.2). Indeed, a proof as that of Theorem 4.1 gives the desired result.

2) The solution $\{v, \rho\}$ obtained is unique. Suppose that $\{v_1, \rho_1\}$ are two solutions of (5.1)–(5.2). Set $v = v_1 - v_2$, $\rho = \rho_1 - \rho_2$. Then from (4.19)–(4.20) we deduce that $v = 0 = \rho$ since $v(\xi, 0) = \rho(\xi, 0) = 0$.

The theorem is proved.

Proof of Theorem 1.1.

1) Let u be in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$ with $u = 0$ on $(0, T)\times\partial\Omega$. Consider the ordinary differential equation

$$\begin{aligned} &\frac{d}{ds}Y(s; x, t) = u(Y(s; x, t), s), \\ &Y(t; x, t) = x \end{aligned} \tag{5.3}$$

for every (x, t) in Q_T and $0 \leqq s \leqq t$.

Since u is in $C^{2+\alpha,(2+\alpha)/2}(Q_T)$, there exists a unique solution curve passing through (s, t). Set

$$Y(0; x, t) = \xi.$$

Then the mapping $(x, t) \to (\xi, t)$ is a one-to-one mapping of Q_T onto Q_T and of $\partial\Omega \times (0, T)$ onto $\partial\Omega \times (0, T)$. The inverse of that mapping we denote by $X(\xi, t) = x$. For $u(x, t)$, $h(x, t)$ we set

$$v(\xi, t) = u(X(\xi, t), t), \qquad \rho(\xi, t) = h(X(\xi, t), t).$$

Equation (5.3) implies that

$$\frac{d}{ds} X(\xi, s) = v(\xi, s), \qquad X(\xi, 0) = \xi. \tag{5.4}$$

Thus,

$$x = X(\xi, t) = \xi + \int_0^t v(\xi, s)\, ds.$$

An elementary computation shows that (0.1)–(0.2) may be rewritten as

$$\frac{\partial \rho}{\partial t}(\xi, t) + \nabla_v \cdot v = 0 \quad \text{on } (0, T) \times \Omega, \qquad \rho(\xi, 0) = \rho_0(\xi) = h_0(x) \quad \text{on } \Omega, \tag{5.5}$$

and

$$\begin{gathered} \frac{\partial v}{\partial t} - \nabla_v^2 v + \nabla_v \rho - \frac{(\nabla_v \rho \cdot \nabla_v v)}{\rho} = 0 \quad \text{on } Q_T, \\ v(\xi, t) = 0 \quad \text{on } (0, T) \times \partial\Omega, \qquad v(\xi, 0) = v_0(\xi) = u_0(x) \quad \text{on } \Omega, \end{gathered} \tag{5.6}$$

with ∇_v as in Theorem 5.1.

2) From Theorem 5.1 we have a unique solution $\{v, \rho\}$ in $C^{2+\alpha,(2+\alpha)/2}(Q_{T^*}) \times C^{1+\alpha,(1+\alpha)/2}(Q_{T^*})$, a solution of (5.5)–(5.6). Then with $u(x, t) = v(X^{-1}(x, t), t)$ and $h(x, t) = (X^{-1}(x, t), t)$ we get $\{u, h\}$ as the unique solution of (0.1)–(0.2) with all the stated properties.

REFERENCES

[1] Y. Y. BELOV AND N. N. YANENKO, *Influence of viscosity on the smoothness of solutions of incompletely parabolic systems*, Math. Notes, 10 (1971), pp. 480–483.

[2] A. FRIEDMAN, *Boundary estimates for second-order parabolic equations and their applications*, J. Math. Mech., 7 (1958), pp. 771–791.

[3] B. GUSTAFSSON AND A. SUNDSTROM, *Incompletely parabolic problems in fluid dynamics*, SIAM J. Appl. Math., 35 (1978), pp. 343–357.

[4] N. ITAYA, *On the Cauchy problem for the system of fundamental equations describing the movement of compressible viscous fluid*, Kodai Math. Sem. Report, 23 (1971), pp. 60–120.

[5] J. NASH, *Le probleme de Cauchy pour les équations differentielles d'un fluide general*, Bull. Soc. Math. France, 90 (1962), pp. 487–497.

[6] V. A. SOLONNIKOV, *Solvability of a problem on the motion of a viscous incompressible fluid bounded by a free surface*, Izv. Akad. Nauk SSSR, 416 (1977), =Math. USSR Izv., 11 (1977), pp. 1323–1357.

[7] A. TANI, *On the first initial boundary value problem of compressible viscous fluid motion*, Publ. Res. Inst. Math. Sci., Kyoto University, 13 (1977), pp. 193–253.

[8] B. A. TON, *Initial boundary value problems for the equations of the theory of shallow waters*, to appear.

[9] ———, *On the initial value problem for compressible fluid flows with vanishing viscosity*, Kodai Math. J., 3 (1980).

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0010 $01.00/0

ON DIRICHLET'S PROBLEM FOR ELLIPTIC EQUATIONS IN SECTIONALLY SMOOTH n-DIMENSIONAL DOMAINS. II.*

A. AZZAM†

Abstract. In a recent paper, [SIAM J. Math. Anal., 11 (1980), pp. 248–253] we studied the Dirichlet problem for elliptic equations in sectionally smooth domains. Conditions sufficient for the solution to be of class $C_\nu(1<\nu<2)$ were given. In the present note this result is improved.

In [1] we studied the Dirichlet problem for the uniformly elliptic equation

$$a_{ij}(x)u_{|ij}+a_i(x)u_{|i}+a(x)u=f(x), \tag{1}$$

in a domain $\Omega\subset R^n$, $n\geqq 2$ with sectionally smooth boundary Γ. Γ consists of $(n-1)$-dimensional surfaces $\Gamma_1,\cdots,\Gamma_k$ of class $C_{2+\alpha}$, $0<\alpha<1$. For simplicity we take $k=2$ and write $\Gamma_1\cap\Gamma_2=S$. Let P be any point on S. When $a_{ij}(P)u_{|ij}=0$ is transformed to canonical form, the angle between Γ_1 and Γ_2 at P will be changed to $\omega(P)$. In [1] we considered the case $\omega(P)<\pi$, and obtained $C_\nu(\bar{\Omega})$ statements for the solution, $1<\nu<2$. In this note, we introduce a new barrier function (cf. [1, Lemma 1]) which enables us to improve the result of [1, Thm. 1] as follows.

THEOREM 1. *Let a_{ij}, a_i, a and f in* (1) *be of class $C_\alpha(\bar{\Omega})$. If the boundary value of u is continuous on Γ and of class $C_{2+\alpha}(\Gamma\backslash S)$, then $u\in C_\nu(\Omega)$, where $\nu=\min(2,\pi/\omega-\varepsilon)$, $\omega=\max_{P\in S}\omega(P)$ and $\varepsilon>0$ is arbitrarily small.*

This result may be accomplished by replacing the barrier function in Lemma 1 by

$$V(x)=-Mr^\nu\cos\lambda\left(\theta-\frac{\pi}{2}\right),$$

where $\nu=\min(2,\pi/\omega-\varepsilon)<((\pi-2\delta)/\omega)=\lambda$, $\varepsilon>0$ arbitrarily small and $\delta>0$ suitably chosen. Note that taking $\beta=(\pi-\omega)/2$ we have $\cos\lambda(\theta-\pi/2)\geqq\sin\delta$ for $\beta\leqq\theta\leqq\omega+\beta$. The rest of the proof of Lemma 1 remains almost unchanged. The proofs of Lemma 2 and Theorem 2 remain unchanged if $\nu\geqq 1$. If $\nu<1$, Lemma 2 is not needed, and in the proof of Theorem 2, we deal with the function u and its Hölder coefficient $H_\nu(u)$ rather than u' and $H_{\nu-1}(u')$.

REFERENCE

[1] A. AZZAM, *On Dirichlet's problem for elliptic equations in sectionally smooth n-dimensional domains*, this Journal, 11 (1980), pp. 248–253.

* Received by the editors April 28, 1980.
† Department of Mathematics, University of Windsor, Windsor, Ontario N9B 3P4, Canada.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0011 $01.00/0

LINEAR DECOMPOSABLE SYSTEMS IN CONTINUOUS TIME*

STEPHEN J. HEGNER†

Abstract. With the aid of mathematical tools from category theory and functional analysis, an algebraic approach to the theory of continuous-time linear systems is presented. The dynamics of these systems are described by infinitely differentiable semigroups of operators. A method of obtaining the canonical input–output behavior of a system is given, based upon the concepts of free and cofree objects in category theory. Reachability and observability are discussed using the categorical concept of image-factorization system. It is shown that in the general case of infinite-dimensional systems, there are multiple concepts of reachability and observability, one pair of such concepts for each distinct image-factorization system. In particular, there is a distinct concept of canonical realization for each image-factorization system. Some special results on the nature of the reachability map, observability map, and state space of systems whose input and output spaces are finite dimensional are also developed.

Introduction. In Arbib and Manes (1974), the framework of decomposable systems was introduced. This framework, which uses only very elementary concepts from category theory, captures discrete-time linear systems, as well as some other types of systems such as group machines, in such a way that the key concepts arise as natural categorical constructions. It is the purpose of this paper to develop an analogous approach for continuous-time systems.

As in the discrete-time approach, the heart of the continuous-time system is its dynamics. In the continuous-time case, the one-step state-transition dynamics is replaced by a differentiable dynamics. The precise characterization of such dynamics is the differentiable semigroup. The full theory of these dynamics is developed in § 1.

In the discrete-time case of linear systems, the natural structure for the inputs and outputs over time is well-known. However, the natural structure of the inputs over time in the case of group machines was not well-known until a framework akin to decomposable systems was used (see Arbib (1973)). Similarly, the natural structure for inputs over time in continuous-time systems was not known, and various structures have been used. In § 2, categorical techniques are employed to show what this natural structure must be. The natural structure for outputs is also developed.

Central concepts in system theory are reachability and observability. They were captured for decomposable systems in Arbib and Manes (1974) by using the categorical concept of an image-factorization system. In the category of vector spaces and linear maps, there is only one such system, and so only one concept each of reachability and observability. However, in the case of continuous-time systems, although the use of image-factorization systems is almost identical, there is more than one such system. Hence, there is more than one natural concept each of reachability and observability. In § 3, these ideas are developed in detail.

Of particular interest in linear system theory are those systems whose input and output spaces are finite dimensional. Some special properties of the reachability map, the observability map, and the state space of such systems are developed in § 4. Finally, § 5 contains a brief comparison of the present work to other work on the general theory of continuous-time linear systems.

* Received by the editors March 3, 1978, and in final revised form July 13, 1980. This work is based in part upon the author's doctoral dissertation at the University of Massachusetts at Amherst, which was supported by the National Science Foundation under grant DCR72-03733 A01.

† Department of Applied Mathematics and Computer Science, Thornton Hall, University of Virginia, Charlottesville, Virginia 22901.

Every effort has been made to keep the amount of knowledge of category theory required for an understanding of this paper to a minimum. An understanding of that provided in Arbib and Manes (1974) should prove sufficient. However, the reader is referred to Arbib and Manes (1975), Herrlich and Strecker (1973), or Schubert (1972) for help whenever necessary.

While the amount of functional analysis required for understanding this paper is necessarily not minimal, an attempt has been made to isolate technical details as lemmas and to explain intuitively some of the crucial constructions, so that the gist of the theory may be comprehended without full understanding of the proofs.

Since the system-theoretic structures used in this paper are based upon those developed in Arbib and Manes (1974), it is assumed that the reader is familiar with that paper. Frequently, its concepts will be used to motivate the presentation given here.

0. Terminology and notation. In this section, some of the terminology and notation used throughout the report is gathered. It is not meant so much to be read as to be referenced when a question arises.

General notation. $\mathbf{R}_+$ denotes the nonnegative reals, with the usual topology. $\mathbf{Q}_+$ denotes the nonnegative rationals. $\mathbf{N}$ denotes the natural numbers.

Locally convex spaces. Schaefer (1971), Köthe (1969) and Treves (1967) should be consulted as general references for locally convex spaces.

$\mathbf{K}$ denotes either the field $\mathbf{R}$ of real numbers or the field $\mathbf{C}$ of complex numbers, each with its usual topology. $\mathbf{K}$ is to be fixed in any particular context. l.c.s. is an abbreviation for locally convex, separated, topological vector space over $\mathbf{K}$. If the topology $\mathcal{T}$ of the l.c.s. E must be explicitly indicated, the notation $E[\mathcal{T}]$ is used. $\mathcal{U}(E)$ denotes the set of all convex neighborhoods of 0 of the l.c.s. E.

A subset U of a l.c.s. E is called a *barrel* if it is closed, absolutely convex, and absorbing; U is called *bornivorous* if it is convex and it absorbs every bounded subset of E. E is called *barreled* (resp. *quasi-barreled*, resp. *bornological*) if every barrel, (resp. bornivorous barrel, resp. bornivorous set) is in $\mathcal{U}(E)$.

E' denotes the (continuous) dual of the l.c.s. E. $\langle E, F\rangle$ denotes that E and F form a dual pair which separates points. Polars in dual pairs are denoted by the symbol $^\circ$.

A continuous linear map $f: E \to F$ of l.c.s.'s is called a *dense map* if $\overline{f(E)} = F$ (overbar denotes closure), a *homomorphism* if it transforms neighborhoods of 0 in E into neighborhoods of 0 in $f(E)$, a *quotient map* if it is a surjective homomorphism, a *near quotient* if $\overline{f(U)} \in \mathcal{U}(F)$ whenever $U \in \mathcal{U}(E)$, an *embedding* if it is an injective homomorphism, and *closed* if $f(E)$ is closed in F.

The completion of the l.c.s. E is denoted $\hat{E}$. If $f: E \to F$, the extension of f to completions is denoted $\hat{f}: \hat{E} \to \hat{F}$. A l.c.s. is *quasi-complete* if each of its closed, bounded subsets is complete. The quasi-completion of E is denoted $\hat{E}$, and the extension of $f: E \to F$ is denoted $\hat{f}: \hat{E} \to \hat{F}$.

The following notations for categories of spaces will be used. **LS** denotes the category of all vector spaces over $\mathbf{K}$, with linear maps as morphisms. **LCS** denotes the category whose objects are the l.c.s.'s over $\mathbf{K}$, and whose morphisms are the continuous linear maps. **CS** (resp. **QC**) denotes the full subcategory of **LCS** consisting of the complete (resp. quasi-complete) l.c.s.'s.

Special locally convex spaces. An (F) *space* is a l.c.s. which is metrizable and complete (also called a Fréchet space). A (DF) *space* is a l.c.s. which admits a fundamental sequence of bounded sets, and for which every bounded subset of its strong dual, which is the union of a sequence of equicontinuous sets, is equicontinuous.

The strong dual of an (F) (*resp.* (DF)) *space is a* (DF) (*resp.* (F)) *space.* A (B) space is complete-l.c.s. whose topology is definable by a single norm. A Banach space has a specific norm associated with it, and is not to be confused with a (B) space. *A l.c.s. is a* (B) *space if and only if it is simultaneously an* (F) *space and a* (DF) *space.* A (S) *space* (also called a Schwartz space) is a l.c.s. for which every continuous linear map from it into a (B) space is compact.

Spaces of linear mappings. Let E and F be l.c.s.'s. $L(E, F)$ denotes the linear space of all continuous linear maps from E into F. Let $\mathfrak{S}$ be a set of bounded subsets of E which *covers* E (i.e., $\cup\mathfrak{S} = E$). On $L(E, F)$, the *topology of* $\mathfrak{S}$ *convergence* is defined to have as fundamental neighborhoods of 0, sets of the form $\{f \mid f(A) \subset U\}$ where $A \in \mathfrak{S}$ and $U \in \mathcal{U}(F)$. $L(E, F)$ with this topology is denoted $L_{\mathfrak{S}}(E, F)$, and the neighborhood $\{f \mid f(A) \subset U\}$ is denoted $\mathcal{N}(A, U)$. When $\mathfrak{S}$ = all finite sets (resp. $\mathfrak{S}$ = all bounded sets), the topology of $\mathfrak{S}$ convergence is called the topology of *simple* (resp. *bounded*) *convergence*; $L_{\mathfrak{S}}(E, F)$ is denoted $L_s(E, F)$ (resp. $L_b(E, F)$). When $E = F$, $L(E, F)$ (resp. $L_{\mathfrak{S}}(E, F)$) is denoted $L(E)$ (resp. $L_{\mathfrak{S}}(E)$).

Bilinear mappings and tensor products. For detailed special information on bilinear mappings and tensor products, refer to Grothendieck (1955) and Treves (1967).

Let E, F, and G be l.c.s.'s, and let $f: E \times F \to G$ be a bilinear map. For the purposes of this paper, f is called *hypocontinuous* if it is hypocontinuous with respect to the bounded sets of each space, i.e., for each $V \in \mathcal{U}(G)$, $A \subset E$ bounded, there is a $U \in \mathcal{U}(F)$ such that $f(A \times U) \subset V$, and for each $V \in \mathcal{U}(G)$, $B \subset F$ bounded, there is a $W \in \mathcal{U}(E)$ such that $f(W \times B) \subset V$. A family of *equihypocontinuous* maps is defined similarly.

$E \otimes F$ denotes the tensor product of the l.c.s.'s E and F. On $E \otimes F$ there are two topologies of importance. The π (or *projective*) *topology* is the strongest locally convex topology making the canonical bilinear map $p: E \times F \to E \otimes F$ continuous, and is denoted $E \otimes_\pi F$. The β (or *hypocontinuous*) *topology* is the strongest locally convex topology making p hypocontinuous and is denoted $E \otimes_\beta F$.

Differentiable functions and distributions. The theory used here is essentially that developed by Schwartz (1954–55), (1957, 59), and (1966).

$\mathscr{E}^m(\mathbf{R}_+, E)$ ($m \in \mathbf{N}$ or $m = \infty$) denotes the space of all m-times continuously-differentiable functions on $\mathbf{R}_+$ with values in the l.c.s. E. The differentiation operator on this space is always denoted by the symbol D. The topology of $\mathscr{E}^m(\mathbf{R}_+, E)$ is defined by the family of seminorms of the form $\varphi \mapsto \sup_{\substack{t \in K \\ q \leqq p}} \alpha(D^q \varphi(t))$ where α is a continuous seminorm on E, $K \subset \mathbf{R}_+$ is compact, and $p \in \mathbf{N}$ with $p \leqq m$. If $m = \infty$, $\mathscr{E}^m(\mathbf{R}_+, E)$ is denoted simply $\mathscr{E}(\mathbf{R}_+, E)$.

A useful tool is the mean-value theorem for $\mathscr{E}^1(\mathbf{R}_+, E)$, which may be stated as follows.

THEOREM 0.1. *Let E be a l.c.s., $f \in \mathscr{E}^1(\mathbf{R}_+, E)$, α a continuous seminorm on E, and $I = [a, b] \subset \mathbf{R}_+$ a compact, connected interval with $b > a$.*

(a) $\alpha(f(b) - f(a)) \leqq (b - a) \sup_{t \in I} \alpha(Df(t))$,

(b) $\alpha[(f(b) - f(a))/(b - a) - Df(a)] \leqq \sup_{t \in I} \alpha(Df(t) - Df(a))$.

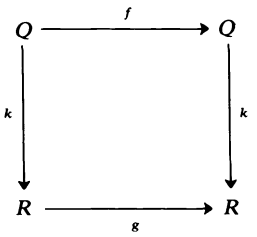
Proof. The proof is essentially the same as the case in which E is a normed space. See, for example, Dieudonné (1960, 8.5.2 and 8.6.2). □

The space $\mathscr{E}(\mathbf{R}_+, \mathbf{K})$ is denoted $\mathscr{E}(\mathbf{R}_+)$. Let $\mathscr{E}(\mathbf{R})$ denote the space of all infinitely differentiable functions from $\mathbf{R}$ to $\mathbf{K}$ with the topology of uniform convergence of each

derivative on compact sets (analogous to the topology on $\mathscr{E}(\mathbf{R}_+)$), and let $\mathscr{E}_-(\mathbf{R})$ denote the closed subspace of $\mathscr{E}(\mathbf{R})$ consisting of those functions which are identically 0 on $]-\infty, 0]$. There is a canonical injection $i: \mathscr{E}(\mathbf{R})/\mathscr{E}_-(\mathbf{R}) \to \mathscr{E}(\mathbf{R}_+)$, given by sending a representative of an equivalence class to its restriction to $[0, \infty[$. This injection is clearly an embedding, and furthermore a surjection (hence an isomorphism), since any $f \in \mathscr{E}(\mathbf{R}_+)$ can be extended to an $\bar{f} \in \mathscr{E}(\mathbf{R})$. (Use Borel's theorem (Treves (1967, Thm. 38.1)) to show this.) Now $\mathscr{E}(\mathbf{R})$ is an (F) and (S) space; hence, $\mathscr{E}(\mathbf{R}_+)$ is also an (F) and (S) space, the strong dual $\mathscr{E}'(\mathbf{R}_+)$ of $\mathscr{E}(\mathbf{R}_+)$ can be identified with $\mathscr{E}_-(\mathbf{R})^\circ$ for the pair $\langle \mathscr{E}(\mathbf{R}), \mathscr{E}'(\mathbf{R}) \rangle$, and $\mathscr{E}'(\mathbf{R}_+)$ is furthermore (DF), (S), bornological, complete (hence barreled), and reflexive (see Grothendieck (1973, Ch. 4, Pt. 4)). As $\mathscr{E}'(\mathbf{R})$ is the space of all distributions on $\mathbf{R}$ with compact support, $\mathscr{E}'(\mathbf{R}_+)$ is the subspace of all distributions which are 0 on each $f \in \mathscr{E}_-(\mathbf{R})$, i.e., those distributions whose support is compact and contained in $[0, \infty[$. The differentiation operator on $\mathscr{E}'(\boldsymbol{R}_+)$ will be denoted by D in lieu of the more cumbersome D'.

1. Differentiable semigroups.

Basic theory. In the discrete-time case of decomposable systems of Arbib and Manes (1974), a central concept is system dynamics. It will be recalled here briefly. Let $\mathscr{K}$ be any category. A *system dynamics* in $\mathscr{K}$ is a pair (Q, f) where Q is a $\mathscr{K}$ object and $f: Q \to Q$ is a $\mathscr{K}$ morphism. A $\mathscr{K}$ morphism $k: Q \to R$ is called a *dynamorphism* for the system dynamics (Q, f) and (R, g) provided that the diagram



commutes. $k: (Q, f) \to (R, g)$ is written to denote this fact. System dynamics in $\mathscr{K}$ and dynamorphisms form a category, denoted Dyn $(\mathscr{K})$.

In the discrete-time interpretation of the system dynamics (Q, f), f is the one-step state-transition map for a system. More precisely, a *decomposable system* in $\mathscr{K}$ is a 6-tuple $M = (Q, f, I, g, Y, h)$ such that (Q, f) is a system dynamics (Q is called the *state space* and f the *state-transition map*), I is a $\mathscr{K}$ object and $g: I \to Q$ a $\mathscr{K}$ morphism (called the *input space* and *input map*, respectively), and Y is a $\mathscr{K}$ object and $h: Q \to Y$ a $\mathscr{K}$ morphism (called the *output space* and *output map*, respectively). In the case that $\mathscr{K}$ is a subcategory of **LS**, the system may be thought of as described by the equations

$$\begin{aligned} q(t+1) &= f(q(t)) + g(i(t)), \\ y(t) &= h(q(t)). \end{aligned} \tag{1}$$

Consult Arbib and Manes (1974) for a complete discussion.

To convert the above equations to continuous time, the one-step transition must be replaced by an infinitesimal transition. That is, the system must now be thought of as described by the equations

$$\begin{aligned} \frac{dg(t)}{dt} &= f(q(t)) + g(i(t)), \\ y(t) &= h(q(t)). \end{aligned} \tag{2}$$

However, it is not quite that simple, because $dq(t)/dt$ does not make sense in an arbitrary subcategory of **LS**, since an arbitrary vector space has no natural topology. Furthermore, it must be guaranteed that f is nice enough so that the above differential equation may be solved. It is this problem of characterizing continuous-time dynamics that is next investigated.

Let E be a l.c.s. A map $T: \mathbf{R}_+ \to L(E)$ is called a *differentiable semigroup* (abbreviated d.s.g.) on E if it has the following four properties:

(s_1) $T(0) = 1_E$, the identity map on E.

(s_2) $T(s+t) = T(s) \circ T(t)$ for each $s, t \in \mathbf{R}_+$.

(s_3) T is pointwise differentiable, i.e., $\lim_{t \to 0} [(T(t)(e) - e)/t]$ exists for each $e \in E$.

(s_4) $\{T(t) \mid 0 \leqq t \leqq \varepsilon\}$ is equicontinuous for some $\varepsilon > 0$.

The function $\mathscr{g}_T \in L(E)$ defined by $\mathscr{g}_T(e) = \lim_{t \to 0} [(T(t)(e) - e)/t]$ is called the *infinitesimal generator* of T.

Semigroups of this form, which also satisfy the condition that $\{(T(t) - 1)/t \mid 0 < t \leqq \varepsilon\}$ is equicontinuous for some $\varepsilon > 0$, were previously studied by Waelbroeck (1964). This condition is not enforced here because it is not necessary.

Let $\mathscr{K}$ be any subcategory of **LCS**. The category of *differentiable system dynamics* in $\mathscr{K}$, denoted D-Dyn ($\mathscr{K}$), is the full subcategory of Dyn ($\mathscr{K}$) whose objects are pairs (Q, f) such that f is the infinitesimal generator of some d.s.g. The morphisms of D-Dyn ($\mathscr{K}$) are thus just the dynamorphisms. A *differentiable decomposable system* in $\mathscr{K}$ is a decomposable system $M = (Q, f, I, g, Y, h)$ in $\mathscr{K}$ such that (Q, f) is a differentiable system dynamics. A differentiable decomposable system may be thought of as governed by (2) above. In order that this internal system description be useful, it is necessary to know that the differential equation is solvable, i.e., that $dq(t)/dt = f(q(t))$ has a unique solution for each initial condition.

Let E be a l.c.s. Recall that $D: \mathscr{E}^1(\mathbf{R}_+, E) \to \mathscr{E}^0(\mathbf{R}_+, E)$ is the differentiation operator. By a *linear differential equation* on E is meant an equation of the form

$$Du(t) = A(u(t)),$$

where $A \in L(E)$. A *solution* to this equation with initial condition $u(0) = x (x \in E)$ is an $f \in \mathscr{E}^1(\mathbf{R}_+, E)$ such that $f(0) = x$, and for each $t \in \mathbf{R}_+$, $Df(t) = A(f(t))$.

PROPOSITION 1.1. *Let E be a l.c.s., and let T be a d.s.g. on E.*

(a) $T \in \mathscr{E}(\mathbf{R}_+, L_s(E))$.

(b) $D^p T(t) = (\mathscr{g}_T)^p \circ T(t) = T(t) \circ (\mathscr{g}_T)^p$, *for each* $p \in \mathbf{N}$.

(c) *For every $e \in E$, the map from $\mathbf{R}_+$ to E given by $t \mapsto T(t)e$ is in $\mathscr{E}(\mathbf{R}_+, E)$.*

(d) *The canonical map $\Lambda_{T,E}: E \to \mathscr{E}(\mathbf{R}_+, E)$ given by $x \mapsto (t \to T(t)x)$ is continuous.*

Proof. Note that (s_3) is equivalent to $T \in \mathscr{E}^1(\mathbf{R}_+, L_s(E))$. Hence, (a) follows from (b), which is routinely verified. (c) is an immediate consequence of (a). To show (d), let $V \in \mathscr{U}(\mathscr{E}(\mathbf{R}_+, E))$. A fundamental such V is of the form $\{\varphi \mid \sup_{\substack{t \in K \\ q \leqq p}} \alpha(D^q(\varphi(t)) \leqq \varepsilon\}$, with α a continuous seminorm on E, $p \in \mathbf{N}$, $K \subset \mathbf{R}_+$ compact, and $\varepsilon > 0$. Let $U = \{x \in E \mid \sup_{\substack{t \in K \\ q \leqq p}} \alpha(D^q T(t)x) \leqq \varepsilon\}$. Clearly $\Lambda_{T,E}(U) \subset V$, so it suffices to show that $U \in \mathscr{U}(E)$. By (s_4), there is a $\gamma > 0$ such that $\{T(t) \mid 0 \leqq t \leqq \gamma\}$ is equicontinuous. Hence, it follows that $\{D^q T(t) \mid q \leqq p \text{ and } t \in K\}$ is also equicontinuous, so there is a continuous seminorm α on E such that for any $x \in E$, if $\beta(x) \leqq 1$, then $\alpha(D^q T(t)x) \leqq \varepsilon$ for any $q \leqq p$ and $t \in K$. Hence, $\{x \mid \beta(x) \leqq 1\} \subset U$, so $U \in \mathscr{U}(E)$. □

THEOREM 1.2. *Let E be a l.c.s., T a d.s.g. on E, and $x \in E$. The differential equation $Du(t) = \mathscr{g}_T(u(t))$ has a unique solution with $u(0) = x$, which is given by $t \mapsto T(t)x$.*

Proof. Clearly $t \mapsto T(t)x$ is a solution of $Du(t) = g_T(u(t))$ with $u(0) = x$. It remains to verify the uniqueness of the solution. Let $g \in \mathscr{C}^1(\mathbf{R}_+, E)$ be a solution with $g(0) = x$. Let $y \in \mathbf{R}_+ \backslash \{0\}$ and define $g^y : [0, y] \to E$ by $t \mapsto T(y-t)g(t)$. For $t \in [0, y]$, $t+h \in [0, y]$,

$$\begin{aligned}\frac{g^y(t+h)-g^y(t)}{h} &= \frac{T(y-(t+h))g(t+h)-T(y-t)g(t)}{h} \\ &= \frac{(T(y-(t+h))-T(y-t))g(t)}{h} + \frac{T(y-t)(g(t+h)-g(t))}{h} \\ &\quad + (T(y-(t+h))-T(y-t))(Dg(t)) \\ &\quad + (T(y-(t+h))-T(y-t))\left(\frac{g(t+h)-g(t)}{h} - Dg(t)\right).\end{aligned}$$

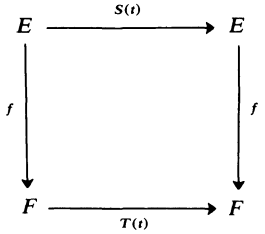
As $h \to 0$, the first term tends to $-(g_T \circ T(y-t))g(t)$, while the second term tends to $T(y-t)Dg(t)$. The third term approaches 0 as $h \to 0$, since $t \mapsto T(y-t)Dg(t)$ is continuous (use Proposition 1.1(c)). To show that the last term goes to 0 as $h \to 0$, choose any $V \in \mathscr{U}(E)$. It suffices to show that this term lies in V for all sufficiently small h. Using the equicontinuity of $\{T(y-(t+h))-T(y-t) \mid 0 \leqq t \leqq \varepsilon\}$ for any $\varepsilon > 0$ (follows from (s_4)) and the fact that $(g(t+h)-g(t))/h - Dg(t) \to 0$ as $h \to 0$, there is an $\varepsilon > 0$ and $U \in \mathscr{U}(E)$ such that for all h, $0 < h \leqq \varepsilon$, $(g(t+h)-g(t))/h - Dg(t) \in U$ and $(T(y-(t+h))-T(y-t))U \subset V$. Thus, the last term can be made arbitrarily small by choosing h small enough. Hence,

$$\begin{aligned}Dg^y(t) &= -(DT(y-t))g(t) + T(y-t)Dg(t) \\ &= -T(y-t) \circ g_T g(t) + T(y-t)Dg(t).\end{aligned}$$

Thus, $Dg^y(t) = 0$, since $g_T g(t) = Dg(t)$. Now by Theorem 0.1, for any seminorm α continuous on E, $\alpha(g^y(y)-g^y(0)) \leqq \sup_{0 \leqq z \leqq y} \alpha(Dg^y(z))$, so $g^y(y) = g^y(0)$. Hence, $T(0)g(y) = T(y)g(0)$, and since y is arbitrary, $g(t) = T(t)g(0) = T(t)x$ for all $t \in \mathbf{R}_+$. Hence, the solution is unique. □

A consequence of the above theorem is that the infinitesimal generator uniquely determines the d.s.g.

THEOREM 1.3. *Let E and F be l.c.s.'s with S a d.s.g. on E and T a d.s.g. on F. For a continuous linear map $f : E \to F$ to be a dynamorphism in* **LCS** *from (E, g_S) to (F, g_T), it is necessary and sufficient that the diagram*



commute for each $t \in \mathbf{R}_+$. In particular, if $g_S = g_T$, then $S = T$.

Proof. If the above diagram commutes for each $t \in \mathbf{R}_+$, then f is clearly a dynamorphism in **LCS**. Conversely, let $x \in E$, and suppose $f \circ g_S = g_T \circ f$. The functions $t \mapsto f(S(t)x)$ and $t \mapsto T(t)(f(x))$ are each solutions to the differential equation $Du(t) = g_T(u(t))$ with $u(0) = f(x)$, as is readily verified. Hence, by Theorem 1.2, they are equal, so that $f \circ S(t) = T(t) \circ f$ for all $t \in \mathbf{R}_+$. □

The d.s.g. determined by the infinitesimal generator g will be denoted $\mathscr{T}_g$.

A characterization theorem. A very important characterization of $\mathscr{E}(\mathbf{R}_+, E)$ and hence d.s.g.'s will now be developed. Denote by $\Delta(\mathbf{R}_+)$ the subspace of $\mathscr{E}'(\mathbf{R}_+)$ consisting of the distributions which have finite support. The elements of $\Delta(\mathbf{R}_+)$ are just finite linear combinations of elements of the form $D^p\delta_t$, where $p \in \mathbf{N}$ and δ_t is the Dirac measure at $t \in \mathbf{R}_+$ (see Schwartz (1966, Ch. III, Thm. XXXV)). The next result gives some important properties of $\Delta(\mathbf{R}_+)$.

LEMMA 1.4. (a) $\Delta(\mathbf{R}_+)$ *is dense in* $\mathscr{E}'(\mathbf{R}_+)$.

(b) *Every bounded subset of* $\mathscr{E}'(\mathbf{R}_+)$ *is contained in the closure of a bounded subset of* $\Delta(\mathbf{R}_+)$.

(c) *The strong dual of* $\Delta(\mathbf{R}_+)$ *is* $\mathscr{E}(\mathbf{R}_+)$.

(d) $\Delta(\mathbf{R}_+)$ *is a* (DF) *space.*

(e) $\mathscr{E}'(\mathbf{R}_+)$ *is a quasi-completion and a completion of* $\Delta(\mathbf{R}_+)$.

Proof. (a) is obvious.

(b) Let $U_0, U_1, U_2, \cdots$ be the fundamental sequence of barreled neighborhoods of 0 in $\mathscr{E}(\mathbf{R}_+)$ given by

$$U_k = \left\{ \varphi \in \mathscr{E}(\mathbf{R}_+) \middle| \sup_{\substack{t \in [0,k] \\ q \leqq k}} |D^q \varphi(t)| \leqq \frac{1}{k+1} \right\}.$$

$\mathscr{E}'(\mathbf{R}_+)$ is the strong dual of $\mathscr{E}(\mathbf{R}_+)$ so $(U_0)^\circ, (U_1)^\circ, (U_2)^\circ, \cdots$ is a fundamental sequence of bounded sets in $\mathscr{E}'(\mathbf{R}_+)$. Denote by B_k the intersection of $(U_k)^\circ$ and $\Delta(\mathbf{R}_+)$. B_k is surely bounded in $\Delta(\mathbf{R}_+)$, and it suffices to show that $(B_k)^\circ \subset U_k$, for then $(B_k)^\circ = U_k$ (all polars taken in $\langle \mathscr{E}'(\mathbf{R}_+), \mathscr{E}(\mathbf{R}_+) \rangle$). However, it is clear that $(k+1)D^j\delta_t \in B_k$ for each $j \leqq k$, $t \in [0, k]$, so that $\varphi \in (B_k)^\circ \Rightarrow |D^j\varphi(t)| \leqq 1/(k+1)$ for each $j \leqq k$, $t \in [0, k]$, i.e., $\varphi \in U_k$.

(c) This follows immediately from (b).

(d) This follows immediately from the definition of (DF) space, since the strong dual of $\Delta(\mathbf{R}_+)$ is an (F) space, by (c).

(e) $\mathscr{E}'(\mathbf{R}_+)$ is complete. Hence, it is a completion of $\Delta(\mathbf{R}_+)$ since $\Delta(\mathbf{R}_+)$ is dense in $\mathscr{E}'(\mathbf{R}_+)$ by (a). $\mathscr{E}'(\mathbf{R}_+)$ is also a quasi-completion of $\Delta(\mathbf{R}_+)$ since $\Delta(\mathbf{R}_+)$ is a (DF) space, by (d), and a quasi-complete (DF) space is complete (Grothendieck (1973, Ch. 4, Pt. 3, Prop. 4, Cor. 2)). □

Let E be a l.c.s. Define the map $\Phi_E : \Delta(\mathbf{R}_+) \times \mathscr{E}(\mathbf{R}_+, E) \to E$ by $(D^p\delta_t, \varphi) \mapsto D^p\varphi(t)$. This map is clearly bilinear. Much more is true, in fact, but first a notation is given. If $\varphi \in \mathscr{E}(\mathbf{R}_+, E)$ and $e' \in E'$, $\langle \varphi, e' \rangle$ denotes the function $t \mapsto \langle \varphi(t), e' \rangle$, a slight abuse of notation. $t \mapsto \langle \varphi(t), e' \rangle$ is clearly an element of $\mathscr{E}(\mathbf{R}_+)$.

LEMMA 1.5. *Let* E *be a l.c.s.*

(a) *For* $x \in \Delta(\mathbf{R}_+)$, $\varphi \in \mathscr{E}(\mathbf{R}_+, E)$, *and* $e' \in E'$, $\langle \Phi_E(x, \varphi), e' \rangle = x(\langle \varphi, e' \rangle)$.

(b) Φ_E *is hypocontinuous.*

Proof. (a) is immediate.

(b) Let $V \in \mathscr{U}(E)$ be a barrel. Let A be an absolutely convex closed bounded subset of $\mathscr{E}(\mathbf{R}_+, E)$, and let B be an absolutely convex closed bounded subset of $\Delta(\mathbf{R}_+)$. It suffices to find $U \in \mathscr{U}(\Delta(\mathbf{R}_+))$ and $W \in \mathscr{U}(\mathscr{E}(\mathbf{R}_+, E))$ such that $\Phi_E(U \times A) \subset V$ and $\Phi_E(B \times W) \subset V$. It suffices also to assume that V is the closed semiball of a continuous seminorm α on E.

A is bounded, hence for any compact $K \subset \mathbf{R}_+$, $p \in \mathbf{N}$,

$$\sup_{\substack{\varphi \in A \\ q \leqq p}} \sup_{x \in K} \alpha(D^q\varphi(x)) = M < \infty.$$

Hence,

$$\sup_{\substack{\varphi\in A\\ e'\in V^\circ}}\sup_{\substack{x\in K\\ q\leqq p}}|\langle\varphi^q(x),e'\rangle|\leqq M;$$

i.e., $\{\langle\varphi,e'\rangle\,|\,\varphi\in A, e'\in V^\circ\}$ is bounded in $\mathscr{E}(\mathbf{R}_+)$. Put $U=\{\langle\varphi,e'\rangle\,|\,\varphi\in A, e'\in V^\circ\}^\circ$. Then,

$$\sup_{\substack{x\in U\\ \varphi\in A\\ e'\in V^\circ}}|\langle\Phi_E(x,\varphi),e'\rangle|=\sup_{\substack{x\in U\\ \varphi\in A\\ e'\in V^\circ}}|x(\langle\varphi,e'\rangle)|=\sup_{\substack{x\in U\\ y\in U^\circ}}|x(y)|=1.$$

Hence, $\Phi_E(U\times A)\subset V$.

Next, B is the polar of a neighborhood of 0 in $\mathscr{E}(\mathbf{R}_+)$ (polar for the pair $\langle\Delta(\mathbf{R}_+),\mathscr{E}(\mathbf{R}_+)\rangle$). Say,

$$B^\circ=\left\{\varphi\in\mathscr{E}(\mathbf{R}_+)\,\middle|\,\sup_{\substack{x\in K\\ q\leqq p}}|D^q\varphi(x)|\leqq\xi\right\},$$

for some $K\subset\mathbf{R}_+$ compact, $p\in\mathbf{N}$, $\xi>0$, without loss of generality. Put

$$W=\left\{\varphi\in\mathscr{E}(\mathbf{R}_+,E)\,\middle|\,\sup_{\substack{x\in K\\ q\leqq p}}\alpha(D^q\varphi(x))\leqq\xi\right\}.$$

Now,

$$\sup_{\substack{x\in B\\ \varphi\in W\\ e'\in V^\circ}}|\langle\Phi_E(x,\varphi),e'\rangle|=\sup_{\substack{x\in B\\ \varphi\in W\\ e'\in V^\circ}}|x(\langle\varphi,e'\rangle)|=1,$$

since $\varphi\in W$, $e'\in V^\circ\Rightarrow\langle\varphi,e'\rangle\in B^\circ$. Hence,

$$\Phi_E(B\times W)\subset V.$$ □

$\Phi_E(x,-)$ is called the *extension of x to vector-valued functions.* If E is quasi-complete (for example), then Φ_E has a unique (hypocontinuous) extension $\hat{\Phi}_E:\mathscr{E}'(\mathbf{R}_+)\times\mathscr{E}(\mathbf{R}_+,E)\to E$, in view of Lemma 1.4(a) and (b). It is easy to see that this extension is exactly the extension of distributions to vector-valued functions, as developed (in an entirely different manner) by Schwartz (1954–55). It is Φ_E and not this extension which is of primary value here, however.

With these preliminary results, the characterization theorem for $\mathscr{E}(\mathbf{R}_+,E)$ may be stated and proved.

THEOREM 1.6. *Let E be a l.c.s. There is an isomorphism $i:\mathscr{E}(\mathbf{R}_+,E)\to L_b(\Delta(\mathbf{R}_+),E)$ given by $\varphi\mapsto(x\mapsto\Phi_E(x,\varphi))$. The inverse of this map is $f\mapsto(t\mapsto f(\delta_t))$.*

Proof. It is easy to see that $f\mapsto(t\mapsto f(\delta_t))$ is both a left and a right inverse to i, so that i is a bijection. The continuity of i is an immediate consequence of the hypocontinuity of Φ_E. To show that i^{-1} is continuous, let $U\in\mathscr{U}(\mathscr{E}(\mathbf{R}_+,E))$. Without loss of generality, it suffices to assume that

$$U=\left\{\varphi\in\mathscr{E}(\mathbf{R}_+,E)\,\middle|\,\sup_{\substack{t\in K\\ q\leqq p}}\alpha(D^q\varphi(t))\leqq 1\right\},$$

for some compact $K\subset\mathbf{R}_+$, $p\in\mathbf{N}$, and α a continuous seminorm on E. Now set V equal to the unit semiball of α, and let

$$A=\{D^q\delta_t\in\Delta(\mathbf{R}_+)\,|\,q\leqq p\text{ and }t\in K\}.$$

Clearly,

$$i^{-1}(\{f \mid f(A) \subset V\}) \subset U,$$

so it suffices to show that A is bounded. However,

$$A^\circ = \left\{ \psi \in \mathscr{E}(\mathbf{R}_+) \,\middle|\, \sup_{\substack{x \in K \\ q \leq p}} |D^q \psi(x)| \leqq 1 \right\},$$

which is in $\mathscr{U}(\mathscr{E}(\mathbf{R}_+))$. Hence A is bounded, and so i^{-1} is continuous. □

Examples of d.s.g.'s. Let E be any l.c.s. On $\mathscr{E}(\mathbf{R}_+, E)$ define the left shift by t ($t \in \mathbf{R}_+$) to be $\mathscr{S}_t : \varphi \mapsto \varphi_t$, where $\varphi_t(s) = \varphi(t+s)$. Clearly, $\mathscr{S}_t \in L(\mathscr{E}(\mathbf{R}_+, E))$. Define $\mathfrak{S}_E : \mathbf{R}_+ \to L(\mathscr{E}(\mathbf{R}_+, E))$ by $\mathfrak{S}_E(t) = \mathscr{S}_t$. Define on $\mathscr{E}'(\mathbf{R}_+)$ a similar right shift by $\mathscr{S}'_t : f \mapsto f_t$, where $f_t(\varphi) = f(\varphi_t)$ for $\varphi \in \mathscr{E}(\mathbf{R}_+)$. $[(f_t - f)/t](\varphi) = f((\varphi_t - \varphi)/t)$, so $\mathscr{S}'_t \in L(\mathscr{E}'(\mathbf{R}_+))$. Define $\mathfrak{S}' : \mathbf{R}_+ \to L(\mathscr{E}'(\mathbf{R}_+))$ by $\mathfrak{S}'(t) = \mathscr{S}'_t$. Finally, note that $\mathscr{S}'_t(\Delta(\mathbf{R}_+)) \subset \Delta(\mathbf{R}_+)$. Define $\mathfrak{S}^\Delta(t) : \mathbf{R}_+ \to L(\Delta(\mathbf{R}_+))$ by $\mathfrak{S}^\Delta(t) = \mathscr{S}'_t|_{\Delta(\mathbf{R}_+)}$.

THEOREM 1.7. (a) *$\mathfrak{S}'$ is a d.s.g. on $\mathscr{E}'(\mathbf{R}_+)$ with infinitesimal generator D.*

(b) *$\mathfrak{S}^\Delta$ is a d.s.g. on $\Delta(\mathbf{R}_+)$ with infinitesimal generator D.*

(c) *If E is a l.c.s., then $\mathfrak{S}_E$ is a d.s.g. on $\mathscr{E}(\mathbf{R}_+, E)$ with infinitesimal generator D.*

Proof. (a) Clearly $\mathfrak{S}'$ satisfies (s_1), (s_2), and (s_3), just by definition. To show (s_4), first note that $\{\mathfrak{S}'(t) \mid 0 \leqq t \leqq \varepsilon\}$ is bounded for any $\varepsilon > 0$, since $t \mapsto \mathfrak{S}'(t)$ is continuous, $[0, \varepsilon]$ is compact in $\mathbf{R}_+$, the continuous image of a compact set is compact, and every compact subset of l.c.s. is bounded. However, $\mathscr{E}'(\mathbf{R}_+)$ is barreled, and so every simply-bounded subset of $L(\mathscr{E}'(\mathbf{R}_+))$ is equicontinuous (Schaefer (1971, Ch. 3, 4.2)). Hence, $\{T(t) \mid 0 \leqq t \leqq \varepsilon\}$ is equicontinuous for any $\varepsilon > 0$.

(b) This follows immediately from (a), since restrictions of equicontinuous sets are equicontinuous (Bourbaki (1966, Ch. X, § 2, Prop. 4)), and $Dx \in \Delta(\mathbf{R}_+)$ for $x \in \Delta(\mathbf{R}_+)$.

(c) Recall from Theorem 1.6 that $\mathscr{E}(\mathbf{R}_+, E) \cong L(\Delta(\mathbf{R}_+), E)$. Furthermore, under the canonical identification $i : \varphi \mapsto (x \mapsto \Phi_E(x, \varphi))$, it is clear that $\mathfrak{S}_E(t)(\varphi) = \varphi_t = i(\varphi) \circ \mathfrak{S}^\Delta(t)$. The result now follows from (b). □

Relationship to (C_0) semigroups. The theory of equicontinuous semigroups of class (C_0) is an alternate approach to semigroups of operators on locally convex spaces. It will now be shown how these semigroups compare to d.s.g.'s. Let E be a sequentially-complete l.c.s., and let $T : \mathbf{R}_+ \to L(E)$ be a map which satisfies axioms (s_1) and (s_2). T is called an *equicontinuous semigroup of class* (C_0) (abbreviated e.s.g.) if the following two axioms are also satisfied (Yosida (1971, Ch. IX, 2)).

(e_1) $\lim_{t \to t_0} T(t)x = T(t_0)x$ for any $t_0 \in \mathbf{R}_+$, $x \in E$.

(e_2) $\{T(t) \mid t \in \mathbf{R}_+\}$ is equicontinuous.

Let $\mathscr{A} = \{x \in E \mid \lim_{t \to 0} [(T(t) - 1)/t](x) \text{ exists}\}$, and define $A : \mathscr{A} \to E$ by $x \mapsto \lim_{t \to 0} [(T(t) - 1)/t](x)$. A is called the infinitesimal generator of T, and $\mathscr{A}$ is dense in E (Yosida (1971, Ch. IX, 2, Th. 1)).

PROPOSITION 1.8. *Let E be a l.c.s., and let $T : \mathbf{R}_+ \to L(E)$.*

(a) *If T is a d.s.g., then T is an e.s.g. if and only if E is sequentially complete and $\{T(t) \mid t \in \mathbf{R}_+\}$ is equicontinuous.*

(b) *If E is sequentially complete and T is an e.s.g., then T is a d.s.g. if and only if the domain of the infinitesimal generator is all of E.*

Proof. The proof is an immediate consequence of the definitions. □

$\{\mathfrak{S}(t) \mid t \in \mathbf{R}_+\}$ is easily seen to be *not* equicontinuous, so not every d.s.g. is an e.s.g. Conversely, let $C(\mathbf{R}_+)$ denote the space of bounded uniformly continuous $\mathbf{K}$-valued functions on $\mathbf{R}_+$ with the sup-norm topology, and define $T(t) : \mathbf{R}_+ \to L(C(\mathbf{R}_+))$ by

$(T(t)f)(x) = f(x+t)$. T is an e.s.g. (Yosida (1971, Ch. IX, 2, Example 2)), but not a d.s.g. since not every uniformly continuous function is differentiable. Hence, not every e.s.g. is a d.s.g. Thus, neither of these semigroup concepts is subsumed by the other.

If E is a (B) space, a map $T: \mathbf{R}_+ \to L(E)$ is called a (C_0) semigroup if (s_1), (s_2), and (e_1) are satisfied. Note that (s_4) is satisfied automatically because a (B) space is barreled and $\{T(t) \mid 0 \leqq t \leqq \varepsilon\}$ is surely bounded, so it is not necessary to require (s_4) in the case of a (C_0) semigroup on a (B) space. The infinitesimal generator of a (C_0) semigroup is defined as for an e.s.g.; its domain is dense in E also. (C_0) semigroups are related to d.s.g.'s by the following result.

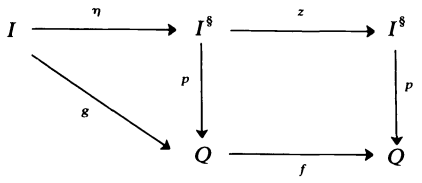
PROPOSITION 1.9. *Let E be a* (B) *space, and let $A \in L(E)$. A is the infinitesimal generator of a unique (C_0) semigroup (which is also a d.s.g.) given by $t \mapsto e^{At}$.*

Proof. Consult Rudin (1973, 13.36). □

For a more complete discussion of e.s.g.'s, consult Yosida (1971).

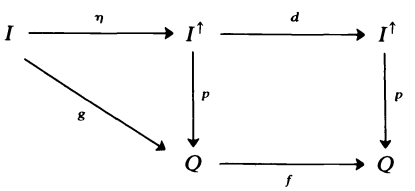
2. Input–output behavior. One of the key features of the approach to discrete-time systems given in Arbib and Manes (1974) is that given a decomposable system M, a *canonical* input–output behavior may be constructed which M realizes. In this section, it is shown that an analogous such canonical behavior exists for continuous-time systems.

Input behavior. Let $\mathcal{K}$ be a category, and I an object of $\mathcal{K}$. A *free system dynamics* over I is a system dynamics $(I^{\S}, z)$ in $\mathcal{K}$ and a $\mathcal{K}$ morphism $\eta: I \to I^{\S}$ such that for any other system dynamics (Q, f) and morphism $g: I \to Q$, there is a unique dynamorphism $\rho: (I^{\S}, z) \to (Q, f)$ such that



commutes. Free system dynamics are unique up to isomorphism, and exist in many categories, including **LS**. In **LS**, a free system dynamics is given by $I^{\S} = \{(\cdots, i_n, \cdots, i_0) \mid i_k \in I$ and only finitely many terms nonzero$\}$, $z: I^{\S} \to I^{\S}$ is left-shift one space, and $\eta: I \to I^{\S}$ is injection into the rightmost position. If $M = (Q, f, I, g, Y, h)$ is a decomposable system in $\mathcal{K}$, the map ρ is called the *reachability map* of M. When $\mathcal{K} = \mathbf{LS}$, ρ is given by $(\cdots, i_n, \cdots, i_0) \mapsto \sum_{k \geqq 0} f^k \circ g(i_k)$. See Arbib and Manes (1974) for details.

In the continuous-time case, the free system dynamics must be augmented with a smoothness condition. Let $\mathcal{K}$ be a subcategory of **LCS**. For a $\mathcal{K}$ object I, a *free differentiable system dynamics* over I is a differentiable system dynamics $(I^{\uparrow}, d)$ in $\mathcal{K}$ and a continuous linear map $\eta: I \to I^{\uparrow}$ such that for any other differentiable system dynamics (Q, f) in $\mathcal{K}$ and $\mathcal{K}$ morphism $g: I \to Q$, there is a unique dynamorphism $\rho: (I^{\uparrow}, d) \to (Q, f)$ such that



commutes. Free differentiable system dynamics are easily seen to be unique up to isomorphism, if they exist. If $M=(Q, f, I, g, Y, h)$ is a differentiable decomposable system in $\mathscr{K}$, ρ is called the *reachability map* of M *for continuous time*. It will be called just the reachability map of M for the rest of this paper; the discrete-time reachability map will be explicitly qualified as such. In this section, it will be shown that free differentiable system dynamics exist in **LCS**, as well as **QC** and **CS**.

The case of $\mathscr{K}=\mathbf{LCS}$ will now be considered. Recall from § 1 that $\Delta(\mathbf{R}_+)$ is the space of all distributions on $\mathbf{R}_+$ which have finite support, i.e., which have the form $\sum_{k=1}^{n} a_k D^{p_k}\delta_{t_k}$. Now in the general case, the inputs in $I^{\uparrow}$ are distributions of finite support, but they are I-valued rather than scalar-valued. That is, they are of the form $\sum_{k=1}^{n} i_k \cdot D^{p_k}\delta_{t_k}$, with $i_k \in I$, $p_k \in \mathbf{N}$, $t_k \in \mathbf{R}_+$, for $1 \leqq k \leqq n$. The proper mathematical way of viewing such I-valued distributions is via the tensor product $\Delta(\mathbf{R}_+)\otimes I$, with the identification $\sum_{k=1}^{n} i_k \cdot D^{p_k}\delta_{t_k} = \sum_{k=1}^{n} D^{p_k}\delta_{t_k}\otimes i_k$.

Recall that $\mathfrak{S}^{\Delta}$ is the d.s.g. on $\Delta(\mathbf{R}_+)$ given by the right-shift operator, and that the infinitesimal generator of this d.s.g. is just the distributional differentiation operator D restricted to $\Delta(\mathbf{R}_+)$. Let I be a l.c.s. A semigroup of operators $\mathfrak{S}^{\Delta}\otimes I$ may be induced on $\Delta(\mathbf{R}_+)\otimes I$ by defining $(\mathfrak{S}^{\Delta}\otimes I)(t)=\mathfrak{S}^{\Delta}(t)\otimes 1_I$. To make this a d.s.g., a topology must be placed on $\Delta(\mathbf{R}_+)\otimes I$. The appropriate topology, as will now be shown, is the hypocontinuous topology $\Delta(\mathbf{R}_+)\otimes_{\beta} I$.

LEMMA 2.1. *Let E and F be l.c.s.'s, $H\subset L(E)$ an equicontinuous family. Then, $H\otimes 1=\{f\otimes 1 \mid f\in H\}$ is an equicontinuous subset of $L(E\otimes_{\beta}F)$.*

Proof. Let $p: E\times F\to E\otimes_{\beta}F$ be the canonical map, which is hypocontinuous by definition. It suffices to show that $p\circ(H\times 1)=\{p\circ(h\times 1)\mid h\in H\}$ is equihypocontinuous. Let $V\in\mathscr{U}(E\otimes_{\beta}F)$ and let $A\subset E$ be bounded. Since H is equicontinuous, it is bounded, so $H(A)$ is also bounded (Grothendieck (1973, Ch. 3, Prop. 3, Cor. 1)). Since p is hypocontinuous, there is a $U\in\mathscr{U}(F)$ such that $p(H(A)\times U)\subset V$. However, $(H\times 1)(A\times U)=(H(A)\times U)$. Hence $p\circ(H\times 1)(A\times U)\subset V$. Next, let $B\subset F$ be bounded. Since p is hypocontinuous, there is a $W\in\mathscr{U}(E)$ such that $p(W\times B)\subset V$. Since H is equicontinuous, there is a $Y\in\mathscr{U}(E)$ such that $H(Y)\subset W$. Hence $p\circ(H\times 1)(Y\times B)\subset p(W\times B)\subset V$. Thus, $p\circ(H\times 1)$ is equihypocontinuous, so $H\otimes 1$ is equicontinuous in $L(E\otimes_{\beta}F)$. □

LEMMA 2.2. *Let I be a l.c.s. $\mathfrak{S}^{\Delta}\otimes_{\beta} I$ is a d.s.g. on $\Delta(\mathbf{R}_+)\otimes_{\beta} I$, with infinitesimal generator $D\otimes 1_I$.*

Proof. (s_1) and (s_2) are trivial. (s_4) follows from Lemma 2.1. Hence, all that need be shown is (s_3). To do this, it suffices to show that for every pair $(x, i)\in\Delta(\mathbf{R}_+)\times I$ and $V\in\mathscr{U}(\Delta(\mathbf{R}_+)\otimes_{\beta} I)$, there is a $t_0\in\mathbf{R}_+$ such that $0<t\leqq t_0$ implies that $p(((\mathfrak{S}^{\Delta}(t)x-x)/t-Dx), i)\in V$, where $p:\Delta(\mathbf{R}_+)\times I\to\Delta(\mathbf{R}_+)\otimes_{\beta} I$ is the canonical map. Since p is hypocontinuous and $\{i\}$ is bounded, $U=\{y\in\Delta(\mathbf{R}_+)\mid (y, i)\in p^{-1}(V)\}\in\mathscr{U}(\Delta(\mathbf{R}_+))$. However, $\lim_{t\to 0}(\mathfrak{S}^{\Delta}(t)x-x)/t-Dx=0$. Hence, there is a t_0 such that $0<t_0\leqq t$ implies that $(\mathfrak{S}^{\Delta}(t)x-x)/t-Dx\in U$. Thus, $p(((\mathfrak{S}^{\Delta}(t)x-x)/t-Dx), i)\in V$ for $0<t\leqq t_0$, so $\mathfrak{S}^{\Delta}\otimes_{\beta} 1_I$ is differentiable with derivative $D\otimes 1_I$. □

To avoid confusion, $\mathfrak{S}^{\Delta}\otimes I$ will be denoted $\mathfrak{S}^{\Delta}\otimes_{\beta} I$ when it operates on $\Delta(\mathbf{R}_+)\otimes_{\beta} I$.

To show that $\mathfrak{S}^{\Delta}\otimes_{\beta} I$ is a free differentiable system dynamics over I, it is necessary to find the reachability map ρ for each differentiable system dynamics (Q, f) in **LCS** and continuous linear map $g: I\to Q$. ρ is constructed as follows. Recall from § 1 that $\Phi_Q:\Delta(\mathbf{R}_+)\times\mathscr{E}(\mathbf{R}_+, Q)\to Q$ is the extension to vector-valued functions given by $\Phi_Q(\sum_{k=1}^{n} a_k D^{p_k}\delta_{t_k}, \varphi)=\sum_{k=1}^{n} a_k(D^{p_k}\varphi)(t_k)$, and that this map is bilinear and hypocontinuous (see Lemma 1.5). Now recall from Proposition 1.1 that $\Lambda_{\mathscr{T}_f, Q}: Q\to\mathscr{E}(\mathbf{R}_+, Q)$ is given by $q\mapsto(t\to\mathscr{T}_f(t)(q))$. That is, $\Lambda_{\mathscr{T}_f, Q}$ takes q to the "natural response" of the

dynamics (Q, f) with initial state q. Define the intermediate map $\tilde{g}_f : \Delta(\mathbf{R}_+) \times I \to Q$ by

$$\Delta(\mathbf{R}_+) \times I \xrightarrow{1 \times g} \Delta(\mathbf{R}_+) \times Q \xrightarrow{1 \times \Lambda_{g_f, Q}} \Delta(\mathbf{R}_+) \times \mathscr{E}(\mathbf{R}_+, Q) \xrightarrow{\Phi_Q} Q.$$

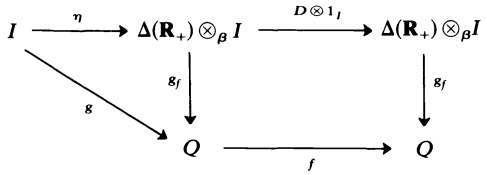
That is, to compute $\tilde{g}_f$ on the pair $(\sum_{k=1}^{n} a_k D^{p_k} \delta_{t_k}, i)$, first compute the state $g(i)$ associated with q, then compute its natural response $t \mapsto \mathscr{T}_f(t)(g(i))$, and finally apply $\sum_{k=1}^{n} a_k D^{p_k} \delta_{t_k}$ to this natural response, to get $\sum_{k=1}^{n} a_k D^{p_k} \circ \mathscr{T}_f(t_k)(g(i))$. The linear map $g_f : \Delta(\mathbf{R}_+) \otimes_\beta I \to Q$ associated with bilinear map $\tilde{g}_f$ is the reachability map ρ. To show this it is first of all necessary to establish that g_f is continuous, which is equivalent to $\tilde{g}_f$ hypocontinuous.

LEMMA 2.3. *Let I and Q be l.c.s.'s, let $g : I \to Q$ be a continuous linear map, and let (Q, f) be a differentiable system dynamics in* **LCS**. *The map $\tilde{g}_f : \Delta(\mathbf{R}_+) \times I \to Q$ is bilinear and hypocontinuous.*

Proof. Clearly $\tilde{g}_f$ is bilinear. Since 1 and $\Lambda_{\mathscr{T}_f, Q}$ are continuous (see Proposition 1.1(d)), they each map bounded sets into bounded sets, whence the hypocontinuity of $\tilde{g}_f$ is immediate from the hypocontinuity of Φ_Q (see Lemma 1.5(b)). □

Define $\eta : I \to \Delta(\mathbf{R}_+) \otimes_\beta I$ to be the canonical injection $i \mapsto \delta_0 \otimes i$. η is clearly linear. It is continuous because it is the restriction of the canonical projection $p : \Delta(\mathbf{R}_+) \times I \to \Delta(\mathbf{R}_+) \otimes_\beta I$ to $\{\delta_0\} \times I$. With these data, the main result may be stated.

THEOREM 2.4. *Let I be a l.c.s., let (Q, f) be a differentiable system dynamics in* **LCS**, *and let $g : I \to Q$ be a continuous linear map. The pair $(\eta, (\Delta(\mathbf{R}_+) \otimes_\beta I, D \otimes 1_I))$ is a free differentiable system dynamics over I. g_f is the unique continuous linear map which makes the diagram below commute.*



Proof. The triangle commutes by definition of η and g_f. The square commutes because $g_f \circ (D \otimes 1_I)(D^p \delta_t \otimes i) = g_f(D^{p+1} \delta_t \otimes i) = D^{p+1} \mathscr{T}_f(t)(g(i)) = f^{p+1} \mathscr{T}_f(t)(g(i)) = f^{p+1} \circ g_f(D^{p+1} \delta_t \otimes i)$, in view of Theorem 1.3. To show g_f is unique, suppose that the above diagram also commutes with k replacing g_f. $(k - g_f)(\delta_0 \otimes i) = 0$, by definition of g_f. $(k - g_f)(D^p \delta_0 \otimes i) = 0$, by induction on p. To show $(k - g_f)(D^p \delta_t \otimes i) = 0$, note that $(k - g_f)(D^p \delta_t \otimes i) = (\mathfrak{S}^\Delta(t) \otimes 1_I)(k - g_f)(D^p \delta_0 \otimes i) = 0$. □

Thus, given a differentiable decomposable system $M = (Q, f, I, g, Y, h)$ in **LCS**, the reachability map of M is g_f. A natural interpretation of g_f will now be given. View the dynamics of the system as being described by

$$\frac{dg(t)}{dt} = f(q(t)) + g(i(t)).$$

An element of $\Delta(\mathbf{R}_+) \otimes_\beta I$ may be viewed (as shown previously) as a finite sum of the form $\sum_{k=1}^{n} i_k \cdot D^{p_k} \delta_{t_k}$. Interpret a typical element of this sum, $i_k \cdot D^{p_k} \delta_{t_k}$, as an instantaneous input occurring at $t = -t_k$ (note the minus sign) with value $i_k \cdot D^{p_k} \delta_0$. Assuming that the system was previously at rest, the state at $-t_k$ will instantaneously jump to $f^{p_k} g(i_k)$. At $t = 0$, it will have decayed to $\mathscr{T}_f(t_k) f^{p_k} g(i_k) = D^{p_k} \mathscr{T}_f(t_k) g(i_k)$, in the absence of any further inputs. For a finite sum of such inputs, merely apply the superposition principle to get the resulting state at time 0.

In the classical case of finite-dimensional linear systems, the state at time t (with zero initial state) is given by $q(t) = \int_{-\infty}^{t} e^{f(t-s)}\, g(i(s))\, ds$ (see Padulo and Arbib (1974, Ch. 6)). In the present case, $e^{f(t-s)}$ is replaced by the more general response $\mathcal{T}_f(t-s)$. $g(i(s))$ is interpreted by $g(\sum_{k=1}^{n} i_k \cdot D^{p_k}\delta_{t_k}) = \sum_{k=1}^{n} g(i_k) D^{p_k}\delta_{t_k}$. If $\int_{-\infty}^{t}$ is interpreted as evaluation in the sense of distributions, then indeed $q(t) = \int_{-\infty}^{t} \mathcal{T}_f(t-s) g(i(s))\, ds$, as is easily seen. Therefore, the above construction does reduce to the classical case for finite-dimensional systems.

The significance of the above result must not be underestimated. Just as Arbib (1973) showed that the correct input construction for group machines is the coproduct construction and not the free monoid construction, so too does the above show that in the case of differentiable decomposable systems in **LCS**, the correct space of input signals is $\Delta(\mathbf{R}_+) \otimes_\beta I$, and not anything else. Nonetheless, the reader may feel a bit disappointed at this point. After all, inputs to continuous-time systems do not consist of just trains of impulses. Rather, they include at least smooth functions with compact support, and hopefully a lot more. Can these somehow be included? The answer is yes, provided that the construction is carried out in the right category. The problem is that there are too many differentiable system dynamics in **LCS**; categories with more completeness properties must be used. The two subcategories of **LCS** which will be considered here are **QC**, the category of all quasi-complete l.c.s.'s, and **CS**, the category of all complete l.c.s.'s. Since the analyses of these two cases are entirely similar, they will be developed completely in parallel.

The obvious approach to extending Theorem 2.3 to the categories **QC** and **CS** is just to apply $\hat{}$ or $\hat{}$ to every space and map on the diagram. However, it is not quite that simple, because the quasi-completion or completion of a d.s.g. is not necessarily a d.s.g. Fortunately, this approach does work in the cases considered here.

Before proceeding further, it is useful to recall that $\mathcal{E}'(\mathbf{R}_+)$ may be identified as both a quasi-completion and a completion of $\Delta(\mathbf{R}_+)$ (see Lemma 1.4(e)). However, one may not immediately assert that $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I \cong \Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$ and $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I \cong \Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$, since the extension of a hypocontinuous bilinear map need not be hypocontinuous. Nonetheless, in this particular case, the necessary result can be shown.

LEMMA 2.5. *Let I be a quasi-complete (resp. complete) l.c.s.*

(a) *Every hypocontinuous bilinear map on $\Delta(\mathbf{R}_+) \times I$ into a quasi-complete (resp. complete) l.c.s. extends uniquely to a hypocontinuous bilinear map on $\mathcal{E}'(\mathbf{R}_+) \times I$.*

(b) *$\mathcal{E}'(\mathbf{R}_+) \otimes_\beta I$ is a dense subspace of $\Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$ (resp. $\Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$).*

Proof. (a) By Lemma 1.4(b), every bounded subset of $\mathcal{E}'(\mathbf{R}_+)$ is contained in the closure of a bounded subset of $\Delta(\mathbf{R}_+)$. The result then follows from a standard extension theorem of hypocontinuous bilinear mappings (see Grothendieck (1973, Ch. 3, Prop. 10)).

(b) This follows immediately from (a). □

From now on, if I is a quasi-complete (resp. complete) l.c.s., $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ and $\Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$ (resp. $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ and $\Delta(\mathbf{R}_+)\hat{\otimes}_\beta I$) shall be identified with each other.

Now it is possible to show that the appropriate extension of $\mathfrak{S}^\Delta \otimes_\beta I$ is the free differentiable system dynamics in the **QC** and **CS** cases. The only step remaining is to show that this extension is a d.s.g., and this is demonstrated next.

LEMMA 2.6. *For any l.c.s. I and any $\varepsilon > 0$,*

$$\left\{ \frac{(\mathfrak{S}' \otimes_\beta I)(t) - 1}{t} \,\middle|\, 0 < t \leqq \varepsilon \right\} \quad \text{and} \quad \left\{ \frac{(\mathfrak{S}^\Delta \otimes_\beta I)(t) - 1}{t} \,\middle|\, 0 < t \leqq \varepsilon \right\}$$

are equicontinuous.

Proof. $\{(\mathcal{E}'(t)-1)/t \mid 0<t\leqq\varepsilon\}\cup\{D\}$ is bounded for any $\varepsilon>0$, since it is compact, being the image of $[0, t]$ under the continuous map $t\mapsto(\mathcal{E}'(t)-1)/t$ for $t>0$ and $0\mapsto D$. Now $\mathcal{E}'(\mathbf{R}_+)$ is barreled, and so every weakly bounded subset of $L(\mathcal{E}'(\mathbf{R}_+))$ is equicontinuous, hence $\{(\mathcal{E}'(t)-1)/t \mid 0<t\leqq\varepsilon\}$ is equicontinuous (see Schaefer (1971, Ch. III, 4.2)). Since restrictions of equicontinuous sets are equicontinuous (Bourbaki (1966, Ch. X, § 2, Prop. 4)), $\{(\mathcal{E}^{\Delta}(t)-1)/t \mid 0<t\leqq\varepsilon\}$ is also equicontinuous. An appeal to Lemma 2.1 now completes the proof. □

LEMMA 2.7. *Let E be a l.c.s., and let T be a d.s.g. on E. If $\{(T(t)-1)/t \mid 0<t\leqq\varepsilon\}$ is equicontinuous for some $\varepsilon>0$, then T has a unique extension $\hat{T}$ (resp. $\hat{T}$) to a d.s.g. on $\hat{E}$ (resp. $\hat{E}$). This extension is given by $\hat{T}(t)(x)=(T(t))\hat{}(x)$ (resp. $\hat{T}(t)(x)=(T(t))\hat{}(x)$), and its infinitesimal generator is $(\mathcal{G}_T)\hat{}$ (resp. $(\mathcal{G}_T\hat{})$).*

Proof. Only the completion case will be given; the quasi-completion case is entirely analogous. Verification of properties (s_1), (s_2) and (s_4) for T is trivial. To verify (s_3), proceed as follows. Extensions of equicontinuous families are equicontinuous (Bourbaki (1966, Ch. X, § 2, Prop. 4)), so $\{(\hat{T}(h)-1)/h \mid 0<h\leqq\varepsilon\}$ is equicontinuous. Let $V\in\mathcal{U}(\hat{E})$, and let $t\in\mathbf{R}_+$. $\{(\hat{T}(t+h)-\hat{T}(t))/h \mid |h|<\varepsilon/2,\ t+h\in\mathbf{R}_+\}$ is also equicontinuous, and so there is a $U\in\mathcal{U}(\hat{E})$ such that

$$\bigcup_{\substack{|h|<\varepsilon/2\\ t+h\in\mathbf{R}_+}}\left(\frac{\hat{T}(t+h)-\hat{T}(t)}{h}-\hat{T}(t)\circ(\mathcal{G}_T)\hat{}\right)(U)\subset V.$$

Let $x\in E$. Since E is dense in $\hat{E}$, there is a $y\in E$ such that $x-y\in U$. Thus,

$$\bigcup_{\substack{|h|\leqq\varepsilon/2\\ t+h\in\mathbf{R}_+}}\left(\frac{\hat{T}(t+h)-\hat{T}(t)}{h}-\hat{T}(t)\circ(\mathcal{G}_T)\hat{}\right)(x-y)\in V.$$

Hence

$$\lim_{h\to 0}\left(\frac{\hat{T}(t+h)-\hat{T}(t)}{h}-\hat{T}(t)\circ(\mathcal{G}_T)\hat{}\right)(x-y)=0,$$

and since

$$\lim_{h\to 0}\left(\frac{\hat{T}(t+h)-\hat{T}(t)}{h}-\hat{T}(t)\circ(\mathcal{G}_T)\hat{}\right)(y)=0,$$

it follows that

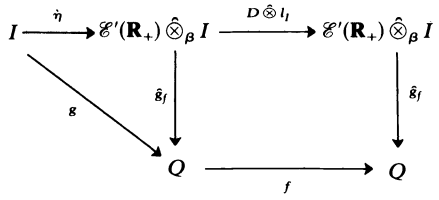
$$\lim_{h\to 0}\left(\frac{\hat{T}(t+h)-\hat{T}(t)}{h}-\hat{T}(t)\circ(\mathcal{G}_T)\hat{}\right)(x)=0.$$

Hence $T\in\mathcal{E}^1(\mathbf{R}_+, L_s(E))$, so (s_3) is satisfied, and T is a d.s.g. □

Thus, if I is a quasi-complete (resp. complete) l.c.s., $\mathcal{E}'\otimes_\beta I$ extends uniquely to a d.s.g. on $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ (resp. $\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$). This extension will be denoted $\mathcal{E}'\hat{\otimes}_\beta I$ (resp. $\mathcal{E}'\hat{\otimes}_\beta I$).

The desired extension of Theorem 2.4, which follows readily from the preceding, may now be stated.

THEOREM 2.8. (a) *Let I be a quasi-complete l.c.s., let (Q, f) be a differentiable system dynamics in* **QC**, *and let $g: I\to Q$ be a continuous linear map. The pair $(\hat{\eta}, (\mathcal{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I, D\hat{\otimes}1_I))$ is a free differentiable system dynamics over I in* **QC**. *$\hat{g}_f$ is the unique continuous linear map which makes the diagram below commute.*



(b) *A completely analogous result holds for* **CS**. □

The problem of interpreting Theorem 2.8 in a system-theoretic context remains. Of course, it is just an extension of Theorem 2.3, but it is helpful to have some sort of further characterization of the spaces $\mathscr{E}'(\mathbf{R}_+)\otimes_\beta I$, $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$, and $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$. To analyze these spaces, $\mathscr{E}'(\mathbf{R}_+)$ is first examined in a bit more detail. A distribution $\lambda \in \mathscr{E}'(\mathbf{R}_+)$ has compact support, and may be interpreted as a generalized-function input signal which starts at (is zero before) $t = -\max(\text{supp}(\lambda))$ (supp means "support of") and ends at time $t = 0$. The space $\mathscr{D}^0(\mathbf{R}_+)$ of all continuous functions with compact support may be canonically embedded in $\mathscr{E}'(\mathbf{R}_+)$. In fact, each element of $\mathscr{E}'(\mathbf{R}_+)$ is of the form $\sum_{k=0}^{n} D^k f_k$, where $f_k \in \mathscr{D}^0(\mathbf{R}_+)$ for $0 \leqq k \leqq n$, and the derivatives are in the generalized sense (see Rudin (1973, 6.27)). Each element of $\mathscr{E}'(\mathbf{R}_+)\otimes_\beta I$ may then be identified as a vector-valued distribution with compact, finite-dimensional support, in a manner analogous to the representation of $\Delta(\mathbf{R}_+)\otimes_\beta I$ given previously. Thus, the space of input signals $I^\uparrow$ consists at least of elements of the form $\sum_{k=1}^{n} i_k \cdot D^{p_k} f_k$, with $i_k \in i$, $p_k \in \mathbf{N}$, $f_k \in \mathscr{D}^0(\mathbf{R}_+)$, for all $1 \leqq k \leqq n$. While this is a rather rich space of inputs, the quasi-completion or completion operator adds quite a bit more, in general.

The space of all I-valued distributions on $\mathbf{R}_+$ with scalarly-compact support is defined to be $L_b(\mathscr{E}(\mathbf{R}_+), I)$ (see Schwartz (1957, 59, Ch. 1, p. 52)). Each element of $\mathscr{E}'(\mathbf{R}_+)\otimes_\beta I$ may be identified with an element of $L_b(\mathscr{E}(\mathbf{R}_+), I)$ via $\sum f_k \otimes i_k = \sum i_k \cdot f_k$. Unfortunately, the topology which $L_b(\mathscr{E}(\mathbf{R}_+), I)$ induces on $\mathscr{E}'(\mathbf{R}_+)\otimes I$ is (by definition) the ε topology (see Grothendieck (1955) for definition and properties of this topology) and not the β topology. Since $\mathscr{E}'(\mathbf{R}_+)$ is nuclear, the ε topology is exactly the π topology (again see Grothendieck (1955)). Hence, the problem of showing when $\mathscr{E}'(\mathbf{R}_+)\otimes_\beta I$ and its quasi-completion and completion may be identified with $L_b(\mathscr{E}(\mathbf{R}_+), I)$ or one of its subspaces reduces to determining when the β and π topologies coincide. The next result answers this question for (DF) spaces.

THEOREM 2.9. *Let I be a quasi-complete* (DF) *space.*

(a) *I is complete.*

(b) $\mathscr{E}'(\mathbf{R}_+)\otimes_\beta I = \mathscr{E}'(\mathbf{R}_+)\otimes_\pi I$.

(c) $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I = \mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I \cong L_b(\mathscr{E}(\mathbf{R}_+), I)$, *the last isomorphism being the completion of the canonical injection.*

(d) *Every element of $L_b(\mathscr{E}(\mathbf{R}_+), I)$ has compact support.*

Proof. (a) Consult Grothendieck (1973, Ch. 4, Pt. 3, Prop. 4, Cor. 2).

(b) It suffices to show that every hypocontinuous bilinear map on $\mathscr{E}'(\mathbf{R}_+)\times I$ is continuous, which is true since $\mathscr{E}'(\mathbf{R}_+)$ and I are (DF) spaces (see Grothendieck (1973, Ch. 4, Pt. 3, Th. 1)).

(c) The projective (π) tensor product of two (DF) spaces is a (DF) space, as is its completion (see Grothendieck (1955, Ch. 1, Prop. 5)). Hence, the first equality follows from (a). $L_b(\mathscr{E}(\mathbf{R}_+), I)$ is complete (see Grothendieck (1973, Ch. 3, Prop. 3)), and so the isomorphism follows from the nuclearity of $\mathscr{E}'(\mathbf{R}_+)$ (which makes the π and ε topologies equal).

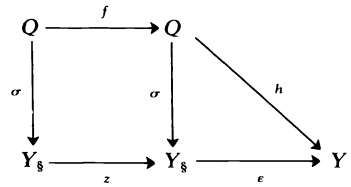
(d) This is a theorem of Schwartz (1957, 59, Ch. I, pp. 62–63). □

The above shows that in case the space I is a (DF) space, then in both the **QC** and **CS** constructions, the space of inputs $I^{\uparrow}$ may be identified with $L_b(\mathscr{E}(\mathbf{R}_+), I)$, and each element has compact support. While this is admittedly a special case, it does cover many important applications. For example, every normable space (and in particular every (B) space) is a (DF) space.

The question of whether, in the case of I a general quasi-complete (resp. complete) l.c.s., $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ (resp. $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$) can be regarded (algebraically) as a subspace of $L(\mathscr{E}(\mathbf{R}_+), I)$ is an injectivity problem, typical to topological-tensor-product theory. Such injectivity problems are in general open questions. See Grothendieck (1955, Ch. I, § 3, no. 2) for a discussion of related problems.

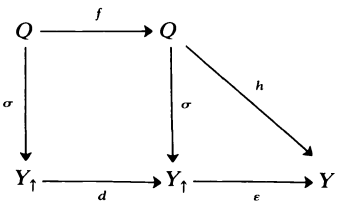
The rest of the interpretation of the **QC** and **CS** cases is a straightforward extension of the **LCS** case. Further details are not given here.

Output behavior. Returning briefly to the discrete-time case, let $\mathscr{K}$ be a category and let Y be an object of $\mathscr{K}$. A *cofree system dynamics* over Y is a system dynamics $(Y_\S, z)$ and a $\mathscr{K}$ morphism $\varepsilon: Y_\S \to Y$ such that for any other system dynamics (Q, f) and $\mathscr{K}$ morphism $h: Q \to Y$, there is a unique dynamorphism $\sigma: (Q, f) \to (Y_\S, z)$ such that



commutes. Cofree system dynamics are unique up to isomorphism, and exist in many categories, including **LS**. In **LS**, a cofree system dynamics is given by $Y = \{(y_0, y_1, \cdots, y_n, \cdots) | y_k \in Y\}$, $z: Y_\S \to Y_\S$ is left shift with dropping the leftmost term, and $\varepsilon: Y_\S \to Y$ is projection of the leftmost factor. If $M = (Q, f, I, g, Y, h)$ is a decomposable system in $\mathscr{K}$, σ is called the *observability map* of M. When $\mathscr{K} = \mathbf{LS}$, σ is given by $q \mapsto (h(q), h \circ f(q), \cdots, h \circ f^n(q), \cdots)$. Consult Arbib and Manes (1974) for details.

The continuous-time case is analogous. Let $\mathscr{K}$ be a subcategory of **LCS**. For a $\mathscr{K}$ object Y, a *cofree differentiable system dynamics* over Y is a differentiable system dynamics $(Y_\uparrow, d)$ in $\mathscr{K}$ and a $\mathscr{K}$ morphism $\varepsilon: Y_\uparrow \to Y$ such that for any other differentiable system dynamics (Q, f) in $\mathscr{K}$ and $\mathscr{K}$ morphism $h: Q \to Y$, there is a unique dynamorphism $\sigma: (Q, f) \to (Y_\uparrow, d)$ such that



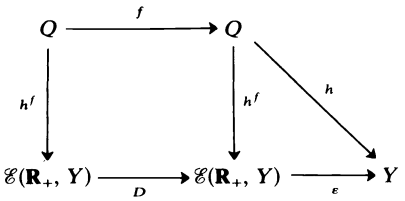
commutes. Cofree differentiable system dynamics are unique up to isomorphism, if they exist. If $M = (Q, f, I, g, Y, h)$ is a differentiable decomposable system in $\mathscr{K}$, σ is called the *observability map* of *M for continuous time.* It will be called just the observability map of M for the rest of this paper; the discrete-time observability map will be explicitly qualified as such. In contrast to the free case, it is a pleasant surprise to discover that the construction of cofree dynamics is very easy. The space $Y_\uparrow$ will turn out to be $\mathscr{E}(\mathbf{R}_+, Y)$ in each of the cases **LCS**, **QC**, amd **CS**. As shown next, all three cases may be treated at once.

LEMMA 2.10. *Let Y be a l.c.s. If Y is quasi-complete (resp. complete), then $\mathscr{E}(\mathbf{R}_+, Y)$ is also quasi-complete (resp. complete).*

Proof. By Theorem 1.6, $\mathscr{E}(\mathbf{R}_+, Y)$ may be identified with $L_b(\Delta(\mathbf{R}_+), Y)$. If Y is quasi-complete, this space may be identified with $L_b(\mathscr{E}'(\mathbf{R}_+), Y)$, since by Lemma 1.4, every bounded subset of $\mathscr{E}'(\mathbf{R}_+)$ is contained in the closure of a bounded subset $\Delta(\mathbf{R}_+)$, and $\mathscr{E}'(\mathbf{R}_+)$ is a quasi-completion of $\Delta(\mathbf{R}_+)$. Since $\mathscr{E}'(\mathbf{R}_+)$ is barreled, $L_b(\mathscr{E}'(\mathbf{R}_+), Y)$ is quasi-complete (see Schaefer (1971, Ch. III, 4.4 Cor.)). Since $\mathscr{E}'(\mathbf{R}_+)$ is also bornological, $L_b(\mathscr{E}'(\mathbf{R}_+), Y)$ is complete whenever Y is complete (see Treves (1967, Thm. 32.2, Cor.)). □

Define $\varepsilon: \mathscr{E}(\mathbf{R}_+, Y) \to Y$ by $\varepsilon(\varphi) = \varphi(0)$. ε is linear and continuous because $\varepsilon(\varphi) = \Phi_Y(\delta_0, \varphi)$. Define $h^f: Q \to \mathscr{E}(\mathbf{R}_+, Y)$ by $q \mapsto (t \mapsto h \circ \mathscr{T}_f(t)q)$. In view of Proposition 1.1(d), h^f is continuous.

THEOREM 2.11. *Let $\mathscr{K} =$ **LCS**, **QC**, or **CS**. Let Y be a l.c.s. in $\mathscr{K}$, let (Q, f) be a differentiable system dynamics in $\mathscr{K}$, and let $h: Q \to Y$ be a continuous linear map. The pair $(\varepsilon, (\mathscr{E}(\mathbf{R}_+, Y), D))$ is a cofree differentiable system dynamics over Y. h^f is the unique continuous linear map which makes the diagram below commute.*



Proof. In view of Lemma 2.10, the proof is identical in each of the three cases. The triangle commutes by definition of h^f and ε. The square commutes because $D \circ h^f(q) = D(t \mapsto h \circ \mathscr{T}_f(t)q) = t \mapsto h \circ \mathscr{T}_f(t) \circ f(q) = h^f \circ f(q)$. Uniqueness is obvious. □

The interpretation is very simple. $h^f(q)$ is the response of the system

$$\frac{dq(t)}{dt} = f(q(t)) + g(i(t)),$$

$$y(t) = h(q(t)),$$

starting in state q at $t = 0$, assuming that no other inputs are applied. Hence, it is just h applied to the natural response, i.e., $t \mapsto h(\mathscr{T}_f(t)q)$.

Behaviors in general. The ability to construct a canonical input–output behavior for a given system is crucial to the rest of this paper. Call a subcategory $\mathscr{K}$ of **LCS** *continuous-time behavioral* if for any $\mathscr{K}$ object E, both free and cofree differentiable system dynamics exist over E. A *continuous-time behavior* in $\mathscr{K}$ is defined to be a dynamorphism from a free differentiable system dynamics to a cofree differentiable system dynamics, i.e., a dynamorphism of the form $k: (I^{\uparrow}, d) \to (Y_{\uparrow}, d)$.

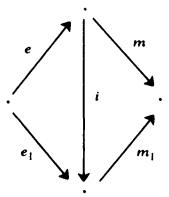
The categories $\mathscr{K} =$ **LCS**, **QC**, and **CS** are continuous-time behavioral, as was shown in this section. System-theoretically, this means that given any differentiable decomposable system $M = (Q, f, I, g, Y, h)$ in $\mathscr{K}$, there is a canonical continuous-time behavior associated with M. Specifically, this behavior is given by $\sigma \circ \rho: (I^{\uparrow}, d) \to (Y_{\uparrow}, d)$, where $\rho: (I^{\uparrow}, d) \to (Q, f)$ is the reachability map (for continuous time) of M and $\sigma: (Q, f) \to (Y_{\uparrow}, d)$ is the observability map (for continuous time) of M.

In the next section, the problem of constructing a canonical realization of a continuous-time behavior will be addressed.

3. Realization, reachability, and observability.

General principles. Let $\mathcal{K} = \mathbf{LS}$, and let $M = (Q, f, I, g, Y, h)$ be a (discrete-time) decomposable system in $\mathcal{K}$. M is *reachable* if its discrete-time reachability map $\rho: I^{\uparrow} \to Q$ is surjective and *observable* if its discrete-time observability map $\sigma: Q \to Y_{\uparrow}$ is injective. In Arbib and Manes (1974), it is shown how to generalize these ideas to arbitrary categories. The key idea is the following.

Let $\mathcal{K}$ be any category. A morphism f in $\mathcal{K}$ is called an *epimorphism* if for any morphisms g and h with $g \circ h$ and $h \circ f$ defined, $g \circ f = h \circ f \Rightarrow g = h$. In **LS**, the epimorphisms are precisely the surjections. Dually, f is a *monomorphism* in $\mathcal{K}$ if for any morphisms g and h with $f \circ g$ and $f \circ h$ defined, $f \circ g = f \circ h \Rightarrow g = h$. In **LS**, the monomorphisms are precisely the injections. An *image-factorization system* for $\mathcal{K}$ is an ordered pair $(\mathbf{E}, \mathbf{M})$ such that $\mathbf{E}$ is a class of epimorphisms in $\mathcal{K}$ and $\mathbf{M}$ is a class of monomorphisms in $\mathcal{K}$, each closed under composition and each containing all isomorphisms, such that each $\mathcal{K}$ morphism f has a factorization $f = m \circ e$ with $e \in \mathbf{E}$ and $m \in \mathbf{M}$ which is unique up to isomorphism in the sense that if $f = m_1 \circ e_1$ is another such factorization, then there is an isomorphism i such that the diagram



commutes. (e, m) is called an $(\mathbf{E}, \mathbf{M})$ *factorization* of f.

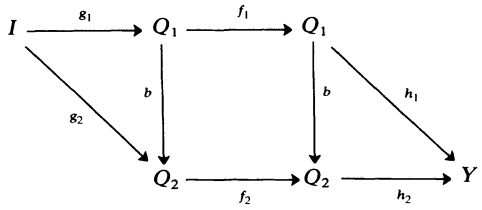
In **LS**, the only image-factorization system is (surjections, injections). Therefore, the following definition is compatible with that given above. Let M be a decomposable system in $\mathcal{K}$, and let $(\mathbf{E}, \mathbf{M})$ be an image-factorization system for $\mathcal{K}$. M is $\mathbf{E}$-*reachable* if its discrete-time reachability map is in $\mathbf{E}$, and $\mathbf{M}$-*observable* if its discrete-time observability map is in $\mathbf{M}$.

The categorical formulation of reachability and observability extends to the continuous-time case by replacing $I^{\S}$ (resp. $Y_{\S}$) by $I^{\uparrow}$ (resp. $Y_{\uparrow}$). Specifically, let $\mathcal{K}$ be a continuous-time behavioral category, let $M = (Q, f, I, g, Y, h)$ be a differentiable decomposable system in $\mathcal{K}$ and let $(\mathbf{E}, \mathbf{M})$ be an image-factorization system for $\mathcal{K}$. M is $\mathbf{E}$-*reachable for continuous time* if its reachability map for continuous time $\rho: I^{\uparrow} \to Q$ is in $\mathbf{E}$, and M is $\mathbf{M}$-*observable for continuous time* if its observability map for continuous time $\sigma: Q \to Y_{\uparrow}$ is in $\mathbf{M}$. The rest of this paper will deal exclusively with continuous time, so the adjective "continuous-time" will be dropped from reachable and observable.

Unlike the **LS** case of discrete time, in each of **LCS**, **QC**, and **CS**, there is more than one image-factorization system, so that reachability and observability will depend upon $(\mathbf{E}, \mathbf{M})$. Image-factorization systems for each of these three categories will be discussed in detail in this section.

The realization problem is the converse of the problem discussed in the previous section. Let $\mathcal{K}$ be a continuous-time behavioral category, and let I and Y be $\mathcal{K}$ objects. Define the category D-Sys$(\mathcal{K}, I, Y)$ to have as objects differentiable decomposable systems in $\mathcal{K}$ which have input space I and output space Y. A morphism $b: (Q_1, f_1, I, g_1, Y, h_1) \to (Q_2, f_2, I, g_2, Y, h_2)$ in D-Sys$(\mathcal{K}, I, Y)$ is a dynamorphism

$b:(Q_1, f_1) \to (Q_2, f_2)$ in D-Dyn($\mathcal{K}$) such that

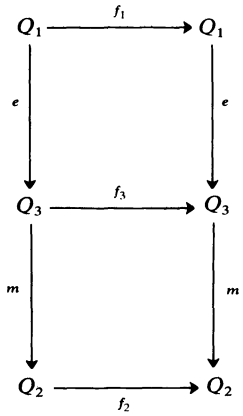


commutes. Given a continuous-time behavior $k:(I^{\uparrow}, d) \to (Y_{\uparrow}, d)$ in $\mathcal{K}$, a *realization* of k is a system in D-Sys($\mathcal{K}$, I, Y) such that $\sigma \circ \rho = k$, where ρ is the reachability map of M and σ is the observability map of M. Let (**E**, **M**) be an image-factorization system for $\mathcal{K}$. M is an (**E**, **M**)-*canonical* realization of k if it is both **E**-reachable and **M**-observable. (**E**, **M**) is *compatible* if every continuous-time behavior in $\mathcal{K}$ has an (**E**, **M**)-canonical realization. Existence and uniqueness of canonical realizations will now be investigated.

In the discrete-time case, the dynamorphic-image lemma was used to guarantee the existence of canonical realizations (see Arbib and Manes (1974, 4.4)). This lemma may also be used in the continuous-time case, although additional conditions must be imposed to insure that the fill-in is a differentiable system dynamics. The details follow.

LEMMA 3.1. *Let $\mathcal{K}$ be a continuous-time behavioral category and* (**E**, **M**) *an image-factorization system for $\mathcal{K}$. Let* (Q_1, f_1) *and* (Q_2, f_2) *be differentiable system dynamics in $\mathcal{K}$, let* $r:(Q_1, f_1) \to (Q_2, f_2)$ *be a dynamorphism in* D-Dyn($\mathcal{K}$) *and let* $Q_1 \xrightarrow{e} Q_3 \xrightarrow{m} Q_2$ *be an* (**E**, **M**) *factorization of r in $\mathcal{K}$.*

(a) *There is a unique $\mathcal{K}$ morphism $f_3: Q_3 \to Q_3$ such that*



commutes.

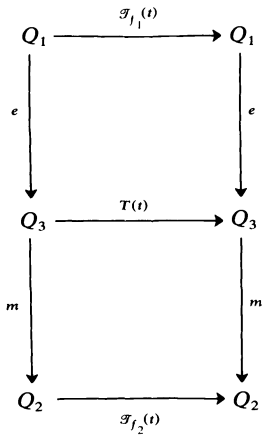
(b) *If either e is a quotient map or m is an embedding, (Q_3, f_3) is in* D-Dyn($\mathcal{K}$).

(c) (**E**, **M**) *induces an image-factorization system for* D-Dyn($\mathcal{K}$) *whenever* $\mathbf{E} \subset$ *quotient maps or* $\mathbf{M} \subset$ *embeddings.*

(d) *If* $\{[\mathcal{T}_{f_1}(t) - 1]/t \mid 0 < t \leqq \varepsilon\}$ *is equicontinuous for some* $\varepsilon > 0$, *and e is a near quotient,* (Q_3, f_3) *is in* D-Dyn($\mathcal{K}$) *and unique up to isomorphism (although* (**E**, **M**) *may not induce an image-factorization system on* D-Dyn($\mathcal{K}$) *in this case).*

Proof. (a) The existence and uniqueness of f_3 is guaranteed by the dynamorphic-image lemma (see Arbib and Manes (1974, 4.4)).

(b) Start by noting that the dynamorphic-image lemma also guarantees the existence of a fill-in $T(t)$ for each $t \in \mathbf{R}_+$ as indicated below.

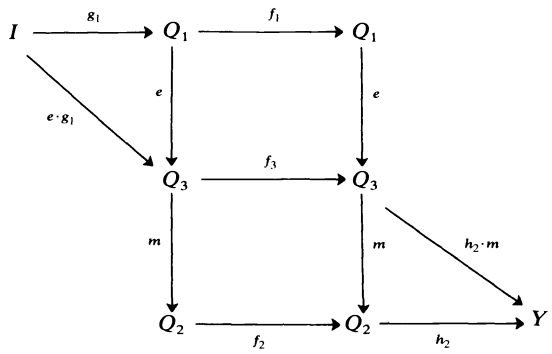


Clearly $T(t)$ satisfies (s_1) and (s_2). If e is surjective, the continuity of e immediately gives $\lim_{t\to 0}[(T(t)x-x)/t]=f_3(x)$ for each $x\in Q_3$, so (s_3) is satisfied. The equicontinuity property (s_4) follows easily from the quotient property. If m is an embedding, T is essentially a sub-d.s.g. of $\mathcal{T}_{f_3}$, so (s_3) and (s_4) are easily seen to be satisfied.

(c) This follows routinely from (b).

(d) Suppose e is a near quotient and $\{(\mathcal{T}_{f_1}(t)-1)/t \mid 0<t\leqq\varepsilon\}$ is equicontinuous, where $\varepsilon>0$. Because e is a near quotient, $\{(\bar{T}(t)-1)/t \mid 0<t\leqq\varepsilon\}$ is equicontinuous, where $\bar{T}$ is T (in the above diagram) restricted to $e(Q_1)$. By (b), $\bar{T}$ is a d.s.g. on $e(Q_1)$ with infinitesimal generator $f|_{Q_3}$. By Lemma 2.7, $\bar{T}$ may be extended to a d.s.g. $\hat{T}$ on $\hat{Q}_3$. By the uniqueness of the fill-in, the infinitesimal generator of $\hat{T}$ must be $\hat{f}_3$. Now restricting $\hat{T}$ to Q_3 yields a d.s.g. T with infinitesimal generator f_3, because $\hat{f}_3$ takes Q_3 into Q_3. The uniqueness of (Q_3, f_3) up to isomorphism is a consequence of the dynamorphic-image lemma. □

Note that the factorizations guaranteed by Lemma 3.1 easily lift from D-Dyn($\mathcal{K}$) to D-Sys($\mathcal{K}$, I, Y) for any I, Y in $\mathcal{K}$, for if $r:(Q_1, f_1, I, g_1, Y, h_1)\to(Q_2, f_2, I, g_2, Y, h_2)$ is a morphism in D-Sys($\mathcal{K}$, I, Y) and $Q_1 \overset{e}{\to} Q_3 \overset{m}{\to} Q_2$ is an (**E**, **M**) factorization of r which lifts to D-Dyn($\mathcal{K}$) with $(Q_1, f_1)\overset{e}{\to}(Q_3, f_3)\overset{m}{\to}(Q_2, f_2)$, this factorization also lifts to D-Sys($\mathcal{K}$, I, Y), as illustrated below.



From this observation and Lemma 3.1 follow readily conditions under which canonical realizations exist and are unique up to isomorphism.

THEOREM 3.2. *Let $\mathcal{K}$ be a continuous-time behavioral category,* (**E**, **M**) *an image-factorization system for $\mathcal{K}$, and (I, Y) a pair of $\mathcal{K}$ objects. Each behavior $k:(I^{\uparrow}, d)\to$*

$(Y_\uparrow, d)$ in D-Sys$(\mathcal{K}, I, Y)$ *has a canonical realization which is unique up to isomorphism in* D-Sys$(\mathcal{K}, I, Y)$ *if any of the conditions below is satisfied.*

(i) $\mathbf{E} \subset$ *quotient maps.*

(ii) $\mathbf{M} \subset$ *embeddings.*

(iii) $\mathbf{E} \subset$ *near quotients and* $(I^\uparrow, d)$ *is such that* $\{(\mathcal{T}_d(t)-1)/t, 0<t\leqq\varepsilon\}$ *is equicontinuous for some* $\varepsilon>0$. □

Examples. The three categories **LCS**, **QC**, and **CS** will now be studied in the context of canonical realizations. The first step is to characterize epimorphisms and monomorphisms in these categories.

PROPOSITION 3.3. *Let $\mathcal{K}$ be a full subcategory of* **LCS** *which contains* **K**, *and let f be a $\mathcal{K}$ morphism.*

(a) *f is an epimorphism if and only if f is dense.*

(b) *f is a monomorphism if and only if f is injective.*

Proof. (a) Suppose $f: E \to F$ is a dense map, and g and h are morphisms such that $g \circ f = h \circ f$. Now $g(x) = h(x)$ for all $x \in f(E)$, by construction. However, $\{x \in F \mid g(x) = h(x)\}$ is closed in F (see Bourbaki (1966, Ch. I, § 8.1, Prop. 2, Cor. 1)). Hence, $g = h$. Conversely, assume $f: E \to F$ is a $\mathcal{K}$ morphism which is not dense. Pick $x \in F \backslash \overline{f(E)}$, and let $g \in F'$ with $g(x) = 1$ and $g(\overline{f(E)}) = 0$ (the existence of such a function is guaranteed by the Hahn–Banach theorem). Now let $h: \mathbf{K} \to E$ be any nonzero linear map (necessarily continuous). $h \circ g$ is a $\mathcal{K}$ morphism, and $(h \circ g) \circ f = 0 \circ f$, yet $h \circ g \neq 0$. Hence, f is not an epimorphism.

(b) Every injection is clearly a monomorphism. Conversely, let $f: E \to F$ be a monomorphism which is not injective. Pick $x \in \ker(f)$, and define $g: \mathbf{K} \to E$ by $a \mapsto a \cdot x$. g is necessarily continuous. Now let $h \in F' \backslash \{0\}$. Clearly, $f \circ (g \circ h) = f \circ 0$, yet $g \circ h \neq 0$. Hence f is not a monomorphism. □

The **LCS** case will be treated first.

THEOREM 3.4. *Each of the following is a compatible image-factorization system for* **LCS**.

(a) (*quotient maps, injections*).

(b) (*surjections, embeddings*).

(c) (*dense maps, closed embeddings*).

Proof. Each of the factorizations is standard (see Köthe (1969, § 15, 4 (3 and 4))). Let $f: E \to F$ be a continuous linear map. $E \xrightarrow{f_1} E/\ker(f) \xrightarrow{f_2} F$ with f_1 the canonical quotient map and $f_2: x \mapsto f(x)$ is a factorization for (a). $E \xrightarrow{f_3} f(E) \xrightarrow{f_4} F$ with $f_3: x \mapsto f(x)$ and f_4 the inclusion map is a factorization for (b). Finally, $E \xrightarrow{f_5} \overline{f(E)} \xrightarrow{f_6} F$ with $f_5: x \mapsto f(x)$ and $f_6: x \mapsto x$ is a factorization for (c). The compatibility in each case follows from Theorem 3.2. □

Thus, in the **LCS** case, there are at least three distinct concepts of reachability. The (quotient maps, injections) case is reachability in the strongest sense; each state may be reached and the topology from the input space is preserved. Correspondingly, observability is in the weakest sense; the observability map is injective and continuous, but no other properties are preserved. In the (surjections, embeddings) case, reachability is weaker in the sense that the reachability map need not be a homomorphism. However, observability is stronger in that the observability map is a homomorphism as well as an injection. Finally, in the (dense maps, closed embeddings) case, the reachability is weakest in the sense that states need not actually be reached, but rather only approached to an arbitrary degree of approximation. Conversely, observability is in the strongest sense; the observability map is an isomorphism of a closed subspace.

An obvious approach to extending the image-factorization systems of Theorem 3.4 to **QC** (resp. **CS**) is to construct the quasi-completion (resp. completion) of the image-factorization for **LCS**. That is, if $f: E \to F$ is a continuous linear map of quasi-complete (resp. complete) l.c.s.'s, and $E \overset{e}{\to} G \overset{m}{\to} F$ is an (**E**, **M**) factorization of f (where (**E**, **M**) is an image-factorization system for **LCS**), regard the factorization $E \overset{\hat{e}}{\to} \hat{G} \overset{\hat{m}}{\to} F$ (resp. $E \overset{\hat{e}}{\to} \hat{G} \overset{\hat{m}}{\to} F$) as the induced factorization in **QC** (resp. **CS**). This technique works for each of the image-factorization systems of Theorem 3.4, although the proof for the (quotient maps, injections) case is not trivial.

Since a closed subspace of a quasi-complete (resp. complete) l.c.s. is itself quasi-complete (resp. complete), **QC** (resp. **CS**) inherits (dense maps, closed embeddings) as an image-factorization system.

THEOREM 3.5. *(dense maps, closed embeddings) is a compatible image-factorization system for* **QC** *(resp.* **CS**). □

Thus, the weakest kind of reachability, together with the strongest kind of observability, extends naturally to **QC** and **CS**. This is not the case for the other image-factorization systems.

Since the completion of a linear subspace of a complete l.c.s. is closed, it is immediate that the extension of the **LCS** image-factorization system (surjections, embeddings) to **CS** is just (dense maps, closed embeddings). The extension for **QC** requires a new definition. Call a **QC** morphism $f: E \to F$ a *quasi-surjection* if F is isomorphic to a quasi-completion of $f(E)$ (when $f(E)$ has the relative topology induced by F), and call f a *quasi-closed embedding* if it is an embedding and $f(E)$ is quasi-closed in F. ($f(E)$ is quasi-closed in F if each closed and bounded subset of $f(E)$ is closed in F.)

THEOREM 3.6. *(quasi-surjections, quasi-closed embeddings) is an image-factorization system for* **QC**.

Proof. The factorization is just the quasi-completion of the (surjections, embeddings) factorization in **LCS**. That is, if $f: E \to F$ is a continuous linear map of quasi-complete l.c.s.'s and $E \overset{e}{\to} G \overset{m}{\to} F$ is a (surjections, embeddings) factorization of f in **LCS**, then $E \overset{e}{\to} \hat{G} \overset{\hat{m}}{\to} F$ is a (quasi-surjections, quasi-closed embeddings) factorization in **QC**. The verification of the image-factorization system properties is routine. The compatibility is a straightforward consequence of Theorem 3.2 and Lemma 2.7. □

The extension of (quotient maps, injections) is substantially more difficult to handle than the other two. The less-than-obvious part of the construction is showing that the quasi-completion (resp. completion) of the monomorphic part of the factorization is monomorphic. That is, show that $\hat{m}$ (resp. $\hat{m}$) is injective. To do this requires the use of some relatively advanced concepts from category theory. The definitions may be found in Schubert (1972). The reader who is not familiar with these concepts should read the definition of quasi-quotient below and then skip to Theorem 3.9.

A **QC** morphism $f: E \to F$ is called a *quasi-quotient* if it is a homomorphism and F is a quasi-completion of $f(E)$ (regarded as a subspace of F).

LEMMA 3.7. *Let E and F be quasi-complete (resp. complete) l.c.s.'s, let $f: E \to F$ be a continuous linear map, and let $G = \{(x, y) \in E \times E \mid x - y \in \ker(f)\}$.*

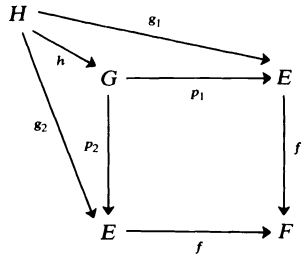
(a) *G is a quasi-complete (resp. complete) l.c.s., regarded as a linear subspace of $E \times E$.*

(b) *$G \overset{p_1, p_2}{\rightrightarrows} E$ is a kernel pair of $E \times E$, where $p_1: (x, y) \to x$ and $p_2: (x, y) \mapsto y$.*

(c) *f is a coequalizer in* **QC** *(resp.* **CS**) *if and only if it is a quasi-quotient (resp. near quotient), and in this case it is a coequalizer of $G \overset{p_1, p_2}{\rightrightarrows} E$.*

Proof. (a) Since the product of quasi-complete (resp. complete) l.c.s.'s is quasi-complete (resp. complete), it suffices to show that G is closed in $E \times E$. However, G is the kernel of the continuous linear map $E \times E \overset{f \times f}{\longrightarrow} F \times F \overset{(-)}{\longrightarrow} F$, whence it is closed.

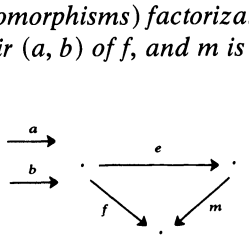
(b) Clearly $f \circ p_1 = f \circ p_2$. Let H be a quasi-complete (resp. complete) l.c.s. and let $g_1 : H \to E$ and $g_2 : H \to E$ be continuous linear maps such that $f \circ g_1 = f \circ g_2$. Define $h : H \to G$ by $x \mapsto (g_1(x), g_2(x))$. h is clearly linear and continuous, and it follows that the diagram



commutes. The uniqueness of h is clear. Hence, $G \overset{p_1,p_2}{\rightrightarrows} E$ is a kernel pair of f in **QC** (resp. **CS**).

(c) It is easy to verify that the quotient maps are precisely the coequalizers in **LCS**, and that the quotient map $q : E \to E/\overline{(p_1 - p_2)(G)}$ is a coequalizer of $G \overset{p_1,p_2}{\rightrightarrows} E$ in **LCS**. Also, note that $\hat{}\,: \mathbf{LCS} \to \mathbf{QC}$ (resp. $\hat{}\,: \mathbf{LCS} \to \mathbf{CS}$) may be regarded as a functor, and that $\hat{}$ (resp. $\hat{}$) has a right adjoint, namely the inclusion functor $\mathbf{QC} \to \mathbf{LCS}$ (resp. $\mathbf{CS} \to \mathbf{LCS}$), and so preserves all colimits, particularly coequalizers. Hence, f is a coequalizer if and only if it is the image under $\hat{}$ (resp. $\hat{}$) of a quotient map. □

LEMMA 3.8. *Let $\mathcal{K}$ be a category which has kernel pairs and coequalizers of kernel pairs, and suppose that the class of all coequalizers is closed under composition. Then (coequalizers, monomorphisms) is an image-factorization system for $\mathcal{K}$, and if f is a $\mathcal{K}$ morphism, a (coequalizers, monomorphisms) factorization of f is given by $f = m \circ e$, where e is a coequalizer of a kernel pair (a, b) of f, and m is the unique morphism making the diagram*



commute.

Proof. Consult Schubert (1972, 18.4.7). □

THEOREM 3.9. (a) *(quasi-quotients, injections) is a compatible image-factorization system for* **QC**.

(b) *(near quotients, injections) is a compatible image-factorization system for* **CS**.

Proof. That each is an image-factorization system follows from Lemmas 3.7 and 3.8. The compatibility follows from Theorem 3.2 and Lemma 2.7. □

Thus, in summary, it is seen that there are several distinct concepts of reachability and observability for infinite-dimensional linear systems. This is true not only of continuous-time systems, but for topologized discrete-time systems as well (see Hegner (1978b)).

4. Finite port systems.

Basic ideas. In this section, some detailed properties of differentiable decomposable systems $M = (Q, f, I, g, Y, h)$ for which I and Y are finite-dimensional will be considered. Such systems will be termed *finite port*, to emphasize that it is the input and output spaces, and not necessarily the state space, with the property of finite dimensionality. Without loss of generality, it shall be assumed that when M is a finite

port differentiable decomposable system, $I = \mathbf{K}^m$ and $Y = \mathbf{K}^p$ for some nonnegative integers m and p. Such a system will also be called m-input p-output. A continuous-time behavior $B: (\mathbf{K}^{m\uparrow}, d) \to (\mathbf{K}^p_\uparrow, d)$ in any **LCS** continuous-time behavioral category will be termed a *finite port behavior.*

LEMMA 4.1. *There are the following natural isomorphisms.*

(a) $\Delta(\mathbf{R}_+) \otimes_\beta \mathbf{K}^m \cong (\Delta(\mathbf{R}_+))^m$.

(b) $\mathscr{E}'(\mathbf{R}_+) \otimes_\beta \mathbf{K}^m \cong \mathscr{E}'(\mathbf{R}_+) \hat{\otimes}_\beta \mathbf{K}^m \cong \mathscr{E}'(\mathbf{R}_+) \hat{\otimes}_\beta \mathbf{K}^m \cong (\mathscr{E}'(\mathbf{R}_+))^m$.

(c) $\mathscr{E}(\mathbf{R}_+, \mathbf{K}^p) \cong (\mathscr{E}(\mathbf{R}_+))^p$.

Proof. (a) For any l.c.s. F whatever, there is a natural map $F \otimes_\beta \mathbf{K}^m \to F^m$ given by $f \otimes (k_1, \cdots, k_m) \mapsto (k_1 f, \cdots, k_m f)$. It is easily verified that this map defines an isomorphism.

(b) From the proof of (a), $\mathscr{E}'(\mathbf{R}_+) \otimes_\beta \mathbf{K}^m \cong (\mathscr{E}'(\mathbf{R}_+))^m$ can be deduced also, whence the result, since $\mathscr{E}'(\mathbf{R}_+)$ is complete.

(c) The natural map $\mathscr{E}(\mathbf{R}_+, \mathbf{K}^p) \to (\mathscr{E}(\mathbf{R}_+))^p$ given by $\mathscr{E}(\mathbf{R}_+, \mathbf{K}^p) \to (\mathscr{E}(\mathbf{R}_+))^p$ given by $f \mapsto (\pi_1 \circ f, \cdots, \pi_p \circ f)$, with π_k the projection onto the kth element, is easily seen to be an isomorphism. □

A finite port differentiable decomposable system, as defined above, has m input lines and p output lines. The signals applied to each input line are elements of $\Delta(\mathbf{R}_+)$ in the **LCS** case, and elements of $\mathscr{E}'(\mathbf{R}_+)$ in the **QC** and **CS** cases. The output signals on each line are elements of $\mathscr{E}(\mathbf{R}_+)$ in all of the cases. As explained in § 2, the input signals are interpreted as occurring in negative time ($-\infty \leqq t \leqq 0$), while the output signals occur for $t \geqq 0$, after nonzero inputs have ceased. Given a specific behavior $B: (\Delta(\mathbf{R}_+))^m \to (\mathscr{E}(\mathbf{R}_+))^p$ (or $B: (\mathscr{E}'(\mathbf{R}_+))^m \to (\mathscr{E}(\mathbf{R}_+))^p$), the $m \times p$ matrix over $\mathscr{E}(\mathbf{R}_+)$ whose (i, j)th entry is $\pi_j \circ B(in_i(\delta_0))$, where in_i is the ith canonical injection and π_j is the jth canonical projection is called the *weighting pattern* W_B of B. In words, the (i, j)th entry of W_B is the response on the jth output line for $t \geqq 0$ due to an impulse at $t = 0$ on the ith input line. Conversely, any $m \times p$ matrix over $\mathscr{E}(\mathbf{R}_+)$ is the weighting pattern of a unique behavior, as is shown below.

THEOREM 4.2. *Let W be any $m \times p$ matrix of elements of $\mathscr{E}(\mathbf{R}_+)$. Then W uniquely determines the behavior B of an m-input, p-output differentiable decomposable system in* **LCS**, **QC**, *and* **CS**.

Proof. Define $B: \Delta(\mathbf{R}_+)^m \to \mathscr{E}(\mathbf{R}_+)^p$ by $\pi_j \circ B(in_i(D^k \delta_t)) = D^k f_t$, where f is the (i, j)th element of W and $f_t(x) = f(t + x)$. B is a behavior by construction. In the cases of **QC** and **CS**, there is a unique extension to $(\mathscr{E}'(\mathbf{R}_+))^m$, since $\Delta(\mathbf{R}_+)$ is dense in $\mathscr{E}'(\mathbf{R}_+)$ (see Lemma 1.4(a)). □

Some properties of realizations. Some specific properties of the reachability map, observability map, and state space of certain realizations of a finite port differentiable decomposable system will now be developed.

LEMMA 4.3. *Let Q be a l.c.s., and let $k: (\mathscr{E}'(\mathbf{R}_+))^m \to Q$ be a continuous linear map.*

(a) *If k is a quotient map, then Q is complete.*

(b) *If k is surjective and Q is barreled, then f is a quotient map.*

Proof. (Consult Schaefer (1971, Ch. IV, § 8) for definitions and results used in this proof.)

(a) Since $(\mathscr{E}'(\mathbf{R}_+))^m$ is the strong dual of the reflexive (F) space $(\mathscr{E}(\mathbf{R}_+))^m$, it is B-complete. The quotient of a B-complete space is B-complete, hence complete, so Q is complete.

(b) This follows from (a) and the homomorphism theorem. □

Part (a) of Lemma 4.3 shows that for a finite port behavior B, the (quasi-quotients, injections) factorization in **QC** and the (near quotients, injections) in **CS** are each simply

the (quotient maps, injections) factorization. Furthermore, since $(\mathscr{E}(\mathbf{R}_+))^p$ is an (F) space, a quasi-closed subspace is already closed. Hence, the (quasi-surjections, quasi-closed embeddings) factorization in **QC** reduces to the (dense maps, closed embeddings) factorization. Thus, for the image-factorization systems considered in this report, there is no difference between the **QC** and **CS** cases for finite port systems.

THEOREM 4.4. *Let $\mathscr{K} = \mathbf{LCS}$, **QC**, or **CS**, let B be the behavior of a finite port differentiable decomposable system in $\mathscr{K}$, and let $M = (Q, f, I, g, Y, h)$ be a realization of B. If the reachability map ρ is a quotient map or the observability map σ is an embedding, then the state space Q is normable only if it is finite dimensional.*

Proof. Consult Grothendieck (1955, Ch. II) for definitions and results used in this proof. $\mathscr{E}(\mathbf{R}_+)$ and $\mathscr{E}'(\mathbf{R}_+)$ are each nuclear, and an appeal to the fact that subspaces and quotient spaces of nuclear spaces are nuclear shows that Q must be nuclear. However, a normable nuclear space must be finite dimensional, so Q can be normable only if it is finite dimensional. □

The above theorem has a rather remarkable implication. In each of the categories **LCS**, **QC**, and **CS**, equipped with any of the image-factorization systems developed in § 3, a canonical realization of a finite port differentiable decomposable system can have a normable state space only in the very special case of a finite-dimensional system. This has the further implication that unless the behavior admits a finite-dimensional realization, there are at least two distinct concepts of canonical realization, as is now shown.

THEOREM 4.5. *Let B be a finite port behavior in **LCS**, **QC** or **CS**. The (quotient maps, injection) factorization of B coincides with the (dense maps, closed embeddings) factorization if and only if the resulting state space (in each case) is finite dimensional.*

Proof. The state space in the (quotient maps, injections) case is a (DF) space, because $(\Delta(\mathbf{R}_+))^m$ and $(\mathscr{E}'(\mathbf{R}_+))^m$ are each (DF) spaces, and the quotient of a (DF) space is a (DF) space (Köthe (1969, § 29, 5.(1))). The state space in the (dense maps, closed embeddings) case is an (F) space because $(\mathscr{E}(\mathbf{R}_+))^p$ is an (F) space, and a closed subspace of an (F) space is clearly an (F) space. Hence, for the two realizations to coincide, the state space must be both an (F) space and a (DF) space. However, this implies that it is a (B) space, which, by Theorem 4.4, can happen only if it is finite dimensional. □

Finally, a result which characterizes the case of a barreled state space is presented.

THEOREM 4.6. *Let B be a finite port continuous-time behavior in any of the categories **LCS**, **QC**, or **CS**, let $M = (Q, f, I, g, Y, h)$ be a realization of B, and suppose the state space Q is barreled. Then if the reachability map ρ is surjective, it is necessarily a quotient map.*

Proof. In the **QC** and **CS** cases, when the space of input signals is $(\mathscr{E}'(\mathbf{R}_+))^m$, the result follows immediately from Lemma 4.3(b). In the **LCS** case, let $\hat{\rho}: (\mathscr{E}'(\mathbf{R}_+))^m \to \hat{Q}$ be the completion of the reachability map. It suffices to show that $\hat{\rho}$ is a quotient map, for then ρ will also be a quotient, completing the proof. It is immediate that if $k: E \to F$ is any dense embedding with E barreled, then F must be barreled also. Via the natural embedding $Q \hookrightarrow \rho((\mathscr{E}'(\mathbf{R}_+))^m)$, it then follows that $\rho((\mathscr{E}'(\mathbf{R}_+))^m)$ is barreled. Hence, in view of Lemma 4.3(b), $(\mathscr{E}'(\mathbf{R}_+))^m/\ker(\rho) \cong \rho((\mathscr{E}'(\mathbf{R}_+))^m)$. However, $(\mathscr{E}'(\mathbf{R}_+))^m/\ker(\rho)$ is already complete by Lemma 4.3(a), so it is complete dense subspace of $\hat{Q}$, hence it must be all of $\hat{Q}$. Thus $\hat{\rho}$ is a quotient map. □

COROLLARY 4.7. *Let B be a finite port continuous-time behavior in any of the categories **LCS**, **QC**, or **CS**, and let $M = (Q, f, I, g, Y, h)$ be a realization of B. If the reachability map ρ is surjective, then the state space Q cannot be an infinite-dimensional* (B) *space.* □

In closing, it should be noted that it can be shown that the results 4.4 through 4.7 hold (at least) in the more general case that the input space I is the strong dual of a reflexive nuclear (F) space and the output space Y is a nuclear (F) space. However, these extensions do not seem to be of sufficient interest to justify the additional complications involved in their presentation.

5. Remarks. Other workers in the algebraic theory of continuous-time linear systems have used an $\mathscr{E}'(\mathbf{R}_+)$ module framework rather than a semigroup approach. In order to compare their work to that of the present report, it is necessary to recast the d.s.g. approach in a module framework.

LEMMA 5.1. *$\Delta(\mathbf{R}_+)$ and $\mathscr{E}'(\mathbf{R}_+)$ are each commutative rings with unit δ_0. Addition is just that defined by the vector space structure; multiplication is convolution of distributions and is continuous.*

Proof. Proof of these properties for $\mathscr{E}'(\mathbf{R}_+)$ may be found in Treves (1967, Thm. 27.7). It is easy to see that $\Delta(\mathbf{R}_+)$ is a subring of $\mathscr{E}'(\mathbf{R}_+)$, so it inherits the required properties. □

LEMMA 5.2. (a) *There is a bijective correspondence between differentiable system dynamics and $\Delta(\mathbf{R}_+)$ modules whose action is hypocontinuous, the association being given by*

$$(Q, f) \mapsto (\tilde{1}_Q)_f : \Delta(\mathbf{R}_+) \times Q \to Q,$$

$$b : \Delta(\mathbf{R}_+) \times Q \to Q \mapsto (Q, q \mapsto b(\delta_0^{(1)}, q)).$$

(b) *There is a bijective correspondence between differentiable system dynamics on quasi-complete (resp. complete) spaces and $\mathscr{E}'(\mathbf{R}_+)$ modules on quasi-complete (resp. complete) spaces whose action is hypocontinuous, the association being given by the natural completions of those of* (a).

Proof. (a) Verification that $(\tilde{1}_Q)_f$ is a module action is trivial. Reference to Lemma 2.3 and the preceding discussion provides verification of the bijectivity and continuity properties.

(b) Follows from Lemma 2.5 and part (a). □

LEMMA 5.3. *Let (Q, f) be a differentiable system dynamics, with Q a* (DF) *space. Then the associated module action is continuous.*

Proof. Follows from Theorem 2.9(b). □

Example 5.4. Kalman and Hautus (1972) deal exclusively with finite port systems over the real field. A behavior in their framework is essentially a continuous $\mathscr{E}'(\mathbf{R}_+)$ module morphism $b : (\mathscr{E}'(\mathbf{R}_+))^m \to (\mathscr{E}(\mathbf{R}_+))^p$ for finite $m, p > 0$. (They explicitly reverse the time variable and use $\mathscr{E}'(\mathbf{R}_-)$ rather than $\mathscr{E}'(\mathbf{R}_+)$, but this is merely a matter of convenience.) As long as all spaces involved are at least quasi-complete (which is the case in their framework), a behavior in their sense is equivalent to a finite port behavior in **QC**. For simplicity, the rest of the discussion will be confined to the case $m = p = 1$.

Their approach differs substantially from the one presented here in the way in which they define a realization. Rather than invoking semigroup theory to ask how the behavior describes an internal realization, they attempt to construct a differential equation directly. However, as they observe, it is impossible to define a natural truncation operator $\mathscr{E}'(\mathbf{R}) \to \mathscr{E}'(\mathbf{R}_+)$. Thus, the complete meaning of a behavior in the algebraic sense, namely truncating the input for $t > 0$ and observing only the natural response, cannot be extended without modification to the continuous-time case. To remedy this situation, they define the state trajectory not as a function $t \mapsto q(t)$ of time, but rather as a function $\varphi \mapsto q(\varphi)$ of test functions. Let $\mathscr{D}$ be the space of infinitely differentiable functions $\mathbf{R} \to \mathbf{K}$ with compact support. The input at "time" $\varphi \in \mathscr{D}$ for

$i \in \mathscr{E}'(\mathbf{R}_+)$ is $(i^*\varphi)^\vee \in \mathscr{D} \subseteq \mathscr{E}'(\mathbf{R})$, where $^\vee : \mathscr{D} \to \mathscr{D} : \varphi(\cdot) \mapsto \varphi(-\cdot)$. Since $(i^*\varphi)^\vee \in \mathscr{D}$, truncation is possible. Let Tr: $\mathscr{D} \to \mathscr{E}'(\mathbf{R}_+)$ be defined by $\mathrm{Tr}(\varphi)(t) = \varphi(t)$ for $t \geqq 0$ and 0 for $t < 0$. This lifts to the subspace of $\mathscr{E}'(\mathbf{R}_+)$ consisting of those distributions defined by elements of $\mathscr{D}$. Factoring $b : \mathscr{E}'(\mathbf{R}_+) \to \mathscr{E}(\mathbf{R}_+)$ as an open surjection followed by an injection $\mathscr{E}'(\mathbf{R}_+) \twoheadrightarrow Q \rightarrowtail \mathscr{E}(\mathbf{R}_+)$, the state at "time" $\varphi \in \mathscr{D}$ under input $i \in \mathscr{E}'(\mathbf{R}_+)$ is $\rho(\mathrm{Tr}(i^*\varphi)^\vee)$. The part of the state space Q on which a dynamics is constructed is $(\mathrm{Tr}(\mathscr{D}))$. A differential equation is then defined on this space by

$$\frac{d}{dt} q(\varphi) = F(q(\varphi)) + G(i(\varphi)),$$

where

$$F : (\mathrm{Tr}(\mathscr{D})) \to (\mathrm{Tr}(\mathscr{D})) : (\mathrm{Tr}(\varphi)) \mapsto \left(\mathrm{Tr}\left(\frac{d}{dt}\varphi\right)\right),$$

$$G : \mathbf{R} \to Q : r \mapsto r \cdot \rho(\delta_0).$$

The output side of the dynamics is defined by

$$H : Q \to \mathscr{E}(\mathbf{R}_+) : q \mapsto \sigma(q)(0).$$

There are a number of drawbacks to this approach. First of all, the differential equation is not of the usual variety. It also appears difficult to consider any evolution of solutions of it, since $dq(\varphi)/dt$ need not be in the domain of F. At any rate, the dynamics is only defined on part of Q. The approach of this paper does not share these difficulties. Because all dynamics used are inherently endowed with a differential structure, this truncation problem does not enter into the construction of an internal dynamics. Thus it is possible to define a meaningful ordinary differential equation describing the dynamics on the whole state space. It is easily verified that if $(Q, f, \mathbf{K}, g, \mathbf{K}, h)$ is a **QC** realization of a behavior $b : \mathscr{E}'(\mathbf{R}_+) \to \mathscr{E}(\mathbf{R}_+)$ in the sense of this paper, then by defining $F = f|_{\mathrm{Tr}(\mathscr{D})}$, $G = g$, $H = h$, the triple (F, G, H) describing the internal dynamics in the sense of Kalman and Hautus is recovered. Thus, their dynamics is a restriction of that developed in this report.

Example 5.5. Bensoussan, Delfour and Mitter (1975) also use the module $\mathscr{E}'(\mathbf{R}_+)$ as a starting point. (Actually, as in the case of Kalman and Hautus, they use $\mathscr{E}'(\mathbf{R}_-)$.) A behavior in their framework is an $\mathscr{E}'(\mathbf{R}_+)$ module morphism $b : L_1(\mathscr{E}(\mathbf{R}_+), I) \to \mathscr{E}(\mathbf{R}_+, Y)$, where I and Y are reflexive separable (B) spaces, and $L_1(\mathscr{E}(\mathbf{R}_+), I)$ is the space of nuclear operators from $\mathscr{E}(\mathbf{R}_+)$ to I. However, since $\mathscr{E}(\mathbf{R}_+)$ is nuclear, $L_1(\mathscr{E}(\mathbf{R}_+), I) = L(\mathscr{E}(\mathbf{R}_+), I) \cong \mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I \cong \mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$. A proof of the first equality may be found in Schaefer (1971, Ch. 3, 7.2); the last two isomorphisms are found in Theorem 2.9. Thus, a behavior in the sense of Bensoussan, Delfour, and Mitter is a special case of the concept of behavior of this report, with the category being either **QC** or **CS**.

They factor the behavior using a (quotient maps, injections) factorization to get a natural state space. To verify that this corresponds to the factorization of this report, the following extension of Lemma 4.3 is needed.

LEMMA 5.6. *Given a reflexive* (B) *space I, a quotient of* $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ *is always complete.*

Proof. In view of the proof of Lemma 4.3, it is only necessary to establish that $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I$ is the strong dual of a reflexive (B) space. For a proof of a more general result which establishes this fact, consult Grothendieck (1955, Ch. 2, 4, Lem. 9). □

Example 5.5 (*continued*). Let $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I \xrightarrow{\pi} Q \xrightarrow{\eta} \mathscr{E}(\mathbf{R}_+, Y)$ be a (quotient maps, injections) factorization of a behavior b in the framework of Bensoussan, Delfour and Mitter. They define an internal representation (Q, F, I, G, Y, H) by identifying Q with $\mathscr{E}'(\mathbf{R}_+)\hat{\otimes}_\beta I/\ker(b)$ and defining $F: Q \to Q$ by $F(\rho(i)) = \delta_0^{(1)} \cdot \rho(i)$, $G(i) = \rho(\delta_0 \otimes i)$, $H(\rho(i)) = \rho(\sigma(i))(0)$. It is easy to see that this is exactly a (quasi-quotients, injections) canonical realization in **QC** and a (near quotients, injections) realization in **CS**, in view of Lemma 5.6. Because they use a direct construction rather than universal constructions to arrive at the internal description, a rather involved procedure using infinite series representation of nuclear operators is needed to show that the system behavior satisfies the differential equation. Nonetheless, their results do correspond exactly to realizations in **QC** and **CS** of systems whose input and output spaces are reflexive separable (B) spaces.

Neither the Kalman and Hautus paper nor the Bensoussan, Delfour and Mitter paper consider the problem of constructing a natural behavior from an internal description, a key feature of the work reported here.

Kamen (1971) has also used an $\mathscr{E}'(\mathbf{R}_+)$-module approach to describe continuous-time linear systems. However, his work is primarily concerned with a module approach to structure, and does not directly represent the internal dynamics with a differential equation.

Because an $\mathscr{E}'(\mathbf{R}_+)$ module gives rise under suitable conditions to a differentiable system dynamics, approaches using such modules as dynamics have a substantial common base with the work of this report, and so relatively detailed comparisons have been presented. On the other hand, there are numerous other approaches to the theory of continuous-time linear systems. Two of the more recent are discussed in the next examples. They differ fundamentally from differentiable decomposable systems in that they do not require that the dynamics be differentiable. Instead, other assumptions are made to provide a tractable theory. Only those aspects of these approaches which admit a reasonably compact comparison to the present framework are discussed. Characterization of systems within a categorical framework whose dynamics more closely resemble those in the following two examples is the subject of other reports (Hegner (1980a), (1980b), (1980c)), to which the reader is referred for more detailed discussions.

Example 5.7. Baras, Brockett, and Fuhrmann (1974), Baras and Brockett (1975), and Baras and Dewilde (1976) consider systems of the form $M = (Q, f, I, g, Y, h)$, where Q is a Hilbert Space, f is the infinitesimal generator of a (C_0) semigroup of operators on Q, I and Y are finite dimensional, and $g: I \to Q$ and $h: Q \to Y$ are continuous linear maps. Such systems are termed regular.

The weighting pattern of a regular system with $I = \mathbf{K}^m$ and $Y = \mathbf{K}^p$ is the $m \times p$ matrix of continuous functions from $\mathbf{R}_+$ to $\mathbf{K}$ whose (i, j)th entry is $t \mapsto \pi_j \circ h \circ T(t) \circ g \circ in_i$, where in_i is the injection into the ith component, π_j is the projection of the jth component, and T is the (C_0) semigroup generated by f. The elements of the weighting pattern matrix need not be infinitely differentiable, so there are regular systems which are not differentiable. On the other hand, a necessary condition that a matrix of continuous functions be the weighting pattern of a regular system is that it be of exponential order (Baras and Brockett (1975, Thm. 4)). Since there are infinitely differentiable functions which are not of exponential order, there are finite port differentiable decomposable systems which are not regular. Of course, the framework of this report also includes systems which are not finite port, whereas those cited above do not.

A more interesting comparison is that of the concepts of reachability, observability, and canonicity. In Baras and Dewilde (1976), the system M is called reachable

if for any $q \in Q$, $g' \circ T'(t)q = 0$, for all $t \geqq 0$, implies $x = 0$. Using duality, this transforms to the equivalent condition that $\{T(t) \circ g(i) \mid i \in I \text{ and } t \in \mathbf{R}_+\}$ is dense in Q. This set corresponds exactly to those states which can be reached by applying inputs which consist entirely of finite linear combinations of impulses occurring for $t \leqq 0$. In the context of differentiable decomposable systems, this transforms to the reachability map ρ being dense when restricted to the subspace of $(\Delta\mathbf{R}_+))^m$ consisting of those elements which are in the linear span of $\{\delta_t \mid t \in \mathbf{R}_+\}^m$. It is not difficult to show that this subspace is dense in $(\Delta(\mathbf{R}_+))^m$, so this concept of reachability just corresponds to ρ being a dense map. In the same paper, M is called observable if $h \circ T(t)q = 0$, for all $t \geqq 0$, implies $q = 0$. This is easily seen to be equivalent to the observability map $\sigma: q \mapsto h \circ T(t)q$ being injective. Finally, a canonical system in the framework of Baras and Dewilde is one which is both reachable and observable. Since in any reasonable category of l.c.s.'s, every epimorphism is dense and every monomorphism is injective (Proposition 3.3), this definition cannot distinguish between various types of canonicity. Theorem 4.5 shows that if this definition is used, there are always nonisomorphic canonical realizations of differentiable decomposable systems in **LCS**, **QC**, and **CS**, except in the case of a finite-dimensional state space. Thus, if one accepts the restriction that a canonical realization should be unique up to isomorphism, then this definition cannot be satisfactory. On the other hand, the concepts of canonical realization introduced in this report, by their very nature, guarantee uniqueness up to isomorphism of canonical realizations.

Example 5.8. Yamamoto (1978) has recently developed a theory of realization for continuous-time linear systems. He defines an m-port input signal to be an element of $(\varinjlim L^2[-n, 0])^m$, and a p-port output signal to be an element of $(\varprojlim L^2[0, n])^p$. Here, $L^2[a, b]$ is the Hilbert space of all square integrable functions on the interval $[a, b]$, and $\varinjlim$ (resp. $\varprojlim$) denotes inductive (resp. projective) limit. A behavior in his framework is a continuous linear map $b: (\varinjlim L^2[-n, 0])^m \to (\varprojlim L^2[0, n])^p$ which commutes with the natural shift operators on these spaces, and which is subject to the additional technical constraint that continuous functions be mapped to continuous functions in an appropriately smooth fashion.

Because the input and output signals are not parts of free or cofree constructions, Yamamoto's approach does not always yield a differential equation description such as (2) of § 1 as internal dynamics. Rather, a function of the form States × Inputs over $[0, t] \to$ States must serve as the fundamental definition of internal dynamics, with the existence of a differential equation describing the dynamics only occurring under certain conditions.

On the one hand, since certain of the systems realizable within Yamamoto's framework do not have differentiable dynamics, and on the other hand, since all of his systems are finite port, his approach overlaps somewhat that presented here, but neither subsumes the other.

In constructing a canonical realization, Yamamoto correctly observes that there need not be a unique (dense maps, injections) realization of a behavior. He then argues that the state should be continuous determinable from the output, or equivalently, that the observability map should be a closed embedding. Thus, he advocates the use of (dense maps, closed embeddings) as the correct factorization to yield a canonical realization, which does exist uniquely up to isomorphism. In the work presented here, (dense maps, closed embeddings) is an image-factorization system for the categories used, and so yields a concept of canonical realization. Thus, Yamamoto's concept of canonical realization is a particular member of the family of concepts of canonical realization presented in this report.

In closing, it is emphasized that while others have studied the problem of going from behavior to internal dynamics of continuous-time linear systems, this paper presents for the first time a *canonical* way to go from internal dynamics to behavior. The concepts of reachability and observability depend fundamentally upon the way in which the behavior is defined, and it is only after a satisfactory definition of canonical behavior is found that a satisfactory definition of canonical realization may be found.

It is also possible to develop a general duality theory of differentiable decomposable systems. However, it is necessary to introduce substantially more category theory to do so. Consult Hegner (1978a) for details.

Acknowledgment. The author wishes to thank M. A. Arbib, E. G. Manes and T. A. Cook for many helpful discussions during the course of this work.

REFERENCES

M. A. ARBIB (1973), *Coproducts and decomposable machines*, J. Comput. System Sci., 7, pp. 278–287.

M. A. ARBIB AND E. G. MANES (1974), *Foundations of system theory: decomposable systems*, Automatica-J. IFAC, 10, pp. 285–302.

——— (1975), *Arrows, Structures, and Functors: The Categorical Imperative*, Academic Press, New York.

J. S. BARAS AND R. W. BROCKETT (1975), *H^2-functions and infinite-dimensional realization theory*, SIAM J. Control, 13, pp. 221–241.

J. S. BARAS, R. W. BROCKETT AND P. A. FUHRMANN (1974), *State-space models for infinite-dimensional systems*, IEEE Trans. Automat. Control, 19, pp. 693–700.

J. S. BARAS AND P. DEWILDE (1976), *Invariant subspace methods in linear multivariable distributed systems and lumped-distributed network synthesis*, Proc. IEEE, 64, pp. 160–178.

A. BENSOUSSAN, M. C. DELFOUR AND S. K. MITTER (1975), *Representation and qualitative properties of infinite-dimensional linear systems*, part I, Report ESL-P-602, Electronic Systems Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

N. BOURBAKI (1966), *Elements of Mathematics, General Topology* (2 vol.), Addison-Wesley, Reading, MA.

J. DIEUDONNÉ (1960), *Foundations of Modern Analysis*, Academic Press, New York.

A. GROTHENDIECK (1955), *Produits Tensoriels Topologiques et Espaces Nucléaires*, American Mathematical Society, Providence, RI.

——— (1973), *Topological Vector Spaces*, Gordon and Breach, New York.

S. J. HEGNER (1978a), *Structured action approach to continuous-time systems*, Proc. 1978 Conf. on Information Sciences and Systems, The Johns Hopkins University, March 29–31, pp. 152–157.

——— (1978b), *Duality theory for discrete-time linear systems*, J. Comput. System Sci., 17, pp. 116–143.

——— (1980a), *An algebraic approach to continuous-time linear systems defined over Banach spaces*, Proc. 14th Annual Conf. on Information Sciences and Systems, Princeton University, 26–28 March, 1980.

——— (1980b), *Categories of (C_0) semigroups and realization theory*, to appear.

——— (1980c), *Linear systems over rings with approximate identity*, to appear.

H. HERRLICH AND G. E. STRECKER (1973), *Category Theory*, Allyn and Bacon, Boston, MA.

R. E. KALMAN AND M. L. J. HAUTUS (1972), *Realization of continuous-time linear dynamical systems: rigorous theory in the style of Schwartz*, Ordinary Differential Equations, 1971 NRL-MRC Conference, L. Weiss, ed., Academic Press, New York, pp. 151–164.

E. W. KAMEN (1971), *A distribution module-theoretic representation of linear dynamical continuous-time systems*, Information Systems Laboratory, Stanford Electronics Laboratories, Technical Report No. 6560–24, Stanford, CA.

G. KÖTHE (1969), *Topological Vector Spaces I*, Springer, New York.

L. PADULO AND M. A. ARBIB (1974), *System Theory: A Unified State-Space Approach to Continuous and Discrete Systems*, Hemisphere Publishers, Washington, DC.

W. RUDIN (1973), *Functional Analysis*, McGraw-Hill, New York.

H. H. SCHAEFER (1971), *Topological Vector Spaces*, Springer, New York.

H. SCHUBERT (1972), *Categories*, Springer, New York.

L. SCHWARTZ (1954–55), *Espaces de fonctions différentiables à valeurs vectorielles*, J. Analyse Math., 4, pp. 88–148.

——— (1957, 1959), *Théorie des distributions à valeurs vectorielles*, Ann. Inst. Fourier (Grenoble), 7, pp. 1–141; 8, pp. 1–209.

——— (1966), *Théorie des Distributions*, 3^{e} edition, Hermann, Paris.

F. TREVES (1967), *Topological Vector Spaces, Distributions, and Kernels*, Academic Press, New York.

L. WAELBROECK (1964), *Les semi-groupes différentiables*, Deuxieme Colloque sur l'Analyse Fonctionelle, Centre Belge de Recherche Mathématique, pp. 97–103.

Y. YAMAMOTO (1978), *Realization Theory of Infinite-Dimensional Linear Systems*, Dissertation, U. of Florida, Gainesville FL.

K. YOSIDA (1971), *Functional Analysis*, 3rd edition, Springer, New York.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0012 $01.00/0

DESCRIBING THE BEHAVIOR OF EIGENVECTORS OF RANDOM MATRICES USING SEQUENCES OF MEASURES ON ORTHOGONAL GROUPS*

JACK W. SILVERSTEIN†

Abstract. A conjecture has previously been made on the chaotic behavior of the eigenvectors of a class of n-dimensional random matrices, where n is very large [J. Silverstein, SIAM J. Appl. Math., 37 (1979), pp. 235–245]. Evidence supporting the conjecture has been given in the form of two limit theorems, as $n \to \infty$, relating the random matrices to matrices formed from the Haar measure, h_n, on the orthogonal group $\mathcal{O}_n$.

The present paper considers a reformulation of the conjecture in terms of sequences of the form $\{\mu_n\}$, where for each n, μ_n is a Borel probability measure on $\mathcal{O}_n$. A characterization of μ_n being "close" to h_n for n large is developed. It is suggested that before a definition of what it means for $\{\mu_n\}$ to be *asymptotic Haar* is decided, properties $\{h_n\}$ possess should first be proposed as possible necessary conditions. The limit theorems are converted into properties on $\{\mu_n\}$. It is shown (Theorem 1) that one property is a consequence of the other. Another property is proposed resulting in the construction of measures on $D = D[0, 1]$ which converge weakly. It is shown (Theorem 2) that under this necessary condition for asymptotic Haar, not only is the conjecture in general not true, but that the behavior of the eigenvectors of large dimensional sample covariance matrices deviates significantly from being Haar distributed when the i.i.d. standardized components making up the matrix differ in the fourth moment from 3.

1. Toward a definition of asymptotic Haar. In [6], a class of large dimensional, symmetric, positive semidefinite random matrices resulted from a model for the generation of neural connections of a hypothetical organism at birth. Denote by W_n one of these random matrices which is $n \times n$, where n is very large. Briefly, W_n is of the form $(1/C_n)V_nV_n^T$, where $V_n = (v_{ij})$ is $n \times dn$ and d is fixed; the v_{ij}'s are independent; v_{ij} is 1 or -1 with equal probability, or zero; $P = (P_{ij})$ is $n \times dn$, where $P_{ij} = \text{Prob}\,(v_{ij}^2 = 1)$, is formed under rather general conditions, and, in particular, every row of P is a rotation of the first row; and C_n is the sum of the first row of P. It is shown in [6] that if $C_n \to \infty$ as $n \to \infty$, then the empirical distribution function $F_n(x)$ of the eigenvalues of W_n converges in probability as $n \to \infty$ for each x to a fixed continuous distribution function $F(x)$. This result complements those on large dimensional random matrices (see for example [2], [3], [5], [7], [9], [11], [12], [13]), in particular, results on sample covariance matrices and matrices associated with the statistical theory of spectra.

In [10], a question is raised as to the behavior of the eigenvectors of W_n. It has been conjectured that this behavior is completely chaotic, and an attempt at formalizing this conjecture has been the following: for each n let $\mathcal{O}_n$ denote the orthogonal group consisting of $n \times n$ orthogonal matrices, and let $O_n \in \mathcal{O}_n$ be distributed according to the normalized Haar measure, h_n, on $\mathcal{O}_n$. Let D_n be a nonrandom $n \times n$ diagonal matrix with diagonal elements arranged in nondecreasing order and such that the spectrum of D_n approaches F as $n \to \infty$. The conjecture is that, for n large, the distribution of $W_n' \equiv O_nD_nO_n^T$ is close (in some sense) to the distribution of W_n.

Evidence supporting the conjecture is provided in [10] in the form of results which demonstrate that W_n and W_n' have similar properties. Let $\{P_a(M^n)\}_{a=0}^{\infty}$ be the spectral family of $M^n = W_n$ or W_n', let $\{x_n\}$, $x_n \in \mathbb{R}^n$, be any fixed sequence of unit vectors, and let M_1^n, M_2^n be two independent generations of M^n. Then it is proven in [10] that for $M^n = W_n$ or W_n',

$$x_n^T P_a(M^n) x_n \xrightarrow{\text{i.p.}} F(a) \quad \text{as } n \to \infty \text{ for every } a \in [0, \infty), \tag{1.1}$$

* Received by the editors May 28, 1980, and in revised form September 3, 1980.
† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

and

$$(1.2)\quad \frac{1}{n}\operatorname{tr}\left[(P_{a_1}(M_1^n)-P_{a_2}(M_2^n))^2\right] \xrightarrow{\text{i.p.}} F(a_1)+F(a_2)-2F(a_1)F(a_2) \quad \text{as } n\to\infty,$$

for every $a_1, a_2 \in [0, \infty)$ where tr is the trace function. The belief has been that these two results are enough to prove the conjecture.

The validity of the conjecture will imply certain properties of the neural model. However, the same question can be asked of other classes of large dimensional random matrices, at least for those not constructed from Gaussian variables. It is known, for example, that the Wishart matrix $W(I, dn)$ behaves like W_n' except D_n is random (see [1, Chapt. 13]). We remark here that the results in [10] are true for sample covariance matrices in which the elements in the sample vectors are i.i.d., mean 0, having moments of all orders (the results in [10] rely totally on [6, Lemma 1], and the proof of this lemma can be slightly modified to include these cases). It is also believed that these results are valid for more general random matrices. Thus, statements concerning W_n are relevant for a large class of random matrices.

The present paper continues the investigation of the eigenvectors of W_n primarily by developing some ideas toward a well-defined statement of the conjecture. To begin with, it seems more fitting to shift the attention from W_n to the measure it induces on $\mathcal{O}_n$. Let $O_n \in \mathcal{O}_n$ be random, defined on the same probability space as W_n, and such that $O_n^T W_n O_n = \Lambda_n$, where Λ_n is diagonal with its diagonal elements arranged in nondecreasing order. We may as well assume that the distribution of O_n is the same as that of $O_n J$ for each diagonal J containing ± 1's along its diagonal. Also we may assume that, conditioned on any collection of subsets of eigenvalues of W_n being equal within each subset, the distribution of O_n is the same as that of $O_n K$ whenever $K \in \mathcal{O}_n$ transforms only among each subset of columns of O_n corresponding to a subset of equal eigenvalues, and leaves all other columns unchanged. Let ν_n be the Borel probability measure induced by O_n.

The conjecture can now be expressed in terms of ν_n and h_n being "close" for n large. We will use the expression *asymptotic Haar* to describe this, at present a vague property on sequences $\{\mu_n\}$ where, for each n, μ_n is a Borel probability measure on $\mathcal{O}_n$.

The most obvious and by far the strongest statement of asymptotic Haar is: for every $\varepsilon > 0$, we have for all n sufficiently large $|\mu_n(A) - h_n(A)| < \varepsilon$, for every $A \in \mathbb{B}_n \equiv$ the collection of Borel sets of $\mathcal{O}_n$ (the metric on $\mathcal{O}_n$ being induced from the operator norm). This definition is too restrictive if we do not want to exclude from being asymptotic Haar all sequences $\{\mu_n\}$ of atomic measures. If we let $S_{n,\delta}$ represent the collection of all open balls on $\mathcal{O}_n$ having Haar measure δ, then another definition which would allow certain sequences of atomic measures is: for every $\varepsilon > 0$, $1 \geqq \delta > 0$, we have for all n sufficiently large $|\mu_n(B) - h_n(B)| < \varepsilon$ for every $B \in S_{n,\delta}$. Several alternative definitions can certainly be proposed along the same lines.

It is the author's view that, instead of initially focusing on one definition of asymptotic Haar, attention should be drawn on intuitive and reasonable consequences of the definition. Various properties $\{h_n\}$ possess should be considered as necessary conditions for asymptotic Haar. Also, examples of sequences that should not be asymptotic Haar need to be found. For example, (1.1) and (1.2) can be restated in terms of the following properties.

We say that $\{u_n\}$ satisfies property I if for any sequence of unit vectors $\{x_n\}$, $x_n \in \mathbb{R}^n$, any number b such that $0 \leqq b \leqq 1$, and any sequence of integers $\{m_n\}$ satisfying

$0 \leqq m_n \leqq n$ and $m_n/n \to b$ as $n \to \infty$, we have,

$$x_n^T O_n D(n, m_n) O_n^T x_n \xrightarrow{\text{i.p}} b \quad \text{as } n \to \infty, \tag{1.3}$$

where O_n is μ_n-distributed, and where $D(n, m_n)$ is $n \times n$ and has 0 for all its entries except for 1's in the first m_n diagonal entries.

We say that μ_n satisfies property II if for any b_1, b_2 such that $0 \leqq b_1, b_2 \leqq 1$, and any two sequences of integers $\{m_n^1\}$, $\{m_n^2\}$ satisfying $0 \leqq m_n^i \leqq n$ and $m_n^i/n \to b_i$ as $n \to \infty$, $i = 1, 2$, we have,

$$\frac{1}{n} \operatorname{tr}\left[(O_n D(n, m_n^1) O_n^T - O_n' D(n, m_n^2) O_n'^T)^2\right] \xrightarrow{\text{i.p.}} b_1 + b_2 - 2b_1 b_2 \quad \text{as } n \to \infty, \tag{1.4}$$

where O_n and O_n' are independent and μ_n-distributed.

The sequence $\{h_n\}$ satisfies I. The easiest way of seeing this is to use the fact that $x_n^T O_n D(n, m_n) O_n^T x_n$ is beta-distributed with mean m_n/n which goes to b, and variance $2m_n(n - m_n)/n^3$ which goes to 0 (see [10, proof of Theorem 1]).

The sequence $\{\nu_n\}$ also satisfies I. The proof is elementary and technical and will be omitted.

Theorem 1 in the next section shows that II is a consequence of I, a somewhat surprising result. Thus, we have so far only one necessary condition for asymptotic Haar.

At this stage, we are in a position to consider whether I is enough to characterize asymptotic Haar. For each n let μ_n be absolutely continuous with respect to h_n, having density f_n. Let $\{x_n\}$, $\{m_n\}$ be as in I. Using the fact that $x_n^T O_n D(n, m_n) O_n^T x_n$ is beta-distributed when O_n is h_n-distributed, we get from the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\left(\int_{\mathcal{O}_n} \left| x_n^T O_n D(n, m_n) O_n^T x_n - \frac{m_n}{n} \right| f_n(O_n)\, dh_n(O_n)\right)^2 \\
&\leqq \left[\int_{\mathcal{O}_n} \left(x_n^T O_n D(n, m_n) O_n^T x_n - \frac{m_n}{n}\right)^2 dh_n(O_n)\right]\left[\int_{\mathcal{O}_n} f_n^2(O_n)\, dh_n(O_n)\right] \\
&= 2\frac{m_n}{n^2}\left(\frac{n - m_n}{n}\right) \int_{\mathcal{O}_n} f_n^2(O_n)\, dh_n(O_n).
\end{aligned} \tag{1.5}$$

Thus, if $\int_{\mathcal{O}_n} f_n^2(O_n)\, dh_n(O_n) = o(n)$, then we get L^1-convergence in (1.3) so that $\{\mu_n\}$ satisfies I. This is true if $\{f_n\}$ is any uniformly bounded sequence of densities. For example, if $f_n = 2$ on a closed subset of $\mathcal{O}_n$ having Haar measure $\frac{1}{2}$, and 0 elsewhere, then $\{\mu_n\}$ satisfies I. Under the quite reasonable assumption that the above sequence should not be considered to be asymptotic Haar, then we must conclude that I is *not* enough to characterize asymptotic Haar.

Other properties of $\{h_n\}$ therefore, need to be considered.

The remainder of this paper is devoted to developing another property, and considering the consequences, if this property is to be a necessary condition for asymptotic Haar.

For $O_n \in \mathcal{O}_n$ Haar-distributed and any unit vector $x_n \in \mathbb{R}^n$, we have $O_n^T x_n$-distributed like $(\zeta_1, \zeta_2, \cdots, \zeta_n)/(\sum_{i=1}^n \zeta_i^2)^{1/2}$, where $\zeta_1, \zeta_2, \cdots, \zeta_n$ are i.i.d. $n(0, 1)$. Form

$$\begin{aligned}
X_n(t) &= \frac{\sqrt{n}}{\sqrt{2}}\left(\frac{\sum_{i=1}^{[nt]} \zeta_i^2}{\sum_{i=1}^n \zeta_i^2} - \frac{[nt]}{n}\right) \\
&= \frac{n}{\sum_{i=1}^n \zeta_i^2} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{[nt]} (\zeta_i^2 - 1) - \frac{[nt]}{n} \sum_{i=1}^{h} (\zeta_i^2 - 1)\right),
\end{aligned} \tag{1.6}$$

where $[s]$ is the greatest integer $\leqq s$. We have $X_n(t)$ a random element of $D = D[0, 1]$ (the space of all r.c.l.l. functions on $[0, 1]$) and from straightforward applications of Donsker's Theorem and the theory of weak convergence of measures [4], we have:

$$X_n \xrightarrow{\mathcal{D}} W^0, \tag{1.7}$$

where W^0 is a Brownian bridge. Hence, another necessary condition for asymptotic Haar:

We say that $\{\mu_n\}$ satisfies property III if, for every sequence $\{x_n\} x_n \in \mathbb{R}^n$ of unit vectors, if $(\zeta_1, \zeta_2, \cdots, \zeta_n) \equiv O_n^T x_n$ where O_n is μ_n-distributed, and if $X_n(t)$ is as in (1.6), then (1.7) holds.

This property seems to be a reasonable necessary condition for asymptotic Haar. It ensures that $O_n^T x_n$ be close to being uniformly distributed on the unit sphere in $\mathbb{R}^n$. In fact, $O_n^T x_n$ need only have a distribution resembling the distribution of $(Y_1, Y_2, \cdots, Y_n)/(\sum_{i=1}^n Y_i^2)^{1/2}$ where the Y_i's are i.i.d. with $E(Y_1^2) = 1$ and var $(Y_1^2) = 2$.

It would also seem reasonable that the behavior of the eigenvectors of large dimensional sample covariance matrices be a prime example for asymptotic Haar. But with the inclusion of III as a necessary condition, this is not the case. Let $\{u_{ij}\}$, $i, j = 1, 2, \cdots$, be i.i.d. random variables having mean 0, variance 1, and satisfying $E(|u_{11}|^m) \leqq m^{\alpha m}$ for all integers $m > 2$ and for some α. For each n let $U_n = (u_{ij})$, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, s$, where $n/s \to y > 0$ as $n \to \infty$, and let μ_n be the measure on $\mathcal{O}_n$ induced from $(1/s) U_n U_n^T$. Theorem 2 in the next section shows that if $E(u_{ij}^4) \neq 3$, then $\{\mu_n\}$ does not satisfy III. The proof relies on standard tools used in the theory of weak convergence on metric spaces, along with a recent result on the almost sure convergence of the largest eigenvalue of sequences of sample covariance matrices [5], where the above growth condition on the moments of $|u_{11}|$ is assumed.

In the formation of W_n, letting $P_{ij} = p$ for all i, j where $p \neq \frac{1}{3}$, we are in the above case with $E(u_{11}^4) = E((v_{11}/\sqrt{p})^4) = 1/p$. We must therefore conclude that with III as a necessary condition for asymptotic Haar, the original conjecture is, in general, false. It may be argued that III is too strong, and it may be possible to find interesting properties shared by $\{\nu_n\}$ and $\{h_n\}$. Moreover, $\{\nu_n\}$ may still satisfy III when $p = \frac{1}{3}$ or when the P_{ij}'s are not all the same. However, we feel that failure to satisfy III indicates significant departure from Haar measure.

The requirement that $E(u_{11}^4) = 3$ suggests that for sample covariance matrices, in order to satisfy III, the u_{ij}'s have to be near to being Gaussian distributed, as in the Wishart case. It appears worthwhile to determine what conditions on the u_{ij}'s are needed to ensure III.

In conclusion, it should be emphasized that one purpose of this paper is to begin an investigation on how to characterize the closeness of measures on $\mathcal{O}_n$ to Haar measure, where n is large. The considerations given are clearly the author's view on how to proceed in defining asymptotic Haar. We suggest continuing the characterization by finding other mappings of $\mathcal{O}_n$ onto a common metric space S, resulting in weak convergence of the measures on S induced by $\{h_n\}$. Intuitively, the mappings $F_n: \mathcal{O}_n \to S$ should all be similar, sort of invariant across dimensions. They should also illuminate the intrinsic uniformity of Haar measure.

We find it interesting that the W_n's do not in general fall into the present characterization of asymptotic Haar. Still, $\{\nu_n\}$ and $\{h_n\}$ are similar, and a first step

toward determining just how similar they are would be to understand those sequences $\{\mu_n\}$ satisfying I.

The fact that sequences $\{\mu_n\}$ arising from sample covariance matrices do not in general satisfy III is of even greater interest, and this suggests a behavior of the eigenvectors of these matrices for large n which runs counter to our intuition. A description of this behavior is important to multivariate theory, and work in this area should be pursued.

2. The theorems.

THEOREM 1. I $\rightarrow$ II.

Proof. Assume $\{\mu_n\}$ satisfies I. Let $P(m_n, O_n) \equiv O_n D(n, m_n) O_n^T$. Convergence of $x_n^T P(m_n, O_n) x_n$ to b in probability is equivalent to

$$\text{(2.1)} \qquad \mathrm{E}(x_n^T P(m_n, O_n) x_n) = \int_{\mathcal{O}_n} x_n^T P(m_n, O_n) x_n \, d\mu_n(O_n) \to b \quad \text{as } n \to \infty$$

and

$$\text{(2.2)} \qquad \mathrm{E}((x_n^T P(m_n, O_n) x_n)^2) = \int_{\mathcal{O}_n} (x_n^T P(m_n, O_n) x_n)^2 \, d\mu_n(O_n) \to b^2 \quad \text{as } n \to \infty.$$

The expected values in (2.1) and (2.2) are polynomials in the components of x_n and are therefore continuous in x_n. Let $\{x_n'\}$ and $\{x_n''\}$ be the sequences such that $\mathrm{E}(x_n^T P(m_n, O_n) x_n)$ attains its maximum at x_n' and its minimum at x_n''. Since (2.1) holds for *all* sequences of unit vectors, it is certainly true for $\{x_n'\}$ and $\{x_n''\}$. Therefore,

$$\text{(2.3)} \quad \mathrm{E}(x_n^T P(m_n, O_n) x_n) = b + \alpha_n(x_n) \quad \text{where } |\alpha_n(x_n)| \leqq \alpha_n \text{ and } \alpha_n \to 0 \text{ as } n \to \infty.$$

Similarly,

$$\text{(2.4)} \quad \mathrm{E}((x_n^T P(m_n, O_n) x_n)^2) = b^2 + \beta_n(x_n) \quad \text{where } |\beta_n(x_n)| \leqq \beta_n \text{ and } \beta_n \to 0 \text{ as } n \to \infty.$$

Also, for any two sequences $\{x_n\}, \{y_n\}$ we have

$$\text{(2.5)} \qquad (x_n^T P(m_n, O_n) x_n)(y_n^T P(m_n, O_n) y_n) \xrightarrow{\text{i.p.}} b^2 \quad \text{as } n \to \infty,$$

and as above we have

$$\text{(2.6)} \qquad \mathrm{E}((x_n^T P(m_n, O_n) x_n)(y_n^T P(m_n, O_n) y_n)) = b^2 + \gamma_n(x_n, y_n),$$

where $|\gamma_n(x_n, y_n)| \leqq \gamma_n$ and $\gamma_n \to 0$ as $n \to \infty$. Let $\{m_n^1\}, \{m_n^2\}$ be as in II. Since

$$\text{(2.7)} \qquad \begin{aligned} \frac{1}{n} \operatorname{tr} [(P(m_n^1, O_n) - P(m_n^2, O_n'))^2] = \frac{m_n^1}{n} + \frac{m_n^2}{n} - \frac{1}{n} \operatorname{tr} P(m_n^1, O_n) P(m_n^2, O_n') \\ - \frac{1}{n} \operatorname{tr} P(m_n^2, O_n') P(m_n^1, O_n), \end{aligned}$$

it is sufficient to prove

$$\text{(2.8)} \qquad \frac{1}{n} \operatorname{tr} P(m_n^1, O_n) P(m_n^2, O_n') \xrightarrow{\text{i.p.}} b_1 b_2 \quad \text{as } n \to \infty.$$

We have

$$\begin{aligned}\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n') &= \frac{1}{n}\operatorname{tr} P(m_n^1, O_n)O_n'D(n, m_n^2)O_n'^T \\ &= \frac{1}{n}\operatorname{tr} O_n'^T P(m_n^1, O_n)O_n'D(n, m_n^2) \\ &= \frac{1}{n}\sum_{i=1}^{m_n^2}(O_n'^T P(m_n^1, O_n)O_n')_{ii} \\ &= \frac{1}{n}\sum_{i=1}^{m_n^2} o_{\cdot i}'^T P(m_n^1, O_n)o_{\cdot i}',\end{aligned} \tag{2.9}$$

where $o_{\cdot i}'$ is the ith column of O_n'. For fixed O_n' we have, from (2.3),

$$\int_{\mathcal{O}_n}\frac{1}{n}\sum_{i=1}^{m_n^2} o_{\cdot i}'^T P(m_n^1, O_n)o_{\cdot i}'\,d\mu_n(O_n) = \frac{m_n^2}{n}b_1 + \frac{1}{n}\sum_{i=1}^{m_n^2}\alpha_n(o_{\cdot i}') \tag{2.10}$$

and

$$\frac{1}{n}\left|\sum_{i=1}^{m_n^2}\alpha_n(o_{\cdot i}')\right| \leqq \alpha_n.$$

Therefore,

$$\mathrm{E}\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right) = \left(\frac{m_n^2}{n}\right)b_1 + \xi_n, \tag{2.11}$$

where $\xi_n \to 0$ as $n \to \infty$, and so

$$\mathrm{E}\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right) \to b_1 b_2 \quad \text{as } n \to \infty. \tag{2.12}$$

We have

$$\begin{aligned}\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right)^2 &= \left(\frac{1}{n}\right)^2\left(\sum_{i=1}^{m_n^2} o_{\cdot i}'^T P(m_n^1, O_n)o_{\cdot i}'\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{m_n^2}(o_{\cdot i}'^T P(m_n^1, O_n)o_{\cdot i}')^2 \\ &\quad + \frac{1}{n^2}\sum_{i_1 \neq i_2}^{m_n^2}(o_{\cdot i_1}'^T P(m_n^1, O_n)o_{\cdot i_1}')(o_{\cdot i_2}'P(m_n^1, O_n)o_{\cdot i_2}').\end{aligned} \tag{2.13}$$

For fixed O_n' we have

$$\begin{aligned}&\int_{\mathcal{O}_n}\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right)^2 d\mu_n(O_n) \\ &\quad = \left(\frac{m_n^2}{n}\right)^2 b_1^2 + \frac{1}{n^2}\sum_{i=1}^{m_n^2}\beta_n(o_{\cdot i}') + \frac{1}{n^2}\sum_{i_1 \neq i_2}^{m_n^2}\gamma_n(o_{\cdot i_1}', o_{\cdot i_2}'),\end{aligned} \tag{2.14}$$

from (2.4) and (2.6). The absolute value of the sum of the last two terms is bounded by $\beta_n + \gamma_n$. Therefore,

$$\mathrm{E}\left(\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right)^2\right) = \left(\frac{m_n^2}{n}\right)^2 b_1^2 + \eta_n, \tag{2.15}$$

where $\eta_n \to 0$ as $n \to \infty$, so that

$$\mathrm{E}\left(\left(\frac{1}{n}\operatorname{tr} P(m_n^1, O_n)P(m_n^2, O_n')\right)^2\right) \to (b_1 b_2)^2 \quad \text{as } n \to \infty. \tag{2.16}$$

From (2.12) and (2.16) we get (2.8) and we are done.

THEOREM 2. Let $\{u_{ij}\} i, j = 1, 2, \cdots$, be i.i.d. random variables having mean 0, variance 1, and satisfying $\mathrm{E}(|u_{11}|^m) < m^{\alpha m}$ for all integers $m > 2$, and for some α. For each n let $U_n = (u_{ij})$, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, s$, where $(n/s) \to y > 0$ as $n \to \infty$, and let μ_n be the measure on $\mathcal{O}_n$ induced from $M_n \equiv (1/s)U_n U_n^T$.

If $\{\mu_n\}$ satisfies III, then $\mathrm{E}(u_{11}^4) = 3$.

Proof. Let $F_n(a)$, $a \in [0, \infty)$ be the empirical distribution function of the eigenvalues of M_n. Let $F_y(a)$ be the limiting distribution function which is given in Theorem 2.1 of [7]. Since $F_y(a)$ is continuous for $a \in [0, \infty)$ we can conclude from Theorem 3.2 of [7] that

$$\sup_{a \in [0,\infty)} |F_n(a) - F_y(a)| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty. \tag{2.17}$$

The functions $F_n(a)$ and $F_y(a)$ are elements of $D_0 = D_0[0, \infty) = \{x \in D[0, \infty)$: $\lim_{t \to \infty} x(t)$ exists and is finite$\}$ [8]. From (2.17), it follows that

$$F_n(a) \xrightarrow{\text{a.s.}} F_y(a) \quad \text{as } n \to \infty \text{ in } D_0. \tag{2.18}$$

Assume III and let $\{x_n\}$ be given. For our purpose $X_n(t)$ of (1.6) can be constructed directly from M_n. In fact, we have

$$X_n(F_n(a)) = \frac{\sqrt{n}}{\sqrt{2}}(x_n^T P_a(M_n) x_n - F_n(a)), \tag{2.19}$$

where $\{P_a(M_a)\}$ is the spectral family of M_n. A simple extension of the material in [4, pp. 144–145] to nondecreasing functions in $D_0[0, \infty)$ and [4, Theorem 4.4] leads us to conclude that III and (2.18) imply

$$X_n(F_n(a)) \xrightarrow{\mathcal{D}} W^0_{F_y(a)} \equiv W^y_a \quad \text{in } D_0. \tag{2.20}$$

For every positive integer r, we have

$$\frac{\sqrt{n}}{\sqrt{2}}\left(x_n^T M_n^r x_n - \frac{1}{n}\operatorname{tr} M_n^r\right) = \int_0^\infty a^r \, dX_n(F_n(a)) = -\int_0^\infty r a^{r-1} X_n(F_n(a)) \, da, \tag{2.21}$$

where we have used the fact that with probability 1, $X_n(F_n(a))$ vanishes outside a bounded set.

For any $b > 0$, the mapping that takes $x \in D_0$ to $\int_0^b r a^{r-1} x(a) \, da$ is continuous. Therefore, from [4, Theorem 5.1],

$$\int_0^b r a^{r-1} X_n(F_n(a)) \, da \xrightarrow{\mathcal{D}} \int_0^b r a^{r-1} W_a^y \, da. \tag{2.22}$$

With the growth condition on $\mathrm{E}(|u_{11}|^m)$ we have from [5] that the maximum eigenvalue of M_n converges almost surely to $(1 + \sqrt{y})^2$. Therefore, when $b > (1 + \sqrt{y})^2$ we have

$$\int_0^\infty r a^{r-1} X_n(F_n(a)) \, da - \int_0^b r a^{r-1} X_n(F_n(a)) \, da \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty. \tag{2.23}$$

Therefore,

$$(2.24)\qquad \frac{\sqrt{n}}{\sqrt{2}}\left(x_n^T M_n^r x_n - \frac{1}{n}\operatorname{tr} M_n^r\right) \xrightarrow{\mathcal{D}} -\int_0^b r a^{r-1} W_a^y\, da = -\int_0^{(1+\sqrt{y})^2} r a^{r-1} W_a^y\, da.$$

The limiting distribution is thus Gaussian, with mean and variance only depending on W_a^y.

Let $r = 1$. We have

$$(2.25)\qquad \sqrt{n}\left(\frac{1}{n}\operatorname{tr} M_n - 1\right) = \frac{1}{\sqrt{n}s}\sum_{i,j}(u_{ij}^2 - 1),$$

which has mean 0 and variance $(1/s)(\mathrm{E}(u_{11}^4) - 1) \to 0$ as $n \to \infty$. Therefore, we need only consider $(\sqrt{n}/\sqrt{2})(x_n^T M_n x_n - 1)$. Let $x_n = (1, 0, \cdots, 0)$. Then

$$(2.26)\qquad \frac{\sqrt{n}}{\sqrt{2}}(x_n^T M_n x_n - 1) = \frac{\sqrt{n}}{\sqrt{2}}\left(\frac{1}{s}\sum_j u_{1j}^2 - 1\right) = \frac{\sqrt{n}}{\sqrt{s}}\frac{1}{\sqrt{2}\sqrt{s}}\sum_j (u_{1j}^2 - 1),$$

which from the Central Limit Theorem converges in distribution to $n(0, (y/2)(\mathrm{E}(u_{11}^4) - 1))$. Therefore, III depends on the value of $\mathrm{E}(u_{11}^4)$ which must be 3, because in the Wishart case, u_{11} is $n(0, 1)$.

We remark that from preliminary work, it is believed that $\mathrm{E}(u_{11}^4) = 3$ is enough to ensure (2.24) for all $r \geqq 1$.

REFERENCES

[1] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, 1958.

[2] L. V. ARHAROV, *Limit theorems for the characteristic roots of a sample covariance matrix*, Soviet Math. Dokl., 12 (1971), pp. 1206–1209.

[3] L. ARNOLD, *On Wigner's semicircle law for the eigenvalues of random matrices*, Z. Wahrsch. Verw. Gebiete, 19 (1971), pp. 191–198.

[4] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[5] S. GEMAN, *A limit theorem for the norm of random matrices*, Ann. Probab., 8 (1980), pp. 252–261.

[6] U. GRENANDER AND J. W. SILVERSTEIN, *Spectral analysis of networks with random topologies*, SIAM J. Appl. Math., 32 (1977), pp. 499–519.

[7] D. JONSSON, *Some limit theorems for the eigenvalues of a sample covariance matrix*, Uppsala University, Department of Mathematics, Report No. 6, Uppsala, Sweden.

[8] T. LINDVALL, *Weak convergence of probability measures and random functions in the function space $D[0, \infty)$*, J. Appl. Probab., 10 (1973), pp. 109–121.

[9] V. A. MARCENKO AND L. A. PASTUR, *Distribution of eigenvalues for some sets of random matrices*, Math. USSR-Sb., 1 (1967), pp. 457–483.

[10] J. W. SILVERSTEIN, *On the randomness of eigenvectors generated from networks with random topologies*, SIAM J. Appl. Math., 37 (1979), pp. 235–245.

[11] K. W. WACHTER, *The strong limits of random matrix spectra for sample matrices of independent elements*, Ann. Probab., 6 (1978), pp. 1–18.

[12] E. P. WIGNER, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. of Math., 62 (1955), pp. 548–564.

[13] ———, *On the distribution of the roots of certain symmetric matrices*, Ann. of Math., 67 (1958), pp. 325–327.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, March 1981

0036-1410/81/1202-0013 $01.00/0

BOUNDS AND MAXIMUM PRINCIPLES FOR THE SOLUTION OF THE LINEAR TRANSPORT EQUATION*

EDWARD W. LARSEN†

Abstract. Pointwise bounds are derived for the solution of time-independent linear transport problems with surface sources in convex spatial domains. Under specified conditions, upper bounds are derived which, as a function of position, decrease with distance from the boundary. Under other conditions, lower bounds are derived which increase with distance from the boundary. Also, sufficient conditions are obtained for the existence of maximum and minimum principles, and a counterexample is given which shows that such principles do not always exist.

1. Introduction. The purpose of this article is to derive pointwise bounds on the solution of time-independent linear transport problems for convex subcritical spatial domains D with nonnegative surface sources and no interior sources. Our results have potential applications for radiative transfer, neutron transport and other linear neutral-particle transport processes.

In general, the bounds we derive are space-dependent. For certain problems, we obtain upper bounds which are decreasing functions of distance from the boundary, and for other problems we obtain lower bounds which are increasing functions of distance from the boundary. We begin by treating forward and adjoint energy-dependent transport problems, and then as a special case we consider one-speed problems. Also, if the coefficients of the transport equation satisfy a certain condition, we show that a maximum (or minimum) principle holds: the solution $\psi(\mathbf{x}, \mathbf{\Omega}, E)$ of the forward transport equation is bounded from above (below) by the supremum (infimum) of its values over $\mathbf{x} \in \partial D$, $\mathbf{\Omega}$ pointing into D, and all E. For the adjoint equation, the sup (inf) is taken over $\mathbf{x} \in \partial D$, $\mathbf{\Omega}$ pointing out of D, and all E. To conclude this article, we present an example which shows that a maximum principle does not always exist.

The derivation of pointwise bounds and maximum principles for the solution of linear transport equations has been considered previously by other authors [1–3]. In [1], Case and Zweifel derive a pointwise bound in order to prove the convergence of a Neumann series. We show later that this bound is considerably different from the bounds derived here. In [2], Bensoussan, Lions and Papanicolaou, and in [3], Williams, derive a maximum principle for the solution ψ^* of the adjoint time-dependent transport equation in a "locally conservative" medium (one for which the collision process conserves particles at each spatial point). This maximum principle has the following form: for $0 \leqq t \leqq T$, $\psi^*(\mathbf{x}, \mathbf{\Omega}, E, t)$ is bounded from above by the larger of the following two quantities: (1) the supremum of ψ^*, over all $\mathbf{x}$, $\mathbf{\Omega}$, and E, of its values for $t = 0$, and (2) the supremum of ψ^* over $\mathbf{x} \in \partial D$, all $\mathbf{\Omega}$, all E, and $0 \leqq t \leqq T$. Although the analyses in [2] and [3] are inherently time-dependent, the maximum principle for the analogous time-independent transport equation, which is derived here, is implicitly contained in these results when combined with some analysis in [4]. (This was pointed out to the author by M. Williams.)

In this article, we consider forward and adjoint energy-dependent transport problems in § 2, and we specialize our results to one-group transport problems in § 3. Then in § 4 we discuss a subcritical transport problem for which a maximum principle does not exist.

* Received by the editors January 4, 1980 and in revised form September 8, 1980. This work was performed under the auspices of the U.S. Department of Energy.

† Theoretical Division, University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545.

2. The energy-dependent transport equation. Let D denote an open, convex spatial domain with boundary ∂D possessing a piecewise-continuously varying unit outer normal $\mathbf{n}$, let S denote the unit sphere, and $\psi(\mathbf{x}, \mathbf{\Omega}, E)$ denote the flux of particles at the point $\mathbf{x} \in D$, travelling in the direction $\mathbf{\Omega} \in S$, with energy $E \in B = [E_0, E_1]$. To simplify notation, we introduce the following phase spaces:

$$P = \{(\mathbf{x}, \mathbf{\Omega}, E) | \mathbf{x} \in D, \mathbf{\Omega} \in S, E \in B\},$$

$$\Gamma^- = \{(\mathbf{x}, \mathbf{\Omega}, E) | \mathbf{x} \in D, \mathbf{\Omega} \cdot \mathbf{n} < 0, E \in B\},$$

$$R = \{(\mathbf{x}, E) | \mathbf{x} \in D, E \in B\}.$$

The transport problem which ψ satisfies is then [5], [6]

$$\mathbf{\Omega} \cdot \nabla_x \psi + \Sigma_T \psi = \Sigma_S \psi, \qquad (\mathbf{x}, \mathbf{\Omega}, E) \in P, \tag{2.1}$$

$$\psi = f, \qquad (\mathbf{x}, \mathbf{\Omega}, E) \in \Gamma^-, \tag{2.2}$$

where the operators Σ_T and Σ_S are defined by

$$\Sigma_T \psi(\mathbf{x}, \mathbf{\Omega}, E) = \sigma_T(\mathbf{x}, E) \psi(\mathbf{x}, \mathbf{\Omega}, E), \tag{2.3}$$

$$\Sigma_S \psi(\mathbf{x}, \mathbf{\Omega}, E) = \int \sigma_S(\mathbf{x}, \mathbf{\Omega} \cdot \mathbf{\Omega}', E' \to E) \psi(\mathbf{x}, \mathbf{\Omega}', E') \, d\mathbf{\Omega}' \, dE'. \tag{2.4}$$

(For neutron transport problems, the operator Σ_S accounts for all scatterings; elastic, inelastic, and fission.) We require f, σ_T, and σ_S to be nonnegative measurable functions satisfying

$$\sup_{\Gamma^-} f(\mathbf{x}, \mathbf{\Omega}, E) < \infty$$

and

$$\int \sigma_S(\mathbf{x}, \mathbf{\Omega} \cdot \mathbf{\Omega}', E' \to E) \, d\Omega' \, dE' \leqq c \sigma_T(\mathbf{x}, E), \qquad (\mathbf{x}, E) \in R, \tag{2.5}$$

where c is a nonnegative constant. We seek ψ in the Banach space X of bounded, measurable functions $h(\mathbf{x}, \mathbf{\Omega}, E)$, defined for $(\mathbf{x}, \mathbf{\Omega}, E) \in P \cup \Gamma^-$, and satisfying

$$\|h\| = \sup_{P \cup \Gamma^-} |h(\mathbf{x}, \mathbf{\Omega}, E)| < \infty.$$

By integrating along a characteristic curve, we may invert the operator on the left side of (2.1) and obtain an integral equation for ψ:

$$(I - L)\psi = Q. \tag{2.6}$$

Here

$$L\psi(\mathbf{x}, \mathbf{\Omega}, E) = \int_0^{d(\mathbf{x}, \mathbf{\Omega})} \exp\left(-\int_0^t \sigma_T(\mathbf{x} - s\Omega, E) \, ds\right) (\Sigma_S \psi)(\mathbf{x} - t\mathbf{\Omega}, \mathbf{\Omega}, E) \, dt, \tag{2.7}$$

$$Q(\mathbf{x}, \mathbf{\Omega}, E) = f[\mathbf{x} - d(\mathbf{x}, \mathbf{\Omega})\mathbf{\Omega}, \mathbf{\Omega}, E] \exp\left(-\int_0^{d(\mathbf{x}, \mathbf{\Omega})} \sigma_T(\mathbf{x} - s\mathbf{\Omega}, E) \, ds, \right. \tag{2.8}$$

and $d(\mathbf{x}, \mathbf{\Omega})$ denotes the distance from $\mathbf{x} \in D$ to ∂D in the direction of $-\mathbf{\Omega}$ (see Fig. 1).

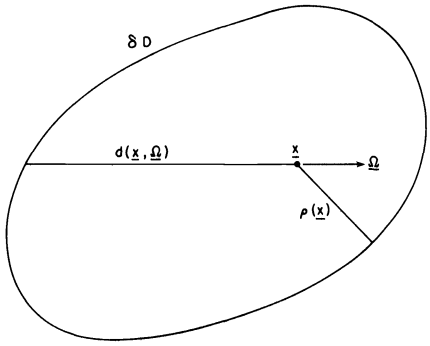


FIG. 1. *The functions* $d(\mathbf{x}, \mathbf{\Omega})$ *and* $\rho(\mathbf{x})$.

If τ denotes the maximum optical thickness of D, i.e.,

$$\tau = \sup_P \int_0^{d(\mathbf{x},\mathbf{\Omega})} \sigma_T(\mathbf{x} - s\mathbf{\Omega}, E)\, ds,$$

then for any $h \in X$, (2.4), (2.5), and (2.7) give

$$|Lh(\mathbf{x}, \mathbf{\Omega}, E)| \leqq \int_0^{d(\mathbf{x},\mathbf{\Omega})} \exp\left(-\int_0^t \sigma_T(\mathbf{x} - s\mathbf{\Omega}, E)\, ds\right) c\sigma_T(\mathbf{x} - t\mathbf{\Omega}, E)\|h\|\, dt$$

$$= c\left[1 - \exp\left(-\int_0^{d(\mathbf{x},\mathbf{\Omega})} \sigma_T(\mathbf{x} - s\mathbf{\Omega}, E)\, ss\right]\|h\|,$$

and so,

$$\|L\| \leqq c(1 - e^{-\tau}). \tag{2.9}$$

Thus, L is bounded, and if spr (L), the spectral radius of L, is less than 1, (2.6) can be solved by means of a Neumann series,

$$\psi = \sum_{n=0}^{\infty} L^n Q. \tag{2.10}$$

The nth term of this series has the physical interpretation of being the flux of particles which have undergone exactly n collisions.

Throughout this article, we shall assume spr $(L) < 1$, which means physically that the transport problem (2.1), (2.2) is "subcritical" [6]. The case spr $(L) = 1$ [or spr $(L) > 1$] corresponds physically to a "critical" [or "supercritical"] transport problem [6], which does not have a physically meaningful solution for $f > 0$ even if a mathematical solution exists.

Since f is nonnegative and bounded, Q is nonnegative and an element of X. Then since L is a positive operator, (2.10) implies $\psi \geqq 0$. Thus, we have:

THEOREM 1. *There exists a unique, nonnegative solution* $\psi \in X$ *of problem* (2.1), (2.2).

THEOREM 2. *For any* $h \in X$, $h \geqq 0$, *there exists a unique, nonnegative solution* $\phi \in X$ *of the integral equation* $(I - L)\phi = h$.

To proceed, let us define the function $\rho(\mathbf{x})$, for $\mathbf{x} \in D$, as the shortest distance from $\mathbf{x}$ to ∂D. Equivalently,

$$\rho(\mathbf{x}) = \inf_{\mathbf{\Omega} \in S} d(\mathbf{x}, \mathbf{\Omega}), \tag{2.11}$$

which implies

$$\rho(\mathbf{x}) \leqq d(\mathbf{x}, \boldsymbol{\Omega}), \qquad \mathbf{x} \in D, \quad \boldsymbol{\Omega} \in S. \tag{2.12}$$

Also,

$$\rho(\mathbf{x} - s\boldsymbol{\Omega}) \geqq \rho(\mathbf{x}) - s \quad \text{for } \mathbf{x}, (\mathbf{x} - s\boldsymbol{\Omega}) \in D. \tag{2.13}$$

(This inequality is obvious for $s \geqq \rho(\mathbf{x})$, since then the right side is nonpositive. For $s < \rho(\mathbf{x})$, Fig. 2 and the triangle inequality give

$$\rho(\mathbf{x} - s\boldsymbol{\Omega}) + s \geqq |\mathbf{x}_0 - \mathbf{x}| \geqq \rho(\mathbf{x}),$$

which is equivalent to (2.13).)

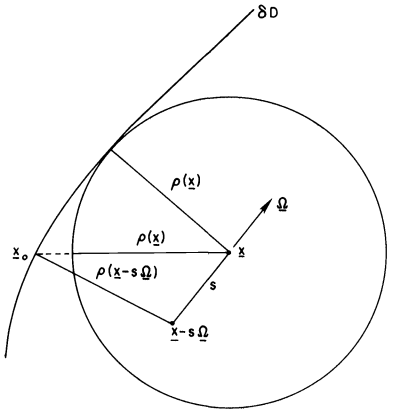


FIG. 2. *Geometrical properties of* $\rho(\mathbf{x})$.

Now we can prove our main result:

THEOREM 3. *Let there exist a function of* $\Psi(E) \in X$ *and a constant* $\alpha \geqq 0$ *such that*

$$f(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \Psi(E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in \Gamma^-, \tag{2.14}$$

and

$$\alpha\Psi(E) + (\Sigma_S\Psi)(\mathbf{x}, E) \leqq \sigma_T(\mathbf{x}, E)\Psi(E), \qquad (\mathbf{x}, E) \in R. \tag{2.15}$$

Then the solution ψ of problem (2.1), (2.2) satisfies

$$0 \leqq \psi(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \Psi(E)\, e^{-\alpha\rho(\mathbf{x})}, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P. \tag{2.16}$$

Proof. ψ exists and is bounded and nonnegative by Theorem 1. To obtain the right half of the inequality (2.16), we shall derive

$$(I - L)(\Psi(E)\, e^{-\alpha\rho(\mathbf{x})}) = Q + W, \tag{2.17a}$$

where

$$W \in X \quad \text{and} \quad W \geqq 0. \tag{2.17b}$$

We subtract (2.6) from (2.17a) to obtain

$$(I - L)(\Psi(E)\, e^{-\alpha\rho(\mathbf{x})} - \psi(\mathbf{x}, \boldsymbol{\Omega}, E)) = W,$$

to which Theorem 2 applies, giving the desired result.

Now we shall derive (2.17). First let us note that $\Psi(E)\, e^{-\alpha\rho(\mathbf{x})} \in X$, and since $L: X \to X$ is bounded, $(Q+W) \in X$. Then since $Q \in X$, we have $W \in X$. Therefore, it remains to show $W \geqq 0$, or equivalently,

$$(I-L)(\Psi(E)\, e^{-\alpha\rho(\mathbf{x})}) \geqq Q(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P. \tag{2.18}$$

To derive (2.18), we use the inequality (2.13) and get

$$-L(\Psi(E)\, e^{-\alpha\rho(\mathbf{x})}) = -\int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \exp\left(-\int_0^t \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right)(\Sigma_S\Psi)(\mathbf{x}-t\boldsymbol{\Omega}, E)\, e^{-\alpha\rho(\mathbf{x}-t\boldsymbol{\Omega})}\, dt$$

$$\geqq -e^{-\alpha\rho(\mathbf{x})} \int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \exp\left(\alpha t - \int_0^t \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right)(\Sigma_S\Psi)(\mathbf{x}-t\boldsymbol{\Omega}, E)\, dt.$$

However, the inequality (2.15) can be written as

$$-(\Sigma_S\Psi)(\mathbf{x}, E) \geqq (\alpha - \sigma_T(\mathbf{x}, E))\Psi(E),$$

and so

$$-L(\Psi(E)\, e^{-\alpha\rho(\mathbf{x})})$$

$$\geqq e^{-\alpha\rho(\mathbf{x})} \int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \exp\left(\alpha t - \int_0^t \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right)(\alpha - \sigma_T(\mathbf{x}-t\boldsymbol{\Omega}, E))\Psi(E)\, dt$$

$$= \Psi(E)\, e^{-\alpha\rho(\mathbf{x})}\left\{\exp\left(\alpha d(\mathbf{x}, \boldsymbol{\Omega}) - \int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right) - 1\right\}.$$

Rearranging this inequality and making use of (2.12), (2.14) and (2.8) gives

$$(I-L)(\Psi(E)\, e^{-\alpha\rho(x)}) \geqq e^{\alpha(d(\mathbf{x},\boldsymbol{\Omega})-\rho(\mathbf{x}))}\Psi(E) \exp\left(-\int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right)$$

$$\geqq \Psi(E) \exp\left(-\int_0^{d(\mathbf{x},\boldsymbol{\Omega})} \sigma_T(\mathbf{x}-s\boldsymbol{\Omega}, E)\, ds\right)$$

$$\geqq Q(\mathbf{x}, \boldsymbol{\Omega}, E).$$

This establishes the inequality (2.18) and completes the proof of the theorem. □

At this point, we shall briefly discuss the existence of $\Psi(E)$ and α, as described in Theorem 3. For the special case $\Sigma_S = 0$, which corresponds to a purely absorbing medium, we can obviously take

$$\alpha = \inf_R \sigma_T(\mathbf{x}, E),$$

$$\Psi(E) = \sup_{\mathbf{x}\in\partial D, \boldsymbol{\Omega}\cdot\mathbf{n}<0} f(\mathbf{x}, \boldsymbol{\Omega}, E).$$

If $\Sigma_S \neq 0$ and D is homogeneous, i.e., if σ_T and σ_S are independent of $\mathbf{x}$, then it is useful to consider the eigenvalue problem[1]

$$\lambda\chi(E) = -\sigma_T(E)\chi(E) + \int \sigma_S(E' \to E)\chi(E')\, dE', \tag{2.19}$$

where we have defined

$$\sigma_S(E' \to E) = \int \sigma_S(\boldsymbol{\Omega}' \cdot \boldsymbol{\Omega}, E' \to E)\, d\boldsymbol{\Omega}'.$$

[1] This problem is closely related to that of computing the time eigenvalues for an infinite medium, for which a lengthy discussion is given in [7].

Rearranging (2.19) gives

$$\chi(E) = A(\lambda)\chi(E) \equiv \int \frac{\sigma_S(E' \to E)}{\sigma_T(E) + \lambda} \chi(E')\, dE'. \tag{2.20}$$

The condition (2.5) implies that for

$$\lambda > -\sigma_m \equiv -\inf_B \sigma_T(E),$$

$A(\lambda)$ is a bounded, positive operator on the space of bounded functions of E with sup norm. Moreover, $A(\lambda_1) > A(\lambda_2)$ for $-\sigma_m < \lambda_1 < \lambda_2$, and $A(\lambda) \to 0$ as $\lambda \to \infty$.

If[2] σ_S and σ_T are such that for each λ, $A(\lambda)$ possesses an eigenvalue $\mu(\lambda)$ with a corresponding positive eigenfunction $\chi_\lambda(E)$, then $\mu(\lambda)$ is a continuous, decreasing function of λ with $\mu(+\infty) = 0$. If also $\mu(\lambda) > 1$ as $\lambda \to -\sigma_m$ then there exists a unique value $\lambda_0 > -\sigma_m$ such that $\mu(\lambda_0) = 1$, and then (2.19) is satisfied with $\lambda = \lambda_0$ and $\chi(E) = \chi_{\lambda_0}(E)$. If, in addition, $\lambda_0 \leqq 0$, then α and $\Psi(E)$ in Theorem 3 exist and may be taken to be

$$\alpha = -\lambda_0, \qquad \Psi(E) = a\chi_{\lambda_0}(E),$$

where a is a suitable nonnegative constant.

The following is a simple consequence of Theorem 3.

THEOREM 4. (Maximum Principle). *If*

$$\sigma_T(\mathbf{x}, E) \geqq \int \sigma_S(\mathbf{x}, \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}', E' \to E)\, d\boldsymbol{\Omega}'\, dE', \qquad (\mathbf{x}, E) \in R, \tag{2.21}$$

then

$$\psi(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \sup_{\Gamma^-} f, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P. \tag{2.22}$$

Proof. The inequality (2.21) implies that in Theorem 3 we may take

$$\Psi(E) = \sup_{\Gamma^-} f(\mathbf{x}, \boldsymbol{\Omega}, E)$$

and $\alpha = 0$. Then Theorem 3 immediately gives (2.22). □

The proofs of the following two theorems are nearly identical to those of Theorems 3 and 4.

THEOREM 5. *Let there exist a function* $\Phi(E) \in X$ *and a constant* $\beta \geqq 0$ *such that*

$$\Phi(E) \leqq f(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in \Gamma^-,$$

and

$$\beta\Phi(E) + \sigma_T(\mathbf{x}, E)\Phi(E) \leqq (\Sigma_S\Phi)(\mathbf{x}, E), \qquad (\mathbf{x}, E) \in R.$$

Then the solution ψ of the (assumed subcritical) problem (2.1), (2.2) satisfies

$$\Phi(\mathrm{E})\, e^{\beta\rho(\mathbf{x})} \leqq \psi(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P.$$

THEOREM 6. (Minimum Principle). *If*

$$\sigma_T(\mathbf{x}, E) \leqq \int \sigma_S(\mathbf{x}, \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}', E' \to E)\, d\boldsymbol{\Omega}'\, dE', \qquad (\mathbf{x}, E) \in R,$$

[2] Conditions on $A(\lambda)$ which guarantee this are given in [8], [9], [10]. Generally, if $\mu(\lambda)$ exists, it is not less in magnitude than any other eigenvalue of $A(\lambda)$. If the given transport problem is multigroup [6], then σ_S and σ_T are step functions of E' and E with a finite number of jumps, and $\mu(\lambda)$, $\chi_\lambda(E)$ exist provided inf $\sigma_S > 0$.

then

$$\inf_{\Gamma^-} f \leqq \psi(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P.$$

For certain problems, the existence of β and $\Phi(E)$, as described in Theorem 5, can be shown. For example, if D is homogeneous and (2.19) has a solution λ_0, $\chi_{\lambda_0}(E)$ with $\lambda_0 \geqq 0$ and $\chi_{\lambda_0}(E) > 0$, then we may take

$$\beta = \lambda_0, \qquad \Phi(E) = a\chi_{\lambda_0}(E),$$

where a is an appropriate nonnegative constant. Also, we show later in § 3 that the assumptions of Theorems 5 and 6 are not necessarily inconsistent with the subcriticality condition on D, and that the assumptions of Theorems 3–6 reduce to explicit conditions for one-group problems.

Now we shall consider adjoint energy-dependent neutron transport problems. These can be written as [6]

$$-\boldsymbol{\Omega} \cdot \nabla_x \psi^* + \Sigma_T \psi^* = \Sigma_S^* \psi^*, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P,$$

$$\psi^*(\mathbf{x}, \boldsymbol{\Omega}, E) = g(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in \Gamma^+,$$

where Σ_S^* is defined by (2.4) with E' and E interchanged in σ_S, and $\Gamma^+ = \{(\mathbf{x}, \boldsymbol{\Omega}, E) | \mathbf{x} \in \partial D, \boldsymbol{\Omega} \cdot \mathbf{n} > 0, E \in B\}$. If we define

$$\eta(\mathbf{x}, \boldsymbol{\Omega}, E) = \psi^*(\mathbf{x}, -\boldsymbol{\Omega}, E), \tag{2.23}$$

then the problem for η is

$$\boldsymbol{\Omega} \cdot \nabla_x \eta + \Sigma_T \eta = \Sigma_S^* \eta, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P, \tag{2.24}$$

$$\eta(\mathbf{x}, \boldsymbol{\Omega}, E) = g(\mathbf{x}, -\boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in \Gamma^-. \tag{2.25}$$

Thus, η satisfies a "forward" transport problem of the type we have considered above for ψ. In fact, the only difference between problems (2.24), (2.25) and (2.1), (2.2) is that E and E' are interchanged in the kernel of the scattering operator Σ_S. Thus, all of the above results for ψ can be applied to η, of course provided one everywhere replaces $\sigma_S(\mathbf{x}, \boldsymbol{\Omega}' \cdot \boldsymbol{\Omega}, E' \to E)$ by $\sigma_S(\mathbf{x}, \boldsymbol{\Omega}' \cdot \boldsymbol{\Omega}, E \to E')$. Since the resulting pointwise bounds on η are independent of $\boldsymbol{\Omega}$, then by (2.23) these bounds apply to ψ^* also. Finally, it is clear that if η is bounded from above (or below) by the supremum (infimum) of its values on Γ^-, then ψ^* is bounded from above (or below) by the supremum (infimum) of its values on Γ^+.

At this point, we can make contact with the ideas in [2] and [3], mentioned in § 1. If D is a spatial medium with the property that for every particle which enters a collision, exactly one particle is emitted, then the medium D is termed "locally conservative" in [2] and [3]; a necessary and sufficient condition for this is

$$\sigma_T(\mathbf{x}, E) = \int \sigma_S(\mathbf{x}, \boldsymbol{\Omega}' \cdot \boldsymbol{\Omega}, E \to E') \, d\boldsymbol{\Omega}' \, dE', \qquad (\mathbf{x}, E) \in R. \tag{2.26}$$

Under this condition, Theorems 4 and 6 both apply to the adjoint transport problem, and so

$$\inf_{\Gamma^+} g \leqq \psi^*(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \sup_{\Gamma^+} g, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P. \tag{2.27}$$

A weaker version of the right half of this inequality, described in § 1, is derived for time-dependent problems in [2] and [3].

The following simple direct proof of (2.27) can be given. Let

$$\xi(\mathbf{x}, \boldsymbol{\Omega}, E) = \psi^*(\mathbf{x}, \boldsymbol{\Omega}, E) - \inf_{\Gamma^+} g,$$

and

$$\theta(\mathbf{x}, \boldsymbol{\Omega}, E) = \sup_{\Gamma^+} g - \psi^*(\mathbf{x}, \boldsymbol{\Omega}, E).$$

The functions ξ and θ satisfy the adjoint transport equation, due to the condition (2.26), and are nonnegative on Γ^+, due to their definitions. Hence, they are nonnegative, and so we obtain (2.27).

In [1], a bound on the solution ψ of (2.6) is obtained in the following way. First, it is clear that (in the sup norm)

$$\|Q\| = \sup_{\Gamma^-} f \equiv M.$$

Next, the following estimate for $\|L\|$ is obtained:

$$\|L\| \leqq c.$$

(See (2.9).) With c required to be less than 1, one obtains immediately from (2.10),

$$\psi(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \|\psi\| \leqq \frac{M}{1-c}, \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P.$$

This bound is a constant, i.e., independent of $\mathbf{x}$, $\boldsymbol{\Omega}$, and E, but it appears to be applicable to problems for which Theorem 3 does not apply (since Theorem 3 depends on the existence of a suitable function $\Psi(E)$). However, this bound gives no information about spatial decay from boundaries, and one cannot derive a maximum principle from it.

3. The one-speed transport equation. We consider the one-speed transport equation for two reasons. First, it is a model transport equation (all particles are assumed to have the same energy) which has received wide theoretical attention. Second, our results here are more general and explicit than in the energy-dependent case.

We take D, ∂D, $\mathbf{n}$, and S to be defined as in § 2. Then the one-speed transport problem is [5], [6]

(3.1)

$$\boldsymbol{\Omega} \cdot \nabla \psi(\mathbf{x}, \boldsymbol{\Omega}) + \sigma(\mathbf{x})\psi(\mathbf{x}, \boldsymbol{\Omega}) = \sigma(\mathbf{x})c(\mathbf{x}) \int k(\mathbf{x}, \boldsymbol{\Omega}' \cdot \boldsymbol{\Omega})\psi(\mathbf{x}, \boldsymbol{\Omega}')\, d\boldsymbol{\Omega}', \qquad \mathbf{x} \in D, \quad \boldsymbol{\Omega} \in S,$$

(3.2)
$$\psi(\mathbf{x}, \boldsymbol{\Omega}) = f(\mathbf{x}, \boldsymbol{\Omega}), \qquad \mathbf{x} \in \partial D, \quad \boldsymbol{\Omega} \cdot \mathbf{n} < 0.$$

Here, $\psi(\mathbf{x}, \boldsymbol{\Omega})$ is the flux of particles at $\mathbf{x} \in D$ travelling in the direction $\boldsymbol{\Omega} \in S$, $\sigma^{-1}(\mathbf{x})$ is the mean free path at $\mathbf{x}$, $c(\mathbf{x})$ is the average number of particles emitted per collision at $\mathbf{x}$ ($c(\mathbf{x}) = 1$ in a conservative medium), k is a nonnegative measurable scattering integral normalized by

(3.3)
$$\int k(\mathbf{x}, \boldsymbol{\Omega}' \cdot \boldsymbol{\Omega})\, d\boldsymbol{\Omega}' = 1, \qquad \mathbf{x} \in D,$$

and f is bounded, nonnegative, and measurable. We also require $\sigma(\mathbf{x})$ and $c(\mathbf{x})$ to be nonnegative and measurable.

Equations (3.1) and (3.2) can be converted into an integral equation which is completely analogous to (2.6), and by repeating the analysis in § 2 with very minor modifications, we obtain the following results:

THEOREM 7. *If*

$$\sup_{\mathbf{x}\in\partial D,\boldsymbol{\Omega}\cdot\mathbf{n}<0} f(\mathbf{x},\boldsymbol{\Omega}) = M, \tag{3.4}$$

and

$$\inf_{\mathbf{x}\in D} \sigma(\mathbf{x})[1-c(\mathbf{x})] = \alpha \geqq 0,$$

then

$$0 \leqq \psi(\mathbf{x},\boldsymbol{\Omega}) \leqq M e^{-\alpha\rho(\mathbf{x})}, \qquad \mathbf{x}\in D, \quad \boldsymbol{\Omega}\in S,$$

and consequently ψ is bounded from above by the supremum of its values on ∂D for $\boldsymbol{\Omega}\cdot\mathbf{n}<0$.

THEOREM 8. *If D is subcritical,*

$$\inf_{\mathbf{x}\in\partial D,\boldsymbol{\Omega}\cdot\mathbf{n}<0} f(\mathbf{x},\boldsymbol{\Omega}) = m, \tag{3.5}$$

and

$$\sup_{\mathbf{x}\in D} \sigma(\mathbf{x})[c(\mathbf{x})-1] = \beta \geqq 0,$$

then

$$m e^{\beta\rho(\mathbf{x})} \leqq \psi(\mathbf{x},\boldsymbol{\Omega}), \qquad \mathbf{x}\in D, \quad \boldsymbol{\Omega}\in S,$$

and consequently ψ is bounded from below by the infimum of its values on ∂D for $\boldsymbol{\Omega}\cdot\mathbf{n}<0$.

We note that Theorem 7 holds if $0 \leqq c(\mathbf{x}) \leqq 1$, Theorem 8 holds if $1 \leqq c(\mathbf{x}) \leqq c_0$ (where c_0 is the constant critical value of c), and the results of both theorems are independent of the scattering kernel k. We also have:

THEOREM 9. *If $\sigma(\mathbf{x})[1-c(\mathbf{x})]=0$ for each $\mathbf{x}\in D$, then with m and M defined as in* (3.4) *and* (3.5), *ψ satisfies*

$$m \leqq \psi(\mathbf{x},\boldsymbol{\Omega}) \leqq M, \qquad \mathbf{x}\in D, \quad \boldsymbol{\Omega}\in S.$$

Theorem 9 applies in a conservative region $[c(\mathbf{x})=1]$, a vacuum $[\sigma(\mathbf{x})=0]$, or a composite region consisting of a conservative part and a vacuum part. Also, Theorem 9 has a simple direct proof which is analogous to that for (2.27).

The remarks made at the end of § 2 concerning adjoint problems apply here also, but now with the simplification that $\eta(\mathbf{x},\boldsymbol{\Omega})$ satisfies *exactly* (3.1) with a boundary condition of the type (3.2). Therefore, all of the results in this section apply directly to η, and hence to ψ^*, with no need to redefine α or β.

We conclude this section with three example problems for which the bounds in Theorems 7 and 8 are "sharp".

First, we consider (3.1) and (3.2), $c(\mathbf{x})=1$ and $f(\mathbf{x},\boldsymbol{\Omega})=M=$ constant. Then the solution of (3.1) and (3.2) is exactly $\psi(\mathbf{x},\boldsymbol{\Omega})=M$, which agrees with Theorems 7 and 8 since $\alpha=\beta=0$.

Second, we take $c(\mathbf{x})=0$, $\sigma(\mathbf{x})=\sigma=$ constant and $f(\mathbf{x},\boldsymbol{\Omega})=M=$ constant. Then

$$\psi(\mathbf{x},\boldsymbol{\Omega}) = M e^{-\sigma d(\mathbf{x},\boldsymbol{\Omega})},$$

and so

$$\sup_{\mathbf{\Omega}\in S} \psi(\mathbf{x}, \mathbf{\Omega}) = M e^{-\sigma\rho(\mathbf{x})}.$$

This is exactly the upper bound on ψ given in Theorem 7, since $\alpha = \sigma$.

Third, we take $c(\mathbf{x}) = c = \text{constant}$, $\sigma(\mathbf{x}) = \sigma = \text{constant}$, $f(\mathbf{x}, \mathbf{\Omega}) = M = \text{constant}$, and $k(\mathbf{x}, \mathbf{\Omega}' \cdot \mathbf{\Omega}) = \delta(1 - \mathbf{\Omega}' \cdot \mathbf{\Omega})$. [This (singular) kernel describes a scattering process in which the direction of a particle is unchanged in a collision.] Then for any $c \geqq 0$,

$$\psi(\mathbf{x}, \mathbf{\Omega}) = M e^{-\sigma(1-c)d(\mathbf{x},\mathbf{\Omega})}.$$

For $c \leqq 1$,

$$\sup_{\mathbf{\Omega}\in S} \psi(\mathbf{x}, \mathbf{\Omega}) = M e^{-\sigma(1-c)\rho(\mathbf{x})},$$

which is exactly the bound on ψ given in Theorem 7, since $\alpha = \sigma(1-c)$. For $c \geqq 1$,

$$\inf_{\mathbf{\Omega}\in S} \psi(\mathbf{x}, \mathbf{\Omega}) = M e^{\sigma(c-1)\rho(\mathbf{x})},$$

which is exactly the bound on ψ given in Theorem 8, since $\beta = \sigma(c-1)$ and $m = M$.

4. A counterexample. Here we describe a simple example problem which shows that a maximum principle need not exist in a subcritical medium. We consider the transport problem (2.1), (2.2) with:

$$f(\mathbf{x}, \mathbf{\Omega}, E) = M = \text{constant},$$

$$\sigma_T(\mathbf{x}, E) = 1 \qquad (\text{and so } \Sigma_T = I),$$

$$\Sigma_S\psi(\mathbf{x}, \mathbf{\Omega}, E) = \frac{1}{\psi\pi} \int \frac{\theta(E)\phi(E')}{E_{1/2} - E_0} \psi(\mathbf{x}, \mathbf{\Omega}', E')\, d\mathbf{\Omega}'\, dE',$$

where

$$\theta(E) = \begin{cases} 1, & E_0 \leqq E \leqq E_{1/2}, \\ 0, & E_{1/2} < E \leqq E_1, \end{cases}$$

$$\phi(E) = \begin{cases} 0, & E_0 \leqq E \leqq E_{1/2}, \\ 1, & E_{1/2} < E \leqq E_1, \end{cases}$$

and $E_{1/2}$ is any fixed number between E_0 and E_1.

Since $\Sigma_s^2 = 0$, then one can easily show $L^2 = 0$. Hence, spr $(L) = 0$, and so D is subcritical. Also, the above definitions imply

$$\begin{aligned} \int_{E_0}^{E_1} \int \Sigma_S\psi(\mathbf{x}, \mathbf{\Omega}, E)\, d\mathbf{\Omega}\, dE &= \int_{E_{1/2}}^{E_1} \int \psi(\mathbf{x}, \mathbf{\Omega}, E)\, d\mathbf{\Omega}, dE \\ &< \int_{E_0}^{E_1} \int \psi(\mathbf{x}, \mathbf{\Omega}, E)\, d\mathbf{\Omega}\, dE \\ &= \int_{E_0}^{E_1} \int \Sigma_T\psi(\mathbf{x}, \mathbf{\Omega}, E)\, d\mathbf{\Omega}\, dE. \end{aligned}$$

Therefore, the average number of particles which are scattered out of any collision is less than one.

The explicit solution of this transport problem is

$$\psi(\mathbf{x}, \boldsymbol{\Omega}, E) = \begin{cases} Me^{-d(\mathbf{x},\boldsymbol{\Omega})}, & E_{1/2} < E \leqq E_1, \\ M\left[e^{-d(\mathbf{x},\boldsymbol{\Omega})} + \left(\dfrac{E_1 - E_{1/2}}{E_{1/2} - E_0}\right) \displaystyle\int_0^{d(\mathbf{x},\boldsymbol{\Omega})} e^{-t} H(\mathbf{x} - t\boldsymbol{\Omega})\, dt\right], & E_0 \leqq E \leqq E_{1/2}, \end{cases}$$

where

$$H(\mathbf{x}) = \frac{1}{\psi\pi} \int_{\cdot} e^{-d(\mathbf{x},\boldsymbol{\Omega})}\, d\boldsymbol{\Omega}.$$

The function $H(\mathbf{x})$ satisfies $0 < H(\mathbf{x}) \leqq 1$ for all $\mathbf{x} \in D$. For any point $\mathbf{x}_0 \in D$ and direction $\boldsymbol{\Omega}_0 \in S$, $E_{1/2}$ can be chosen close enough to E_0 to make the ratio $(E_1 - E_{1/2})/(E_{1/2} - E_0)$ sufficiently large so that $\psi(\mathbf{x}_0, \boldsymbol{\Omega}_0, E) > M$ for $E_0 \leqq E \leqq E_{1/2}$. In such a situation, the maximum principle cannot hold. The mechanism which causes this is the scattering law, which sends every scattered particle initially outside the interval $[E_0, E_{1/2}]$ into this interval, thus (potentially) causing the particle flux in this interval to be greater than the largest value of the incident flux on the boundary.

Although the above problem does not, in general, possess a maximum principle, we note that Theorem 4 applies to the adjoint problem, and so the adjoint flux ψ^* does satisfy a maximum principle:

$$\psi^*(\mathbf{x}, \boldsymbol{\Omega}, E) \leqq \sup_{\Gamma^+} \psi^*(\mathbf{x}, \boldsymbol{\Omega}, E), \qquad (\mathbf{x}, \boldsymbol{\Omega}, E) \in P.$$

Acknowledgment. The author wishes to thank Professor M. Williams for several discussions regarding the results in [2] and [3] and the referee for suggesting the direct proof of (2.27).

REFERENCES

[1] K. M. CASE AND P. F. ZWEIFEL, *Existence and uniqueness theorems for the neutron transport equation*, J. Math. Phys., 4 (1963), pp. 1376–1385.

[2] A. BENSOUSSAN, J. L. LIONS AND G. C. PAPANICOLAOU, *Boundary layers and homogenization of transport processes*, Lecture Notes, Dept. of Mathematics, University of Utah, Salt Lake City, Utah, 1975; also J. Publ. RIMS, Kyoto University, 15 (1979), pp. 53–157.

[3] M. WILLIAMS, *Homogenization of linear transport problems*, Thesis Dissertation, New York University, 1976. See also: M. Williams, *The validity of certain homogenization expansions*, Ann. Nucl. Energy, 7 (1980), pp. 257–266.

[4] G. C. PAPANICOLAOU, *Asymptotic analysis of transport processes*, Bull. Amer. Math. Soc., 81 (1975), pp. 330–392.

[5] K. M. CASE AND P. F. ZWIEFEL, *Linear Transport Theory*, Addison-Wesley Publishing Co., Reading, Mass., 1967.

[6] G. I. BELL AND S. GLASSTONE, *Nuclear Reactor Theory*, Van Nostrand-Reinhold, New York, 1970.

[7] M. M. R. WILLIAMS, *Mathematical Methods in Particle Transport Theory*, Wiley-Interscience, New York, 1971, pp. 265–273.

[8] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Amer. Math. Soc. Transl., Series 1, 10 (1962).

[9] M. A. KRASNOSEL'SKII, *Positive Solutions of Operator Equations*, P. Noordhoff Ltd., Groningen, the Netherlands, 1964.

[10] T. E. HARRIS, *The Theory of Branching Processes*, Springer-Verlag, Berlin, 1963.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0001 $01.00/0

MARKOV CHANNELS ARE ASYMPTOTICALLY MEAN STATIONARY*

JOHN C. KIEFFER† AND MAURICE RAHE‡

Abstract. A type of discrete channel is defined which includes the finite-state channel of Blackwell et al. (Ann. Math. Statist., 29 (1958), pp. 1209–1220) and the finite-state source encoder of Shannon (Bell System Tech. J., 27 (1948), pp. 379–423, 623–656) as special cases. It is shown that if the input source to the Markov channel is asymptotically mean stationary in the sense of Gray and Kieffer, then the resulting input-output pair measure is asymptotically mean stationary also. An application to probability theory is given regarding the asymptotic behavior of a sequence of random stochastic matrices.

Introduction. Let $(\Omega, \mathcal{F})$ be a measurable space and $T: \Omega \to \Omega$ a measurable transformation. Following Gray and Kieffer (1980), we say a probability measure μ on $\mathcal{F}$ is *asymptotically mean stationary* (AMS) if $\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \mu(T^{-i}E)$ exists, for all $E \in \mathcal{F}$.

We state some properties of AMS measures which will be needed in this paper. We refer the reader to Gray and Kieffer (1980) for their proof.

Properties of AMS measures. (In the following, μ is a fixed probability measure on $(\Omega, \mathcal{F})$.)

1. μ is AMS if and only if there exists a probability measure $\bar{\mu}$ on $(\Omega, \mathcal{F})$, stationary with respect to T, such that $\mu(E) = \bar{\mu}(E)$ for every invariant set $E \in \mathcal{F}$.

2. μ is AMS if there exists a probability measure $\bar{\mu}$ on $\mathcal{F}$, stationary with respect to T, such that if $E \in \mathcal{F}$ is invariant and $\bar{\mu}(E) = 0$ then $\mu(E) = 0$.

3. If T is invertible, μ is AMS if and only if there exists a probability measure $\bar{\mu}$ on $\mathcal{F}$, stationary with respect to T, such that μ is absolutely continuous with respect to $\bar{\mu}$.

4. If μ is AMS and $f: \Omega \to (-\infty, \infty)$ is a bounded measurable function, then $\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} f \cdot T^i$ exists a.e. $[\mu]$.

Property 4 indicates why AMS measures are important—namely, because the individual ergodic theorem holds for bounded functions. We refer the reader to Gray and Kieffer (1980) where numerous examples are given to show how AMS measures may arise.

We define a *source* $[\Omega, \mu]$ to be the pair consisting of some measurable space Ω and a probability measure μ on Ω. If there is understood to be some measurable transformation $T: \Omega \to \Omega$, we say the source $[\Omega, \mu]$ is *stationary*, *ergodic*, or AMS if μ is respectively stationary, ergodic, or AMS relative to T.

A *channel* is a triple $[\Omega, \Lambda, \nu]$, where Ω, Λ are measurable spaces and $\nu = \{\nu_x : x \in \Omega\}$ is a family of probability measures on Λ, such that for each measurable subset E of Λ, the map $x \to \nu_x(E)$ from $\Omega \to [0, 1]$ is measurable. If $[\Omega, \mu]$ is a source and $[\Omega, \Lambda, \nu]$ a channel, $\mu\nu$ denotes the measure on the product measurable space $\Omega \times \Lambda$ such that

$$\mu\nu(E) = \int_\Omega \nu_x(E_x)\, d\mu(x)$$

for each measurable subset E of $\Omega \times \Lambda$, where $E_x = \{y \in \Lambda : (x, y) \in E\}$. Suppose it is understood that there are certain measurable transformations $T_1 : \Omega \to \Omega$ and $T_2 : \Lambda \to \Lambda$. Let $T_1 \times T_2 : \Omega \times \Lambda \to \Omega \times \Lambda$ be the measurable transformation such that

$$(T_1 \times T_2)(\omega, \lambda) = (T_1\omega, T_2\lambda), \qquad \omega \in \Omega, \ . \ \lambda \in \Lambda.$$

* Received by the editors August 8, 1979.

† Department of Mathematics, University of Missouri-Rolla, Rolla, MO 65401. The research of this author was supported by the National Science Foundation under grant 76-02276-A01.

‡ Department of Mathematics, Texas A. and M. University, College Station, TX 77843.

We define the channel $[\Omega, \Lambda, \nu]$ to be *stationary* if

$$\nu_{T_1 x}(E) = \nu_x(T_2^{-1}E), \qquad x \in \Omega,$$

E a measurable subset of Λ. $[\Omega, \Lambda, \nu]$ is AMS if for every AMS source $[\Omega, \mu]$, $\mu\nu$ is AMS relative to $T_1 \times T_2$. $[\Omega, \Lambda, \nu]$ is *ergodic* AMS if it is AMS and $\mu\nu$ is ergodic relative to $T_1 \times T_2$ for every ergodic AMS source $[\Omega, \mu]$.

If $(\Omega, \mathcal{F})$ is a measurable space, we let $(\Omega_1^\infty, \mathcal{F}_1^\infty)$ be the measurable space of one-sided sequences from Ω; that is, Ω_1^∞ is the set of all sequences $(\omega_1, \omega_2, \cdots)$ from Ω and $\mathcal{F}_1^\infty$ is the usual product σ-field of subsets of Ω_1^∞. We let $(\Omega^\infty, \mathcal{F}^\infty)$ be the measurable space of two-sided sequences from Ω; that is, Ω^∞ is the set of all sequences $\omega = (\omega_i)_{i=-\infty}^{\infty}$ from Ω and $\mathcal{F}^\infty$ is the usual product σ-field. On the sequence spaces Ω_1^∞, Ω^∞, we always fix the measurable transformation to be the appropriate shift. That is, the shift on Ω_1^∞ is the map $T: \Omega_1^\infty \to \Omega_1^\infty$ such that

$$T(\omega_1, \omega_2, \cdots) = (\omega_2, \omega_3, \cdots);$$

the shift on Ω^∞ is the map $T: \Omega^\infty \to \Omega^\infty$ such that $T(\omega_i) = (\omega_i')$, where $\omega_i' = \omega_{i+1}$, $i \in Z$, Z denoting the set of all integers. We use the symbol T to denote a shift; the context should make clear what space the shift operates on. If $x \in \Omega_1^\infty$ or Ω^∞, x_i denotes the ith coordinate of x, and $[x]_j^k$ denotes $(x_j, x_{j+1}, \cdots, x_k)$.

Sources of the form $[\Omega_1^\infty, \mu]$, $[\Omega^\infty, \mu]$ are respectively called *one-sided* and *two-sided* sources. Channels of the form $[\Omega_1^\infty, \Lambda_1^\infty, \nu]$, $[\Omega^\infty, \Lambda^\infty, \nu]$ are respectively called *one-sided* and *two-sided* channels.

We fix for the rest of the paper measurable spaces $(A, \mathcal{A})$, $(B, \mathcal{B})$, where $(A, \mathcal{A})$ is arbitrary but B is a finite set and $\mathcal{B}$ is the set of all subsets of B. We take B of the form $B = \{1, 2, \cdots, b\}$ for some positive integer b. We let $\{Y_i : i \in Z\}$ denote the family of maps from $B^\infty \to B$ such that $Y_i(y) = y_i$, $y \in B^\infty$, $i \in Z$. We let $\{Y_i' : i = 1, 2, \cdots\}$ denote the maps from $B_1^\infty \to B$ such that $Y_i'(y) = y_i$, $y \in B_1^\infty$, $i = 1, 2, \cdots$.

We let $\mathcal{P}$ be the set of all $b \times b$ stochastic matrices $P = \{P(i, j) : i, j = 1, \cdots, b\}$. We get a metric on $\mathcal{P}$ by defining the distance $\|P_1 - P_2\|$ between $P_1, P_2 \in \mathcal{P}$ to be $\sum_{i,j=1}^{b} [P_1(i, j) - P_2(i, j)]^2$. (With this metric, $\mathcal{P}$ may be thought of as a subset of b^2-dimensional Euclidean space.) With this metric, $\mathcal{P}$ is a compact topological space. We adjoin to $\mathcal{P}$ the σ-field $\beta(\mathcal{P})$ of all Borel subsets of $\mathcal{P}$, making $\mathcal{P}$ a measurable space. Hence, the one-sided and two-sided measurable spaces $(\mathcal{P}_1^\infty, \beta(\mathcal{P})_1^\infty)$, $(\mathcal{P}^\infty, \beta(\mathcal{P})^\infty)$ are defined. We can also place on $\mathcal{P}_1^\infty$ or $\mathcal{P}^\infty$ the product topology, which by Tychonoff's theorem is compact. We note that if $\beta(\mathcal{P}_1^\infty)$, $\beta(\mathcal{P}^\infty)$ denote the Borel subsets of the topological spaces $\mathcal{P}_1^\infty$, $\mathcal{P}^\infty$, then $\beta(\mathcal{P}_1^\infty) = \beta(\mathcal{P})_1^\infty$ and $\beta(\mathcal{P}^\infty) = \beta(\mathcal{P})^\infty$. Hence, for example, the open and closed sets in $\mathcal{P}^\infty(\mathcal{P}_1^\infty)$ are in $\beta(\mathcal{P})^\infty(\beta(\mathcal{P})_1^\infty)$.

We define a map $\phi: A_1^\infty \to \mathcal{P}_1^\infty$ to be *stationary* if $\phi \cdot T = T \cdot \phi$. Similarly, we define what it means for a map $\phi: A^\infty \to \mathcal{P}^\infty$ to be stationary.

If $P \in \mathcal{P}_1^\infty$, we let $\mathcal{M}(P)$ be the set of all probability measures on $\mathcal{B}_1^\infty$ with respect to which $Y_1', Y_2', \cdots$ is a Markov chain with transition matrices $P_1, P_2, \cdots$. That is, $\lambda \in \mathcal{M}(P)$ if and only if

$$\lambda[Y_1' = y_1, \cdots, Y_n' = y_n] = \lambda[Y_1' = y_1] \prod_{i=1}^{n-1} P_i(y_i, y_{i+1}),$$
$$y_1, \cdots, y_n \in B, \quad n = 1, 2, \cdots.$$

Similarly, if $P \in \mathcal{P}^\infty$, we let $\mathcal{M}(P)$ be the set of all probability measures λ on $\mathcal{B}^\infty$ such that

$$\lambda[Y_m = y_m, \cdots, Y_n = y_n] = \lambda[Y_m = y_m] \prod_{i=m}^{n-1} P_i(y_i, y_{i+1}),$$
$$m, n \in Z, \quad m \leqq n, \quad y_m, \cdots, y_n \in B.$$

We define a one-sided channel $[A_1^\infty, B_1^\infty, \nu]$ to be *Markov* if there exists a stationary measurable map $\phi: A_1^\infty \to \mathcal{P}_1^\infty$ such that $\nu_x \in \mathcal{M}(\phi(x))$, $x \in A_1^\infty$. Similarly, we define a two-sided channel $[A^\infty, B^\infty, \nu]$ to be Markov if there is a stationary measurable map $\phi: A^\infty \to \mathcal{P}^\infty$ such that $\nu_x \in \mathcal{M}(\phi(x))$, $x \in A^\infty$.

Examples.

(a) Let C, D be finite sets, m be a positive integer, and $E = \{1, 2, \cdots, m\}$. Let $\{P_c: c \in C\}$ be stochastic $m \times m$ matrices. Let $f: E \to D$ be given. Fixing some $s_0 \in E$, we define a channel $[C_1^\infty, (E \times D)_1^\infty, \nu]$ where for all $n = 1, 2, \cdots$, $s_1, \cdots, s_n \in E$, $y_1, \cdots, y_n \in D$, and $x \in C_1^\infty$,

$$\nu_x[(s', y') \in (E \times D)_1^\infty: (s_1', y_1') = (s_1, y_1), \cdots, (s_n', y_n') = (s_n, y_n)]$$

$$= \prod_{i=1}^{n} P_{x_i}(s_{i-1}, s_i) \quad \text{if } y_i = f(s_i), \quad i = 1, \cdots, n,$$

$$= 0 \qquad \text{otherwise.}$$

This is easily seen to be a Markov channel. It is called a *finite-state channel.* It was first defined by Shannon (1948) in a formally different but equivalent form; Blackwell, Breiman, and Thomasian (1958) and Breiman (1960) were the first to prove coding theorems of information theory for this type of channel. The elements of E represent channel states. Initially the channel is in state s_0 and input symbol x_1 is introduced. The channel then moves to state s_1 with probability $P_{x_1}(s_0, s_1)$ and emits the symbol $y_1 = f(s_1)$. Then input x_2 is introduced, the channel moves to state s_2 with probability $P_{x_2}(s_1, s_2)$ and $y_2 = f(s_2)$ is emitted. Successive inputs $x_3, x_4, \cdots$ are processed by the channel in this same manner.

(b) A *finite-state source coder* [Shannon (1948), Ziv (1978)] consists of finite sets C, D, S and maps $f: C \times S \to S$, $g: C \times S \to D$. S represents the set of states of the coder. The coder starts in state $s_1 \in S$, and codes a sequence $x_1, x_2, \cdots$ of input letters from C in the following way: The coder reads in x_1 and then emits the output $y_1 = g(x_1, s_1)$. The coder then moves to state $s_2 = f(x_1, s_1)$. The next symbol x_2 is read in, an output $y_2 = g(x_2, s_2)$ occurs, and then the coder moves to state $s_3 = f(x_2, s_2)$. Proceeding in this way, a sequence $x = (x_1, x_2, \cdots) \in C_1^\infty$ is coded into a sequence $g^*(x) = (y_1, y_2, \cdots) \in D_1^\infty$, while the coder moves through states $f^*(x) = (s_1, s_2, \cdots) \in S_1^\infty$. We may think of the action of this coder as being represented by the Markov channel $[C_1^\infty, (S \times D)_1^\infty, \nu]$, where

$$\nu_x[(s, y) \in (S \times D)_1^\infty: s = f^*(x), y = g^*(x)] = 1, \qquad x \in C_1^\infty.$$

The main result of this paper (Theorems 6 and 7) is that one- and two-sided Markov channels are AMS. Thus, if an AMS source $[A_1^\infty, \mu]$ is the input to a Markov channel $[A_1^\infty, B_1^\infty, \nu]$, the resulting input-output joint distribution $\mu\nu$ is AMS. Thus, the individual ergodic theorem will hold for $\mu\nu$ and also the Shannon-McMillan theorem (see Gray and Kieffer (1980)). From an information-theoretic point of view, this is a useful fact. For example, one of the steps in Breiman's (1960) derivation of the capacity of a finite-state channel was the demonstration that the Shannon-McMillan theorem holds for $\mu\nu$ if $[A_1^\infty, \mu]$ is a Markov source (a type of source defined later in the paper). Since Markov sources are AMS (Theorem 9), we have considerably widened the class of inputs $[A_1^\infty, \mu]$ for the finite-state channel for which the Shannon-McMillan theorem will hold for $\mu\nu$.

Main results. Our starting point on the lengthy road to proving Theorems 6 and 7 is a result due to Blackwell (1945) on the structure of finite state nonhomogeneous Markov chains.

Before stating this result, recall that a subset S of a vector space is a *simplex* if it is the set of all convex combinations of a finite set of affinely independent vectors. There is only one such set of affinely independent vectors which spans S; we will call this set the *basis* for S. The *dimension* of S is the number of vectors in the basis for S.

If Ω is a set and $\mathcal{M}$ is a family of functions from $\Omega \to (-\infty, \infty)$, we let $\mathcal{F}(\mathcal{M})$ denote the smallest σ-field of subsets of Ω with respect to which all the functions in $\mathcal{M}$ are measurable.

Let $\mathcal{T}$ denote the left-hand tail σ-field $\mathcal{T} = \bigcap_{n<0} \mathcal{F}\{Y_i: i \leqq n\}$.

THEOREM 1 (Blackwell, (1945)). *Let $P \in \mathcal{P}^\infty$. Then $\mathcal{M}(P)$ is a simplex. If $\{Q_1, \cdots, Q_k\}$ is the basis for $\mathcal{M}(P)$, then $Q_1, \cdots, Q_k$ are singular over $\mathcal{T}$; that is, there exist disjoint $F_1, \cdots, F_k \in \mathcal{T}$ such that $Q_i(F_i) = 1, i = 1, \cdots, k$. Also, each Q_i is $\mathcal{T}$-trivial; that is, $Q_i(F) = 0$ or $1, F \in \mathcal{T}, i = 1, \cdots, k$.*

DEFINITION. We define $P \in \mathcal{P}^\infty$ to be *recurrent* if for every open subset O of $\mathcal{P}^\infty$ containing P, O also contains $T^i P$ for infinitely many $i > 0$ and infinitely many $i < 0$. Let R be the subset of $\mathcal{P}^\infty$ consisting of all recurrent points. Then R is a measurable subset of $\mathcal{P}^\infty$ (for, in verifying the condition for membership in R, one can restrict oneself to those O in a countable base for the $\mathcal{P}^\infty$-topology). We also note for later use that R is a shift-invariant subset of $\mathcal{P}^\infty$, and that if μ is any probability measure on $(\mathcal{P}^\infty, \beta(\mathcal{P})^\infty)$ stationary with respect to the shift, then $\mu(R) = 1$.

In the following, if C, D are subsets of some common set, let $C \backslash D = \{\omega \in C: \omega \notin D\}$. If $C \subset \Omega$, let $I_C : \Omega \to \{0, 1\}$ be the indicator function for C; that is, I_C is $\{0, 1\}$-valued and is equal to one precisely on C.

LEMMA 1. *Let $P \in \mathcal{P}^\infty$. Let $\{Q_1, \cdots, Q_k\}$ be the basis for $\mathcal{M}(P)$. Then for each $n \in Z$, there are disjoint subsets $B_n^1, \cdots, B_n^k$ of B such that*

$$\lim_{n \to -\infty} Q_j(Y_n \in B_n^j) = 1, \qquad 1 \leqq j \leqq k.$$

Proof. Let $Q = k^{-1} \sum_{j=1}^k Q_j$. Then $Q \in \mathcal{M}(P)$. Fix disjoint $F_1, \cdots, F_k \in \mathcal{T}$ such that $Q_j(F_j) = 1$, $1 \leqq j \leqq k$. By the Markov property,

$$Q(F_j | Y_n) = Q(F_j | Y_n, Y_{n+1}, \cdots) \quad \text{a.e. } [Q], \quad 1 \leqq j \leqq k.$$

So by the martingale convergence theorem

$$Q(F_j | Y_n) \to Q(F_j | \mathcal{B}^\infty) = I_{F_j} \quad \text{a.e. } [Q] \quad \text{as } n \to -\infty, \quad 1 \leqq j \leqq k.$$

It follows that

$$Q(F_j | Y_n) \to 1 \quad \text{a.e. } [Q_j], \quad 1 \leqq j \leqq k,$$

$$Q(F_j | Y_n) \to 0 \quad \text{a.e. } [Q_{j'}], \quad j \neq j', \quad 1 \leqq j, j' \leqq k.$$

For each $n \in Z$, define $B_n^1, \cdots, B_n^k$ as follows:

$$B_n^j = \{y \in B: Q(F_j | Y_n = y) > \tfrac{1}{2}\} \Big\backslash \Big[\bigcup_{t=1}^{j-1} \{y \in B: Q(F_t | Y_n = y) > \tfrac{1}{2}\} \Big], \qquad 1 \leqq j \leqq k.$$

DEFINITION. For each $k = 2, 3, \cdots$, if σ is a probability measure on B, and $P_1, \cdots, P_{k-1} \in \mathcal{P}$, let $\lambda(\sigma, P_1, \cdots, P_{k-1})$ be the probability measure on B^k such that

$$\lambda(\sigma, P_1, \cdots, P_{k-1})(y_1, \cdots, y_k) = \sigma(y_1) \prod_{i=1}^{k-1} P_i(y_i, y_{i+1}), \qquad y_1, \cdots, y_k \in B.$$

THEOREM 2. *Let $P \in R$. Let $\{Q_1, \cdots, Q_k\}$ be the basis for $\mathcal{M}(P)$. For each $n \in Z$ and $1 \leqq j \leqq k$, let*

$$A_n^j = \{y \in B: Q_j(Y_n = y) > 0\}.$$

Then:

(a) *If* $y \in A_n^j$, *then*

$$\sum_{y' \in A_{n+1}^j} P_n(y, y') = 1, \qquad n \in Z, \quad 1 \leqq j \leqq k.$$

(b) *For each* $n \in Z$, *the sets* $A_n^1, \cdots, A_n^k$ *are disjoint.*

Proof. We have

$$1 = Q_j[Y_{n+1} \in A_{n+1}^j] = \sum_{y \in A_n^j} Q_j[Y_n = y] \Big\{ \sum_{y' \in A_{n+1}^j} P_n(y, y') \Big\},$$

from which (a) follows.

Now we show (b). Fix m, a positive integer. Choose $\delta > 0$ so small that if $\hat{P}_1, \cdots, \hat{P}_{2m} \in \mathcal{P}$, $\tilde{P}_1, \cdots, \tilde{P}_{2m} \in \mathcal{P}$, and $\|\hat{P}_i - \tilde{P}_i\| < \delta$, $1 \leqq i \leqq 2m$, then

$$|\lambda(\sigma, \hat{P}_1, \cdots, \hat{P}_{2m})(E) - \lambda(\sigma, \tilde{P}_1, \cdots, \tilde{P}_{2m})(E)| < m^{-1}, \tag{1}$$

for every subset E of B^{2m+1} and every probability measure σ on B. Set $W = \{s < 0: \|P_{s+i} - P_i\| < \delta, -m \leqq i < m\}$. Since P is recurrent, W is infinite. For each $n \in Z$, let $B_n^1, \cdots, B_n^k$ be the disjoint subsets of B given by Lemma 1. Now for each $1 \leqq j \leqq k$,

$$\lim_{s \to -\infty} \lambda(\sigma_s^j, P_{s-m}, \cdots, P_{s+m-1}) \Big(\underset{i=-m}{\overset{m}{\times}} B_{s+i}^j \Big) = 1,$$

where σ_s^j is the distribution of Y_{s-m} under Q_j. Hence, there must exist some $s \in W$ such that

$$\lambda(\sigma_s^j, P_{-m}, \cdots, P_{m-1}) \Big(\underset{i=-m}{\overset{m}{\times}} B_{s+i}^j \Big) > 1 - 2m^{-1}, \qquad 1 \leqq j \leqq k.$$

Doing this for each $m = 1, 2, \cdots$, and then letting $m \to \infty$ through an appropriate subsequence, we will obtain measures $Q^{(1)}, \cdots, Q^{(k)} \in \mathcal{M}(P)$ and for each $n \in Z$, disjoint subsets $E_n^1, \cdots, E_n^k$ of B, such that

$$Q^{(j)} \Big(\underset{i=-\infty}{\overset{\infty}{\times}} E_i^j \Big) = 1, \qquad 1 \leqq j \leqq k.$$

Since $Q^{(1)}, \cdots, Q^{(k)}$ are mutually singular, they are affinely independent, and so, reordering $Q^{(1)}, \cdots, Q^{(k)}$ if necessary, we have $Q_1 = Q^{(1)}, \cdots, Q_k = Q^{(k)}$. We must then have $A_n^j \subset E_n^j$, $n \in Z$, $1 \leqq j \leqq k$, and so (b) follows.

Notation. If Ω is a set and $x \in \Omega^\infty$, then for each $i \in Z$, x_i^∞ denotes the sequence $(x_i, x_{i+1}, \cdots) \in \Omega_1^\infty$. Similarly, if $x \in \Omega_1^\infty$ and $i = 1, 2, \cdots$, x_i^∞ denotes $(x_i, x_{i+1}, \cdots) \in \Omega_1^\infty$.

THEOREM 3. *Let* $P \in R$. *Let* $\{Q_1, \cdots, Q_k\}$ *be the basis for* $\mathcal{M}(P)$. *For each* $n \in Z$, *let* $A_n^1, \cdots, A_n^k$ *be the disjoint subsets of* B *given by Theorem* 2. *Let* $Q \in \mathcal{M}(P_1^\infty)$. *Then*

$$\lim_{n \to \infty} Q\Big[Y_n' \in \bigcup_{j=1}^k A_n^j \Big] = 1.$$

Proof. Let $A_n^{k+1} = B \backslash [\bigcup_{j=1}^k A_n^j]$, $n = 1, 2, \cdots$. By Theorem 2(a),

$$Q[Y_1' \in A_1^{k+1}, \cdots, Y_n' \in A_n^{k+1}] = Q[Y_n' \in A_n^{k+1}], \qquad n = 1, 2, \cdots.$$

Hence, the sequence $\{Q(Y_n' \in A_n^{k+1})\}_{n=1}^\infty$ decreases to a limit α, which we must show to be zero. Suppose $\alpha > 0$. Fix a positive integer $m > 2\alpha^{-1}$. As in the proof of Theorem 2, choose $\delta > 0$ so small that (1) holds. We now define

$$W = \{s > 0: \|P_{s+i} - P_i\| < \delta, -m \leqq i < m\}.$$

By recurrence of P, W is infinite. For each $s \in Z$ and $1 \leqq j \leqq k$, we have

$$\lambda(\sigma_s^j, P_{s-m}, \cdots, P_{s+m-1})\left(\underset{i=-m}{\overset{m}{\times}} A_{s+i}^j\right) = 1,$$

where σ_s^j is the distribution of Y_{s-m} under Q_j. Also, if $s > m$,

$$\lambda(\sigma_s, P_{s-m}, \cdots, P_{s+m-1})\left(\underset{i=-m}{\overset{m}{\times}} A_{s+i}^{k+1}\right) \geqq \alpha,$$

where σ_s is the distribution of Y'_{s-m} under Q. Hence, there must exist $s \in W$ such that

$$\lambda(\sigma_s^j, P_{-m}, \cdots, P_{m-1})\left(\underset{i=-m}{\overset{m}{\times}} A_{s+i}^j\right) > 1 - m^{-1}, \qquad 1 \leqq j \leqq k$$

and

$$\lambda(\sigma_s, P_{-m}, \cdots, P_{m-1})\left(\underset{i=-m}{\overset{m}{\times}} A_{s+i}^j\right) > \frac{\alpha}{2}.$$

If we do this for each m and then let $m \to \infty$ through an appropriate subsequence, we will obtain $Q'_1, \cdots, Q'_{k+1} \in \mathcal{M}(P)$ and disjoint sets $E_1, \cdots, E_{k+1}$ in $\mathcal{B}^\infty$ such that

$$Q'_i(E_i) = 1, \qquad 1 \leqq i \leqq k,$$

and

$$Q'_{k+1}(E_{k+1}) \geqq \frac{\alpha}{2}.$$

This would force the dimension of $\mathcal{M}(P)$ to be greater than k, a contradiction.

In the following, we regard R as a measurable space by adjoining to R the σ-field of subsets of R consisting of those subsets of R which are contained in $\mathcal{B}(\mathcal{P})^\infty$. Since R is shift-invariant, we will regard the shift transformation as a map from $R \to R$.

DEFINITION. Let $[R, B^\infty, \bar{\nu}]$ be the stationary channel such that, if $P \in R$ is such that $\mathcal{M}(P)$ has basis $\{Q_1, \cdots, Q_k\}$, then $\bar{\nu}_P = k^{-1} \sum_{j=1}^k Q_j$.

The channel $[R, B^\infty, \bar{\nu}]$ is clearly stationary, for if $Q_1, \cdots, Q_k$ is the basis for $\mathcal{M}(P)$, then $Q_1 \cdot T^{-1}, \cdots, Q_k \cdot T^{-1}$ is the basis for $\mathcal{M}(TP)$, and so

$$\bar{\nu}_{TP}(E) = k^{-1} \sum_{j=1}^k Q_j \cdot T^{-1}(E) = \bar{\nu}_P(T^{-1}E).$$

The hard part is to show that the map $P \to \bar{\nu}_P(E)$ is a measurable map from $R \to [0, 1]$, for each $E \in \mathcal{B}^\infty$.

THEOREM 4. *For each $E \in \mathcal{B}^\infty$, the map $P \to \bar{\nu}_P(E)$ is a measurable map from $R \to [0, 1]$.*

The proof is lengthy; it will be accomplished by a series of definitions and lemmas.

DEFINITION. If $P \in \mathcal{P}$, and C, D are nonempty subsets of B, we say C P-leads to D if $\sum_{j \in D} P(i, j) = 1$, $i \in C$. Let $\mathcal{B}^*$ be the set of all finite sequences of disjoint nonempty subsets of B. (Note for later use that $\mathcal{B}^*$ is a finite set.) If $P \in \mathcal{P}$ and $(C_1, \cdots, C_j)$, $(D_1, \cdots, D_k) \in \mathcal{B}^*$, we say $(C_1, \cdots, C_j)$ P-leads to $(D_1, \cdots, D_k)$ if and only if $j = k$ and C_i P-leads to D_i, $1 \leqq i \leqq k$. For $N = 0, 1, \cdots$, if $(\tau_{-N}, \cdots, \tau_N) \in (\mathcal{B}^*)^{2N+1}$, we let $E(\tau_{-N}, \cdots, \tau_N)$ be the set of all $P \in R$ such that there exists $\hat{\tau} \in (\mathcal{B}^*)^\infty$ for which $\hat{\tau}_i$ P_i-leads to $\hat{\tau}_{i+1}$, $i \in Z$, and $[\hat{\tau}]_{-N}^N = (\tau_{-N}, \cdots, \tau_N)$.

LEMMA 2. *$E(\tau)$ is a measurable subset of R, $\tau \in (\mathcal{B}^*)^{2N+1}$, $N = 0, 1, \cdots$.*

Proof. Fix N and $(\tau_{-N}, \cdots, \tau_N) \in (\mathcal{B}^*)^{2N+1}$. $E(\tau_{-N}, \cdots, \tau_N)$ will be measurable if it coincides with the set of all $P \in R$ such that for each $K > N$, there exists $(\hat{\tau}_{-K}, \cdots, \hat{\tau}_K) \in (\mathcal{B}^*)^{2K+1}$ such that $\hat{\tau}_i P_i$-leads to $\hat{\tau}_{i+1}$, $-K \leqq i < K$, and $(\hat{\tau}_{-N}, \cdots, \hat{\tau}_N) = (\tau_{-N}, \cdots, \tau_N)$ (for it is easy to see that the latter set is the intersection of a closed subset of $\mathcal{P}^\infty$ with R, and is therefore measurable). Clearly, $E(\tau_{-N}, \cdots, \tau_N)$ is contained in the set just defined. To show the reverse inclusion, let P be in the set just defined. Then we can find, for each $K > N$, a sequence $\tau^{(K)} \in (\mathcal{B}^*)^\infty$ such that $[\tau^{(K)}]_{-N}^N = (\tau_{-N}, \cdots, \tau_N)$ and $\tau_j^{(K)}$ P_j-leads to $\tau_{j+1}^{(K)}$, $-K \leqq j < K$. Placing the discrete topology on $\mathcal{B}^*$ and then the product topology on $(\mathcal{B}^*)^\infty$ shows that $(\mathcal{B}^*)^\infty$ is compact since $\mathcal{B}^*$ is finite. Therefore, we may choose a $\hat{\tau} \in (\mathcal{B}^*)^\infty$ which is a limit of a convergent subsequence of $\{\tau^{(K)}\}_1^\infty$. By definition of convergence in the product topology, we have for each $j = 1, 2, \cdots$ that $[\hat{\tau}]_{-j}^j$ agrees with $[\tau^{(K)}]_{-j}^j$ for infinitely many K. This implies that $\hat{\tau}_i P_i$-leads to $\hat{\tau}_{i+1}$, $i \in Z$, and $[\hat{\tau}]_{-N}^N = (\tau_{-N}, \cdots, \tau_N)$. Hence, $P \in E(\tau_{-N}, \cdots, \tau_N)$.

DEFINITION. For $N = 0, 1, \cdots$, let us call a measurable map $\alpha: R \to (\mathcal{B}^*)^{2N+1}$ *admissible* if $P \in E(\alpha(P))$ for each $P \in R$. If $K > N$, we say $\alpha': R \to (\mathcal{B}^*)^{2K+1}$ *extends* $\alpha: R \to (\mathcal{B}^*)^{2N+1}$ if $\alpha'(P) = (\tau_{-K}, \cdots, \tau_K)$ implies $\alpha(P) = (\tau_{-N}, \cdots, \tau_N)$.

LEMMA 3. *Let N be a nonnegative integer. Let $\alpha: R \to (\mathcal{B}^*)^{2N+1}$ be an admissible measurable map. Then for any $K > N$ there exists an admissible measurable map $\alpha': R \to (\mathcal{B}^*)^{2K+1}$ which extends α.*

Proof. Fix $K > N$. For each $(\tau_{-K}, \cdots, \tau_K) \in (\mathcal{B}^*)^{2K+1}$, let

$$E'(\tau_{-K}, \cdots, \tau_K) = \{P \in R: \alpha(P) = (\tau_{-N}, \cdots, \tau_N)\} \cap E(\tau_{-K}, \cdots, \tau_K).$$

Let $P \in R$. Then $P \in E(\alpha(P))$, since α is admissible. Thus, there exists $\tau \in (\mathcal{B}^*)^\infty$ such that τ_j P_j-leads to τ_{j+1}, $j \in Z$, and $[\tau]_{-N}^N = \alpha(P)$. This implies $P \in E'(\tau_{-K}, \cdots, \tau_K)$. Therefore,

$$\bigcup_{\tau \in (\mathcal{B}^*)^{2K+1}} E'(\tau) = R. \tag{2}$$

Let $\tau^{(1)}, \cdots, \tau^{(s)}$ be an enumeration of the elements of $(\mathcal{B}^*)^{2K+1}$. Let $\alpha': R \to (\mathcal{B}^*)^{2K+1}$ be the measurable map such that

$$\{P \in R: \alpha'(P) = \tau^{(i)}\} = E'(\tau^{(i)}) \Big\backslash \Big[\bigcup_{j<i} E'(\tau^{(j)}) \Big], \qquad i = 1, \cdots, s.$$

(By (2), α' is defined on all of R.) Then $P \in E'(\alpha'(P))$, $P \in R$. It follows that α' is admissible and extends α.

LEMMA 4. *There is a measurable map $\alpha: R \to (\mathcal{B}^*)^\infty$ such that:*

(a) *If $P \in R$, then for each $i \in Z$, $\alpha(P)_i$ has length equal to the dimension of $\mathcal{M}(P)$.*

(b) *If $P \in R$, and the dimension of $\mathcal{M}(P)$ is k, we have, setting $\alpha(P)_i = (\underline{A}_i^1, \cdots, \underline{A}_i^k)$ for each $i \in Z$, that $\underline{A}_i^j P_i$- leads to $\underline{A}_{i+1}^j$, $i \in Z$, $1 \leqq j \leqq k$. There is an ordering of the basis elements $\{Q_1, \cdots, Q_k\}$ of $\mathcal{M}(P)$ so that*

$$Q_j(\times_{i=-\infty}^\infty \underline{A}_i^j) = 1, \qquad 1 \leqq j \leqq k.$$

Proof. Observe that $\bigcup_{\tau \in \mathcal{B}^*} E(\tau) = R$. (In fact, $E(B) = R$.) Let $\tau_1, \cdots, \tau_r$ be an enumeration of $\mathcal{B}^*$ so that if $i < j$ then τ_i has length at least that of τ_j. Define $\alpha_0: R \to \mathcal{B}^*$ to be the admissible measurable map such that

$$\{P \in R: \alpha_0(P) = \tau_i\} = E(\tau_i) \Big\backslash \Big[\bigcup_{j<i} E(\tau_j) \Big].$$

By Lemma 3, pick a sequence $\alpha_1, \alpha_2, \cdots$ of maps such that for each $i = 1, 2, \cdots$, α_i is a measurable admissible map from $R \to (\mathcal{B}^*)^{2i+1}$ which extends α_{i-1}. Define

$\alpha: R \to (\mathcal{B}^*)^\infty$ so that

$$[\alpha(P)]_{-i}^{i} = \alpha_i(P), \qquad i = 0, 1, \cdots.$$

Then α is measurable. Fix $P \in R$. Then by admissibility of each α_i it follows that $\alpha(P)_n$ P_n-leads to $\alpha(P)_{n+1}$, $n \in Z$. Hence, all the $\alpha(P)_n$ have the same length, say k'. We show that k' is the dimension k of $\mathcal{M}(P)$. Let $\{Q_1, \cdots, Q_k\}$ be the basis for $\mathcal{M}(P)$ and for each $n \in Z$, let $A_n^1, \cdots, A_n^k$ be the disjoint subsets of B given by Theorem 2. By Theorem 2 we have $(A_0^1, \cdots, A_0^k) \in \mathcal{B}^*$ and $P \in E(A_0^1, \cdots, A_0^k)$. Hence, by definition of α_0, the length k' of $\alpha_0(P)$ is at least k. Suppose $k' > k$. Letting $\alpha(P)_i = (\underline{A}_i^1, \cdots, \underline{A}_i^{k'})$, $i \in Z$, we have that $\underline{A}_i^j$ P_i-leads to $\underline{A}_{i+1}^j$, $i \in Z$, $1 \leqq j \leqq k'$. Hence, we may construct measures $Q^{(1)}, \cdots, Q^{(k')} \in \mathcal{M}(P)$ such that $Q^{(j)}(\times_{i=-\infty}^{\infty} \underline{A}_i^j) = 1$, $1 \leqq j \leqq k'$. Since $Q^{(1)}, \cdots, Q^{(k')}$ are singular, they are affinely independent; this contradicts the fact that the dimension of $\mathcal{M}(P)$ is only k. Thus, $k' = k$ and so (a) follows. We note that (b) follows by replacing k' by k in the preceding argument.

In the following, if S is a finite set, let $|S|$ denote the cardinality of S.

Proof of Theorem 4. Let $\alpha: R \to (\mathcal{B}^*)^\infty$ be the map of Lemma 4. Let $K: R \to \{1, 2, \cdots\}$ be the map such that $K(P)$ is the dimension of $\mathcal{M}(P)$, $P \in R$. The map K is measurable because α is measurable and $K(P)$ is the length of $\alpha(P)_0$, $P \in R$. Fix $k \in \{1, 2, \cdots\}$. Let $R_k = \{P \in R: K(P) = k\}$. If $P \in R_k$, let

$$\alpha(P)_i = (\underline{A}_i^1(P), \cdots, \underline{A}_i^k(P)).$$

For each $1 \leqq j \leqq k$ and $N = 1, 2, \cdots$, we define a map $Q_j^N: R_k \to \mathcal{M}_N$, where $\mathcal{M}_N$ is the measurable space of probability measures on B^{2N+1} with the obvious σ-field of measurable sets. If $P \in R_k$, $(i_{-N}, \cdots, i_N) \in B^{2N+1}$,

$$Q_j^N(P)(i_{-N}, \cdots, i_N) = \left[\prod_{t=-N}^{N-1} P_t(i_t, i_{t+1})\right] \Big/ |\underline{A}_{-N}^j(P)|, \qquad i_{-N} \in \underline{A}_{-N}^j(P),$$

$$= 0, \qquad \text{otherwise.}$$

Since α is measurable, the maps Q_j^N are measurable. (Measurability of Q_j^N means that for each $E \subset B^{2N+1}$, the map $P \to Q_j^N(P)(E)$ is a measurable map from $R_k \to [0, 1]$.) Note that by Lemma 4(b)

$$Q_j^N(P)\left(\underset{i=-N}{\overset{N}{\times}} \underline{A}_i^j(P)\right) = 1. \tag{3}$$

For each $P \in R_k$ and $1 \leqq j \leqq k$, define $Q_j(P)$ as the element of $\mathcal{M}(P)$ such that, for $M = 1, 2, \cdots$, and $(y_{-M}, \cdots, y_M) \in B^{2M+1}$,

$$Q_j(P)[Y_{-M} = y_{-M}, \cdots, Y_M = y_M]$$
$$= \lim_{N \to \infty} Q_j^N(P)[(\underline{y}_{-N}, \cdots, \underline{y}_N) \in B^{2N+1}: (\underline{y}_{-M}, \cdots, \underline{y}_M) = (y_{-M}, \cdots, y_M)].$$

(The limits all exist; otherwise by (3) there would be two elements of $\mathcal{M}(P)$ concentrated on $\times_{i=-\infty}^{\infty} \underline{A}_i^j(P)$, which is impossible by Lemma 4(b).) It follows that, for $P \in R_k$, $\{Q_1(P), \cdots, Q_k(P)\}$ is the basis for $\mathcal{M}(P)$, and for each $1 \leqq j \leqq k$ and $E \in \mathcal{B}^\infty$, the map $P \to Q_j(P)(E)$ is a measurable map from $R_k \to [0, 1]$. Hence, it follows that the map $P \to \bar{\nu}_P(E)$ is a measurable map from $R \to [0, 1]$ for each $E \in \mathcal{B}^\infty$.

DEFINITION. For each $y \in B$ and $P \in \mathcal{P}_1^\infty$, let $Q^y(\cdot|P)$ be the element of $\mathcal{M}(P)$ such that $Q^y(Y_1' = y|P) = 1$. Then, if $P \in \mathcal{P}^\infty$, $Q \in \mathcal{M}(P)$ and $E \in \mathcal{B}_1^\infty$,

$$Q[(Y_n, \cdots) \in E] = \sum_{y \in B} Q[Y_{n-1} = y] Q^y((Y_2', \cdots) \in E | P_{n-1}^\infty), \qquad n \in Z. \tag{4}$$

If $P \in \mathcal{P}_1^\infty$, $Q \in \mathcal{M}(P)$, $E \in \mathcal{B}_1^\infty$,

$$(5) \qquad Q[(Y'_n, \cdots) \in E] = \sum_{y \in B} Q[Y'_{n-1} = y] Q^y((Y'_2, \cdots) \in E | P_{n-1}^\infty), \qquad n \geqq 2.$$

THEOREM 5. *Let $P \in R$. Let $E \in \bigcap_{n=1}^\infty \mathcal{F}\{Y'_1 : i \geqq n\}$. Suppose $\bar{\nu}_P[(Y_1, Y_2, \cdots) \in E] = 0$. Then $Q(E) = 0$ for every $Q \in \mathcal{M}(P_1^\infty)$.*

Proof. For each $n \in Z$, let $A_n^1, \cdots, A_n^k$ be the disjoint subsets of B given by Theorem 2. By Theorem 3,

$$Q(E) \leqq \sum_{n=1}^\infty Q\left(E \cap \left\{ Y'_n \in \bigcup_{j=1}^k A_n^j \right\}\right).$$

Fix the positive integer n. Since E is a tail event, there exists $E_n \in \mathcal{B}_1^\infty$ such that

$$E = \{(Y'_{n+1}, \cdots) \in E_n\}.$$

By (5),

$$Q\left(E \cap \left\{ Y'_n \in \bigcup_{j=1}^k A_n^j \right\}\right) = \sum_{j=1}^k \sum_{y \in A_n^j} Q(Y'_n = y) Q^y((Y'_2, \cdots) \in E_n | P_n^\infty).$$

By (4),

$$\begin{aligned} 0 = \bar{\nu}_P[(Y_1, \cdots) \in E] &= \bar{\nu}_P[(Y_{n+1}, \cdots) \in E_n] \\ &= \sum_{y \in B} \bar{\nu}_P[Y_n = y] Q^y((Y'_2, \cdots) \in E_n | P_n^\infty). \end{aligned}$$

Since $\bar{\nu}_P[Y_n = y] > 0$ for $y \in \bigcup_{j=1}^k A_n^j$, we must have

$$Q^y((Y'_2, \cdots) \in E_n | P_n^\infty) = 0, \qquad y \in \bigcup_{j=1}^k A_n^j.$$

This implies

$$Q\left(E \cap \left\{ Y'_n \in \bigcup_{j=1}^k A_n^j \right\}\right) = 0, \qquad n = 1, 2, \cdots.$$

Hence, $Q(E) = 0$.

THEOREM 6. *Every one-sided Markov channel is* AMS.

Proof. Let $[A_1^\infty, B_1^\infty, \nu]$ be a one-sided Markov channel. Let $\phi : A_1^\infty \to \mathcal{P}_1^\infty$ be the stationary map such that $\nu_x \in \mathcal{M}(\phi(x))$, $x \in A_1^\infty$. Let $[A_1^\infty, \mu]$ be an AMS source. Let $\bar{\mu}$ be the stationary measure on $\mathcal{A}_1^\infty$ such that $\bar{\mu}(E) = \mu(E)$ for every invariant event E in $\mathcal{A}_1^\infty$. Let $\phi' : A^\infty \to \mathcal{P}^\infty$ be the stationary map such that $\phi'(x)_i = \phi(x_i^\infty)_1$, $i \in Z$, $x \in A^\infty$. Note that

$$(6) \qquad \phi'(x)_1^\infty = \phi(x_1^\infty), \qquad x \in A^\infty.$$

Let $\bar{\mu}^*$ be the stationary measure on $\mathcal{A}^\infty$ such that

$$\bar{\mu}^*[(Y_1, \cdots) \in E] = \bar{\mu}(E), \qquad E \in \mathcal{A}_1^\infty.$$

Let $R' = \{x \in A^\infty : \phi'(x) \in R\}$. Since $\bar{\mu}^* \cdot (\phi')^{-1}$ is a stationary measure on $\mathcal{P}^\infty$, the set R has measure one under this measure. This implies $\bar{\mu}^*(R') = 1$. Let $[A^\infty, B^\infty, \hat{\nu}]$ be a stationary channel such that $\hat{\nu}_x = \bar{\nu}_{\phi'(x)}$, $x \in R'$. Then $\bar{\mu}^* \hat{\nu}$ is a stationary measure on $\mathcal{A}^\infty \times \mathcal{B}^\infty$. Let $(\bar{\mu}^* \hat{\nu})'$ be the stationary measure on $\mathcal{A}_1^\infty \times \mathcal{B}_1^\infty$ induced by $\bar{\mu}^* \hat{\nu}$. Let E be an invariant set in $\mathcal{A}_1^\infty \times \mathcal{B}_1^\infty$ such that $(\bar{\mu}^* \hat{\nu})'(E) = 0$. We will show $\mu\nu(E) = 0$.

(This will then imply $\mu\nu$ AMS, so we will be done.) We have

$$(\bar{\mu}^*\hat{\nu})'(E) = \int_{A^\infty} \hat{\nu}_x[(Y_1, Y_2, \cdots) \in E_{x_1^\infty}]\, d\bar{\mu}^*(x) = 0.$$

Hence, for $\bar{\mu}^*$-almost all $x \in R'$,

$$\bar{\nu}_{\phi'(x)}[(Y_1, \cdots) \in E_{x_1^\infty}] = 0.$$

Using Theorem 5 and the fact that $\bar{\mu}^*(R') = 1$, we have, for $\bar{\mu}^*$-almost all $x \in A^\infty$,

$$Q^y(E_{x_1^\infty} | \phi'(x)_1^\infty) = 0 \quad \text{for all } y \in B.$$

By (6), this reduces to the fact that for $\bar{\mu}$-almost all $x \in A_1^\infty$,

$$Q^y(E_x | \phi(x)) = 0, \qquad y \in B. \tag{7}$$

Now by (5) and the invariance of E, we have, for $n = 1, 2, \cdots$,

$$\begin{aligned} Q^y(E_x | \phi(x)) &= Q^y((Y'_{n+1}, \cdots) \in E_{x_{n+1}^\infty} | \phi(x)) \\ &\leqq \sum_{y \in B} Q^y((Y'_2, \cdots) \in E_{x_{n+1}^\infty} \phi(x_n^\infty)). \end{aligned}$$

Hence,

$$Q^y(E_x | \phi(x)) \leqq \sum_{y \in B} \left\{ N^{-1} \sum_{n=1}^{N} Q^y((Y'_2, \cdots) \in E_{x_{n+1}^\infty} | \phi(x_n^\infty)) \right\}. \tag{8}$$

But

$$Q^y((Y'_2, \cdots) \in E_{x_{n+1}^\infty} | \phi(x_n^\infty)) = 0$$

for $\bar{\mu}$-almost all $x \in A_1^\infty$, since

$$Q^y(E_x | \phi(x)) = Q^y((Y'_2, \cdots) \in E_{x_2^\infty} | \phi(x)) = 0$$

for $\bar{\mu}$-almost all x, by (7). Thus, the right-hand side of (8) converges to zero for each x in some invariant set F of $\bar{\mu}$-measure 1. Since $\mu(F) = \bar{\mu}(F)$, the convergence also occurs with μ-measure 1. This forces $Q^y(E_x | \phi(x)) = 0$ for μ-almost all $x \in A_1^\infty$. Since $\nu_x \in \mathcal{M}(\phi(x))$,

$$\begin{aligned} \nu_x(E_x) = \nu_x((Y'_2, \cdots) \in E_{x_2^\infty}) &\leqq \sum_{y \in B} Q^y((Y'_2, \cdots) \in E_{x_2^\infty} | \phi(x)) \\ &= \sum_{y \in B} Q^y(E_x | \phi(x)), \end{aligned}$$

and so $\nu_x(E_x) = 0$ for μ-almost all $x \in A_1^\infty$. Now

$$\mu\nu(E) = \int_{A_1^\infty} \nu_x(E_x)\, d\mu(x),$$

so $\mu\nu(E) = 0$.

THEOREM 7. *Every two-sided Markov channel is* AMS.

Proof. Let $[A^\infty, B^\infty, \nu]$ be a two-sided Markov channel. Let $\phi: A^\infty \to \mathcal{P}^\infty$ be the stationary map such that $\nu_x \in \mathcal{M}(\phi(x))$, $x \in A^\infty$. Let $R' = \{x \in A^\infty: \phi(x) \in R\}$. Let $[A^\infty, B^\infty, \hat{\nu}]$ be a stationary channel such that

$$\hat{\nu}_x = \bar{\nu}_{\phi(x)}, \qquad x \in R'.$$

Let $[A^\infty, \mu]$ be AMS. By Property 3 of AMS measures, let $\bar{\mu}$ be a stationary measure on

$\mathscr{A}^\infty$ such that μ is absolutely continuous with respect to $\bar{\mu}$. We will show that $\mu\nu$ is AMS by showing that $\mu\nu$ is absolutely continuous with respect to the stationary measure $\bar{\mu}\hat{\nu}$, and then again appealing to Property 3. Suppose $\bar{\mu}\hat{\nu}(E) = 0$. Then

$$\int_{A^\infty} \hat{\nu}_x(E_x)\, d\bar{\mu}(x) = 0.$$

This implies $\hat{\nu}_x(E_x) = 0$ for $\bar{\mu}$-almost all $x \in A^\infty$. If $x \in R'$, then $\hat{\nu}_x = k^{-1}(Q_1 + \cdots + Q_k)$, where $\{Q_1, \cdots, Q_k\}$ is the basis for $\mathscr{M}(\phi(x))$. Since ν_x is a convex combination of $Q_1, \cdots, Q_k$, we have that ν_x is absolutely continuous with respect to $\hat{\nu}_x$, $x \in R'$. Since $\bar{\mu}(R') = 1$, this gives $\nu_x(E_x) = 0$ for $\bar{\mu}$-almost all $x \in A^\infty$. But μ is absolutely continuous with respect to $\bar{\mu}$, so $\nu_x(E_x) = 0$ for μ-almost all x. Thus,

$$\mu\nu(E) = \int_{A^\infty} \nu_x(E_x)\, d\mu(x) = 0.$$

Remark. The foregoing proof is much easier than that given for the one-sided case, because for $P \in \mathscr{P}^\infty$ the sets $\mathscr{M}(P)$ have a simple structure. The sets $\mathscr{M}(P)$, for $P \in \mathscr{P}_1^\infty$, do not have this structure. Intuitively speaking, the two-sided channels are better behaved because the input source starts at time $-\infty$, and therefore if we look at the output at any finite time t, the output has had an infinite amount of time to "settle down" before reaching time t.

If $P \in \mathscr{P}$, recall that a nonempty subset $B' \subset B$ is a *closed* set of states for P if

$$\sum_{y \in B'} P(x, y) = 1, \qquad x \in B'.$$

P is *decomposable* if there exist two disjoint closed sets of states. Otherwise, P is *indecomposable.*

DEFINITION. Let $\phi: A_1^\infty \to \mathscr{P}_1^\infty$ be a stationary measurable map and $[A_1^\infty, B_1^\infty, \nu]$ a one-sided Markov channel such that $\nu_x \in \mathscr{M}(\phi(x))$, $x \in A_1^\infty$. We say the channel is *indecomposable* if, for every $x \in A_1^\infty$ and every positive integer n, the product $\phi(x)_1\phi(x)_2 \cdots \phi(x)_n$ is an indecomposable stochastic matrix. In similar fashion, one can define what it means for a two-sided Markov channel to be indecomposable.

THEOREM 8. *Indecomposable Markov channels are ergodic* AMS *channels.*

Proof. We prove this for a one-sided channel. (The proof for the two-sided case is similar and so is omitted.) Let $[A_1^\infty, B_1^\infty, \nu]$ be a one-sided Markov indecomposable channel. Let $\phi: A_1^\infty \to \mathscr{P}_1^\infty$ be the stationary map such that $\nu_x \in \mathscr{M}(\phi(x))$, $x \in A_1^\infty$. As in the proof of Theorem 6, $\phi': A^\infty \to \mathscr{P}^\infty$ is the map such that

$$\phi'(x)_i = \phi(x_i^\infty)_1, \qquad i \in Z, \quad x \in A^\infty.$$

Also, $R' = (\phi')^{-1}(R)$ and $[A^\infty, B^\infty, \hat{\nu}]$ is a stationary channel such that

$$\hat{\nu}_x = \bar{\nu}_{\phi'(x)}, \qquad x \in R'.$$

Let $[A^\infty, \mu]$ be stationary and ergodic. We show $\mu\hat{\nu}$ is ergodic. (At the end of the proof, we show that this implies that $[A_1^\infty, B_1^\infty, \nu]$ is an ergodic AMS channel.) For each $n \in Z$, let $\hat{X}_n$, $\hat{Y}_n$ be the maps from $A^\infty \times B^\infty$ to A, B respectively, such that

$$\hat{X}_n(x, y) = x_n, \qquad \hat{Y}_n(x, y) = y_n, \qquad (x, y) \in A^\infty \times B^\infty.$$

Let $\mathscr{F}_{-\infty}$ be the tail σ-field $\bigcap_{n<0} \mathscr{F}\{\hat{X}_i, \hat{Y}_i : i \geqq n\}$ of subsets of $\mathscr{A}^\infty \times \mathscr{B}^\infty$. $\mu\hat{\nu}$ will be ergodic if every invariant event in $\mathscr{F}_{-\infty}$ has measure zero or one. (Since $\mu\hat{\nu}$ is stationary, given any invariant event F in $\mathscr{A}^\infty \times \mathscr{B}^\infty$, there is an invariant event E in $\mathscr{F}_{-\infty}$ with $\mu\hat{\nu}(E \Delta F) = 0$; so we need only look in $\mathscr{F}_{-\infty}$ to determine ergodicity.)

Now if $x \in R'$ the dimension of $\mathcal{M}(\phi'(x))$ is 1. (Otherwise, by Theorem 2, there must exist positive integers $m < n$ such that $\phi(x)_m \phi(x)_{m+1} \cdots \phi(x)_n$ is decomposable, contradicting the indecomposability of the channel.) Hence, $\{\hat{\nu}_x\}$ is the basis for $\mathcal{M}(\phi'(x))$, $x \in R'$. Let $E \in \mathcal{F}_{-\infty}$ be invariant. Then

$$\mu\hat{\nu}(E) = \int_{A^\infty} \hat{\nu}_x(E_x)\, d\mu(x).$$

For each x,

$$E_x \in \mathcal{T} = \bigcap_{n<0} \mathcal{F}\{Y_i : i \leqq n\},$$

and so $\hat{\nu}_x(E_x) = 0$ or 1 for μ-almost all x, since $\hat{\nu}_x$ is $\mathcal{T}$-trivial for $x \in R'$ by Theorem 1, and $\mu(R') = 1$. Now the invariance of E and the stationarity of $[A^\infty, B^\infty, \hat{\nu}]$ imply that $\hat{\nu}_x(E_x) = \hat{\nu}_{Tx}(E_{Tx})$ for μ-almost all x. Hence, there must exist a real number α such that $\hat{\nu}_x(E_x) = \alpha$ for μ-almost all x, since μ is ergodic. Thus, either $\hat{\nu}_x(E_x) = 0$ for μ-almost x, or $\hat{\nu}_x(E_x) = 1$ for μ-almost all x. In the one case, this gives $\mu\hat{\nu}(E) = 0$ and in the other, $\mu\hat{\nu}(E) = 1$.

To complete the proof, we now let $[A_1^\infty, \mu]$ be an arbitrary one-sided ergodic AMS source. We show $\mu\nu$ is ergodic. Let $\bar{\mu}$ be the stationary ergodic measure on $\mathcal{A}_1^\infty$ such that $\bar{\mu}$ coincides with μ for invariant events. Let $\bar{\mu}^*$ and $(\bar{\mu}^*\hat{\nu})'$ be as defined in the proof of Theorem 6. Then, by the first part of this proof, $\bar{\mu}^*\hat{\nu}$ is ergodic and so $(\bar{\mu}^*\hat{\nu})'$ is ergodic. But, as observed in the proof of Theorem 6, $(\bar{\mu}^*\hat{\nu})'$ has the property that if E is invariant and $(\bar{\mu}^*\hat{\nu})'(E) = 0$, then $\mu\nu(E) = 0$. Hence, ergodicity of $(\bar{\mu}^*\hat{\nu})'$ implies that of $\mu\nu$.

Remark. Pfaffelhuber (1971) did part of what we did here for the case of a finite-state indecomposable channel $[A_1^\infty, B_1^\infty, \nu]$. He gives an explicit construction of a two-sided stationary channel $[A^\infty, B^\infty, \nu']$ which is roughly the same as our channel $[A^\infty, B^\infty, \hat{\nu}]$ in the sense that $\nu'_x = \hat{\nu}_x$, $x \in R'$. He shows that for a stationary ergodic source $[A^\infty, \mu]$, $\mu\nu'$ is ergodic. However, he does not show that the original channel $[A_1^\infty, B_1^\infty, \nu]$ is an ergodic AMS channel.

DEFINITION. A one-sided source $[A_1^\infty, \mu]$ is a *Markov source* if A is finite, if there exists a finite set S and a map $f: S \to A$, and if there is a homogeneous Markov chain $X_1, X_2, \cdots$ with state space S such that the joint distribution of $f(X_1), f(X_2), \cdots$ is μ. Similarly, one defines two-sided Markov sources.

The following could be deduced directly using properties of Markov chains, but it is interesting to note that it follows from Theorems 6 and 7.

THEOREM 9. *A Markov source is* AMS.

Proof for one-sided case. Let $[A_1^\infty, \mu]$ be a one-sided Markov source. Let $(X_1, X_2, \cdots)$ be the homogeneous S-valued Markov chain such that the distribution of $(f(X_1), f(X_2), \cdots)$ is μ. Let λ be the distribution of $(X_1, X_2, \cdots)$. Define a channel $[A_1^\infty, S_1^\infty, \nu]$ where $\nu_x = \lambda$ for all $x \in A_1^\infty$. This is a Markov channel, so if $[A_1^\infty, \sigma]$ is any stationary source, then $\sigma\nu$ is AMS. Hence, the distribution of the measure induced by $\sigma\nu$ on S_1^∞ is AMS. But this measure is λ. Thus, $(X_1, X_2, \cdots)$ has an AMS distribution. This implies $(f(X_1), f(X_2), \cdots)$ has an AMS distribution. Therefore, μ is AMS.

Application. We conclude the paper by presenting another application, this time to probability theory. We obtain an ergodic theorem for a random sequence of stochastic matrices.

THEOREM 10. *Let $X_1, X_2, \cdots$ be a sequence of $\mathcal{P}$-valued measurable functions defined on some probability space, whose joint distribution is* AMS. *Then the sequence*

$\{n^{-1}\sum_{i=1}^{n} X_1X_2\cdots X_i\}_{n=1}^{\infty}$ *converges almost surely as* $n\to\infty$. (We are using the obvious notion of convergence of matrices here; we say a sequence of $b\times b$ matrices $M_1, M_2, \cdots$ converges if $\lim_{n\to\infty} M_n(i,j)$ exists, $1\leqq i,j\leqq b$.)

Proof. Let $[\mathscr{P}_1^{\infty}, \mu]$ be the AMS source where μ is the joint distribution of $(X_1, X_2, \cdots)$. Fix $y_1\in B$. Let $[\mathscr{P}_1^{\infty}, B_1^{\infty}, \nu]$ be the Markov channel where

$$\nu_P = Q^{y_1}(\cdot\,|P), \qquad P\in\mathscr{P}_1^{\infty}.$$

By Theorem 6, $\mu\nu$ is AMS. Let $\tilde{Y}_1, \tilde{Y}_2, \cdots$ be the maps from $\mathscr{P}_1^{\infty}\times B_1^{\infty}\to B$ such that

$$\tilde{Y}_i(P, y) = y_i, \qquad i=1,2,\cdots, \quad (P, y)\in\mathscr{P}_1^{\infty}\times B_1^{\infty}.$$

If $y_2\in B$, we have that

$$\left\{n^{-1}\sum_{i=2}^{n+1} I_{\{\tilde{Y}_i=y_2\}}\right\},$$

converges almost surely with respect to $\mu\nu$ as $n\to\infty$. (See Property 4 of AMS measures.) Let $\tilde{X}_1, \tilde{X}_2, \cdots$, be the maps from $\mathscr{P}_1^{\infty}\times B_1^{\infty}\to\mathscr{P}$ such that $\tilde{X}_i(P, y) = P_i$, $i = 1, 2, \cdots$, $(P, y)\in\mathscr{P}_1^{\infty}\times B_1^{\infty}$. By a property of conditional expectation,

$$\left\{E_{\mu\nu}\left[n^{-1}\sum_{i=2}^{n+1} I_{\{\tilde{Y}_i=y_2\}}\middle|\tilde{X}_1, \tilde{X}_2, \cdots\right]\right\}$$

converges almost surely as $n\to\infty$, where $E_{\mu\nu}$ denotes expectation with respect to $\mu\nu$. It is easily verified that

$$E_{\mu\nu}\left[n^{-1}\sum_{i=2}^{n+1} I_{\{\tilde{Y}_i=y_2\}}\middle|\tilde{X}_1=P_1, \tilde{X}_2=P_2, \cdots\right] = n^{-1}\sum_{i=2}^{n+1}(P_1P_2\cdots P_{i-1})(y_1, y_2).$$

Translating this back in terms of the original sequence $X_1, X_2, \cdots$ we see that $n^{-1}\sum_{i=1}^{n}(X_1\cdots X_i)(y_1, y_2)$ converges almost surely for each $(y_1, y_2)\in B\times B$. Hence, $\{n^{-1}\sum_{i=1}^{n} X_1\cdots X_i\}$ converges almost surely.

Acknowledgment. The authors would like to thank Professor Robert M. Gray of Stanford University for suggesting the problem attacked here, and also for several helpful conversations regarding the content of this paper.

REFERENCES

D. BLACKWELL (1945), *Finite non-homogeneous chains*, Ann. Math., 46, pp. 594–599.

D. BLACKWELL, L. BREIMAN AND A. J. THOMASIAN (1958), *Proof of Shannon's Transmission Theorem for finite-state indecomposable channels*, Ann. Math. Statist., 29, pp. 1209–1220.

L. BREIMAN (1960), *Finite-state channels*, in Trans. of the Second Prague Conference on Information Theory, Prague, pp. 49–60.

R. M. GRAY AND J. C. KIEFFER (1980), *Asymptotically mean stationary measures*, Ann. Probab., to appear.

E. PFAFFELHUBER (1971), *Channels with asymptotically decreasing memory and anticipation*, IEEE Trans. Inform. Theory, 17, pp. 379–385.

C. E. SHANNON (1948), *A mathematical theory of communication*, Bell System Tech. J., 27, pp. 379–423, 623–656.

J. ZIV (1978), *Coding theorems for individual sequences*, IEEE Trans. Inform. Theory, 24, pp. 405–412.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0002 $01.00/0

HOW TO HOMOGENIZE A NONLINEAR DIFFUSION EQUATION: STEFAN'S PROBLEM*

ALAIN DAMLAMIAN†

Abstract. We study the homogenization of a Stefan problem (i.e., heat conduction with change of phase) when the structure is ε-periodic, and we prove that the constitutive laws of the limit medium do not depend upon the boundary conditions and are those of an anisotropically heat conducting medium which undergoes a change of phase at each temperature of change of phase of the original substances.

1. Introduction. For many physical studies of composite materials, one is more interested in the global or "macroscopic" behavior of a composite medium rather than in a detailed "microscopic" one. To put it in a different way, and for reasons that can also come from numerical analysis when discretization is considered, one is interested in finding the relevant properties (i.e., the constitutive laws or physical parameters) for an idealized homogeneous medium which would have the limit behavior of the composite material when the size of the periodic mesh goes to zero.

Finding the relevant parameters of this idealized limit (when it exists) is the origin and one of the main contributions of homogenization theory. Since this is not intended to be an introduction, let alone a survey, of homogenization theory, the reader is referred to Bensoussan-Lions-Papanicolaou [1] for a complete set of references.

The model problem we are looking at here is the problem of homogenization of a nonlinear heat equation for a composite material consisting of a periodic mixture of media which can undergo changes of phase, (Stefan's problem). One can also apply the present results to electromagnetic composite materials (see Bossavit-Damlamian [1]).

The plan is as follows:

2. The model ε-problem; weak formulation.
3. Estimates for the ε-problem.
4. A short review of elliptic homogenization.
5. The limit problem and its constitutive laws.

2. The model ε-problem. In short, the problem we are looking at is the following:

Let Ω be a given bounded domain in $\mathbb{R}^N$ (usually $N = 3$) with smooth boundary Γ. We restrict ourselves to two media M_1 and M_2. Their distribution in Ω is given according to a periodic structure of mesh size ε proportional to a basic period Y of size 1. The basic period Y is partitioned into two smooth subsets Y_1 and Y_2 corresponding to each medium M_1 and M_2, so that correspondingly Ω is partitioned into $\Omega_{1,\varepsilon}$ and $\Omega_{2,\varepsilon}$. The boundary between Y_1 and Y_2 is denoted by Σ, and its image in Ω is Σ_ε, which supposed to be rigid and perfectly heat conducting. As for $\partial\Omega = \Gamma$, it is split into $\Gamma_{1,\varepsilon}$ and $\Gamma_{2,\varepsilon}$, corresponding to each medium.

It is assumed, as customary, that the variations in volume are negligible for a Stefan problem.

Time will be restricted to an interval $[0, T]$ and it will be shown that the result is independent of T.

In $Q =]0, T[\times \Omega$, each change of phase for each medium will, in the strong formulation, generate a free boundary separating the phases. With our convention that

* Received by the editors July 22, 1980. This work was supported by the U.S. Army under contracts DAAG29-75-C-0024 and DAAG29-80-C-0041.

† Analyse Numérique et Fonctionnelle, C.N.R.S. et Université Paris-Sud, Bat. 425, 91405 Orsay Cedex and Centre de Mathematiques, Ecole Polytechnique 91128 Palaiseau Cedex, France.

there is at most one change of phase for each medium, two boundaries $S_{i,\varepsilon}(i=1,2)$ (each not necessarily connected) are generated.

Before writing the strong formulation, we introduce some notation. For each medium $i(=1,2)$, let $\alpha_i(v)$, θ_i, b_i, k_i denote the specific heat (a function of the temperature v), the temperature of change of phase, the latent heat and the heat conductivity (which is assumed to be strictly positive independent of the temperature v, a limitation for our final result but one which we have not been able to lift so far).

We can now write the strong formulation of the problem. One looks for $v^\varepsilon(t,x)$ (temperature) and the two surfaces $S_{i,\varepsilon}$ satisfying the following conditions.

$$\text{In } (0,T)\times\Omega_{i,\varepsilon}\backslash S_{i,\varepsilon},\qquad \alpha_i(v^\varepsilon)\frac{\partial v^\varepsilon}{\partial t}-\operatorname{div}(k_i\nabla v^\varepsilon)=f, \tag{2.1}$$

where f is an internal heating term.

$$\text{On } (0,T)\times\Sigma_\varepsilon,\quad \text{continuity of } v^\varepsilon,\quad \text{continuity of heat flux, namely: } k_1\nabla v^\varepsilon\cdot\vec{n}=k_2\nabla v^\varepsilon\cdot\vec{n}. \tag{2.2}$$

$$\text{On } S_{i,\varepsilon},\quad v^\varepsilon=\theta_i,\qquad b_i\cos(\vec{n},\vec{t})-\sum_j\left[h_i\frac{\partial v^\varepsilon}{\partial x_j}\cos(\vec{n},\vec{x}_j)\right]_{S_{i,\varepsilon}}=0, \tag{2.3}$$

where $\vec{n}$ is the unit normal to $S_{i,\varepsilon}$ in space/time and $[\cdot]_{S_{i,\varepsilon}}$ indicates the jump across $S_{i,\varepsilon}$ along $\vec{n}$. This is the classical Stefan condition on the free boundary.

Initial condition: $v^\varepsilon(0)$ given in Ω together with the initial boundaries $S_{i,\varepsilon}(0)$; they are assumed to be compatible ($v^\varepsilon(0)=\theta_i$ on $S_{i,\varepsilon}(0)$). (2.4)

As for the lateral boundary conditions, it is known in linear homogenization theory that, provided they are of variational form, they do not interfere with the limiting process. To be complete, we shall take them as linear inhomogeneous of mixed type. We assume a smooth partition of Γ in Γ^+ and Γ^- (Γ^- with nonempty interior) and require the following:

$$v^\varepsilon(t,x)=g^-(t,x)\quad\text{on } \Gamma^-, \tag{2.5}$$

$$k_i\frac{\partial v^\varepsilon}{\partial n}+Pv^\varepsilon=g^+(t,x)\quad\text{on } \Gamma^+\cap\Gamma_{i,\varepsilon}; \tag{2.6}$$

here, P is a nonnegative smooth function measuring the permeability of the boundary Γ^+ to heat flow, and g^- and g^+ are given smooth functions. It is also assumed that the boundary data g^-, g^+ agree with the initial data v_0^ε at $t=0$. It turns out that conditions (2.2) and (2.3) are Rankine-Hugoniot type conditions for the energy balance equation taken in the distribution sense on Q. In order to write this equation, (which gives a weak formulation) we need some notation:

For each i, let γ_i denote the maximal monotone graph defined (up to a constant) by

$$\gamma_i(v)=\int^v\alpha_i(s)\,ds+b_iH(v-\theta_i), \tag{2.7}$$

where H is the Heaviside function. This represents the enthalpy as a function of the temperature.

We also put:

$$\gamma(y, v) = \gamma_i(v) \quad \text{for } y \text{ in } Y_i, \tag{2.8}$$

$$k(y) = k_i \quad \text{for } y \text{ in } Y_i. \tag{2.9}$$

Then (2.1), (2.2), (2.3) reduce to

$$\frac{\partial u^\varepsilon}{\partial t} - \operatorname{div}\left(k\left(\frac{x}{\varepsilon}\right)\nabla v^\varepsilon\right) = f \quad \text{in } \mathscr{D}'(Q), \tag{2.10}$$

$$u^\varepsilon(t, x) \in \gamma\left(\frac{x}{\varepsilon}, v^\varepsilon(t, x)\right). \tag{2.11}$$

The initial conditions (2.4) can be expressed in terms of u alone as an initial condition

$$u^\varepsilon(0) = u_0, \tag{2.12}$$

which we can assume independent of ε.

For the lateral boundary conditions we introduce an auxiliary problem, where the time t is a mere parameter:

Let $g^\varepsilon(t, x)$ be the solution of

$$\begin{aligned} &-\operatorname{div}\left(k\left(\frac{x}{\varepsilon}\right)\nabla g^\varepsilon(t)\right) = 0 \quad \text{in } \Omega, \\ &g^\varepsilon(t) = g^-(t) \quad \text{on } \Gamma^-, \\ &k\left(\frac{x}{\varepsilon}\right)\frac{\partial g^\varepsilon(t)}{\partial n} + Pg^\varepsilon(t) = g^+(t) \quad \text{on } \Gamma^+. \end{aligned} \tag{2.13}$$

Clearly, g^ε is bounded in $H^1(\Omega)$ uniformly in ε. Now (2.1)–(2.6) has the weak formulation

$$\int_Q -\varphi' u^\varepsilon + \int_0^T a^\varepsilon(v^\varepsilon - g^\varepsilon, \varphi) = \int_Q f\varphi + \int_\Omega \varphi(0)u_0, \tag{2.14}$$

for all φ in $C^1(Q)$, $\varphi(T) = 0$ and $\varphi = 0$ on $(0, T)\times\Gamma^-$,

$$v^\varepsilon - g^\varepsilon = 0 \quad \text{on } (0, T)\times\Gamma^-. \tag{2.15}$$

In (2.14) a^ε is the bilinear Dirichlet form given by

$$a^\varepsilon(w, \varphi) = \int_\Omega k\left(\frac{x}{\varepsilon}\right)(\nabla w \cdot \nabla\varphi) + \int_{\Gamma^+} pw \cdot \varphi. \tag{2.16}$$

The smoothness assumptions made for Γ^-, Γ^+ allow for (2.14) to take φ in a larger class, namely, $\varphi \in W^{1,2}(0, T; V)$, where V is the variational space $\{\psi \in H^1(\Omega), \psi/\Gamma^- = 0\}$ (see Damlamian [1] for a detailed study).

DEFINITION 2.17. $(u^\varepsilon, v^\varepsilon)$ is a weak solution for problem (2.1)–(2.6) if and only if they satisfy (2.11), (2.14) and (2.15).

THEOREM 2.18 (see Damlamian [1], [2]). *Under the hypothesis that the α_i's are bounded above and below away from zero, there exists a unique solution $(u^\varepsilon, v^\varepsilon)$ for*

problems (2.11), (2.14) *and* (2.15) *which satisfies*

$$u^\varepsilon \in W^{1,2}(0, T; V^*) \cap L^\infty(0, T; L^2(\Omega)),$$
$$v^\varepsilon - g^\varepsilon \in W^{1,2}(0, T; L^2(\Omega)) \cap L^\infty(0, T; V).$$

Instead of giving a detailed proof, we will give in the following section the idea of how to obtain uniform estimates.

3. Uniform estimates.

PROPOSITION 3.1. *The solutions* $(u^\varepsilon, v^\varepsilon)$, $\varepsilon > 0$, *satisfy the following*:

u^ε *is bounded in* $W^{1,2}(0, T; V^*) \cap L^\infty(0, T; L^2(\Omega))$ *uniformly in* $\varepsilon > 0$,

$v^\varepsilon - g^\varepsilon$ *is bounded in* $W^{1,2}(0, T; L^2(\Omega)) \cap L^\infty(0, T; V)$ *uniformly in* $\varepsilon > 0$.

Proof. To obtain these estimates, it is enough to show them in the case of smooth γ. Then u^ε and v^ε are smooth enough to replace (2.14) by

$$\int_Q -\varphi' u^\varepsilon + \int_0^T a^\varepsilon(v^\varepsilon - g^\varepsilon, \varphi) = \int_Q f\varphi + \int_\Omega \varphi(0)u_0 - \int_\Omega \varphi(T)u^\varepsilon(T), \tag{3.2}$$

for all φ in $W^{1,2}(0, T; V)$. Then taking $\varphi = [A^\varepsilon]^{-1}u_\varepsilon$ (A^ε being the operator associated with a^ε on V) one gets

$$\frac{1}{2}\|u^\varepsilon(t)\|^2_{\varepsilon, V^*} + \int_0^t \int_\Omega v^\varepsilon u^\varepsilon \leqq \frac{1}{2}\|u_0\|_{\varepsilon, V^*} + \int_{(0,t)\times\Omega} u^\varepsilon g^\varepsilon + \int_0^t \|u^\varepsilon\|_{\varepsilon, V^*}\|f\|_{\varepsilon, V^*}, \tag{3.3}$$

where $\|\cdot\|_{\varepsilon, V^*}$ is the dual norm of $(a^\varepsilon(\varphi, \varphi)^{1/2}$ on V, the latter being uniformly equivalent to the standard norm on V.

Also taking $\varphi = [A^\varepsilon]^{-1}\, du^\varepsilon/dt$ one gets

$$c_1|v^\varepsilon(t)|^2_{L^2(\Omega)} + \int_0^t \left\|\frac{du^\varepsilon}{dt}\right\|^2_{\varepsilon, V} \leqq C_2 \qquad \text{(a constant which depends upon } f, v_0, g^+, g^- \cdots). \tag{3.4}$$

From (3.3) and (3.4) one gets (because $v^\varepsilon u^\varepsilon$ can be assumed nonnegative) that v^ε is bounded in $L^\infty(0, T; L^2(\Omega))$, u^ε in $W^{1,2}(0, T; V^*)$.

Then one takes $\varphi = (d(v^\varepsilon - g^\varepsilon)/dt$ to get

$$c_3 \int_0^t \left|\frac{dv^\varepsilon}{dt}\right|_{L^2(\Omega)} dt + a^\varepsilon(v^\varepsilon - g^\varepsilon, v^\varepsilon - g^\varepsilon) \leqq c_4 \qquad \text{(a constant which depends upon } f, g^+, g^-, v_0, \cdots). \tag{3.5}$$

From (3.5) one infers that v^ε stays bounded in $W^{1,2}(0, T; L^2(\Omega))$ and

$$v^\varepsilon - g^\varepsilon \text{ stays bounded in } L^\infty(0, T; V).$$

A detailed proof of the above can be found in Damlamian [1], [2], and in much simpler cases in Brezis [4] and Lions [1].

It is worth noticing that given the estimates of (3.1) (even when not uniform in ε), (2.14) can be replaced by (3.2) or even by

$$\int_0^t a^\varepsilon(v^\varepsilon - g^\varepsilon, \varphi) = \int_Q f\varphi + \int_\Omega \varphi(u_0^\varepsilon - u^\varepsilon(t)) \tag{3.6}$$

for all φ in V (independent of t). This remark (cf. Damlamian [1]) shows that $(u^\varepsilon, v^\varepsilon)$ is the solution of a simpler variational inequality (of the type studied by G. Duvaut [1]).

Another way of looking at (3.6) is the following, since the operator does not depend upon time:

$$\begin{aligned} a^\varepsilon(V^\varepsilon(t), \varphi) &= \int_\Omega (F(t) + u_0 - u^\varepsilon(t))\varphi, \\ V^\varepsilon(t) &= 0 \quad \text{on } \Gamma^-, \end{aligned} \tag{3.7}$$

for all $t \in (0, T)$, all φ in V, where

$$V^\varepsilon(t, x) = \int_0^t (v^\varepsilon(s, x) - g^\varepsilon(s, x))\, ds,$$

$$F(t, x) = \int_0^t f(s, x)\, ds.$$

4. A short review of elliptic homogenization. The purpose of this section is to show how elliptic homogenization works and how it can be applied in the present problem. See Bensoussan-Lions-Papanicolaou [1] (also L. Tartar [1]). With the same notation as above, we consider the operator $A^\varepsilon = -\text{div}\,(k(x/\varepsilon)\nabla)$ on Ω. Let w^ε be the variational solution of

$$\begin{aligned} A^\varepsilon w^\varepsilon &= f^\varepsilon \quad \text{in } \Omega, \\ w^\varepsilon &= g^- \quad \text{on } \Gamma^-, \\ k\left(\frac{x}{\varepsilon}\right)\frac{\partial w^\varepsilon}{\partial n} + P w^\varepsilon &= g^+ \quad \text{on } \Gamma^+, \end{aligned} \tag{4.1}$$

that is

$$a^\varepsilon(w^\varepsilon, \varphi) = \int_\Omega f^\varepsilon \varphi + \int_{\Gamma^+} g^+ \varphi \tag{4.2}$$

for all φ in V, $w^\varepsilon = g^-$ on Γ^-.

We assume that f^ε converges to f^0 weakly in $L^2(\Omega)$.

PROPOSITION 4.3. *As ε goes to zero, w^ε converges weakly in $H^1(\Omega)$ to the solution w^0 of the problem*

$$a^0(w^0, \varphi) = \int_\Omega f^0 \varphi + \int_{\Gamma^+} g^+ \varphi, \tag{4.4}$$

for all φ in V, $w^0 = g^-$ on Γ^-, where a^0 is the bilinear form given by

$$a^0(w, \varphi) = \int_\Omega \sum_{j,l} q_{j,l} \frac{\partial w}{\partial x_j} \frac{\partial \varphi}{\partial x_l} + \int_{\Gamma^+} P w \varphi,$$

with constant coefficients $q_{j,l}$ given by

$$q_{j,l} = \frac{1}{\text{mes}\,(Y)} \int_Y k(y)\nabla(\chi_\Gamma^l y_l)\nabla(\chi^j - y_j), \tag{4.5}$$

where χ^j is the solution (defined uniquely up to a constant) of

$$-\text{div}\,(k(y)\nabla\chi^j) = -\text{div}\,(k(y)e_j), \qquad \chi^j \text{ periodic in } Y; \tag{4.6}$$

e_j is the jth unit vector in $\mathbb{R}^N$, y_j being the coordinate on e_j.

Proof. It is clear that w^ε is bounded in $H^1(\Omega)$ (by coerciveness of a^ε with the Dirichlet boundary condition), so we can assume (via uniqueness of the solution for the

limit problem-to-be) that $w^\varepsilon \to w^0$. Then by a result of Tartar [1] one can see that $k(x/\varepsilon)\,\partial w/\partial x_l$ converges weakly in $L^2_{\text{loc}}(\Omega)$ (hence, in $L^2(\Omega)$) to $\sum_i q_{il}(\partial w^0/\partial x_i)$ (q_{il} given by (4.5), (4.6)), so that (4.2) goes to the limit in (4.4), which is the weak formulation of

$$\begin{aligned} A^0 w^0 &= f^0 \quad \text{in } \Omega, \\ w^0 &= g^- \quad \text{in } \Gamma^-, \\ \frac{\partial w^0}{\partial \gamma_{A^0}} + p w^0 &= g^+ \quad \text{on } \Gamma^+. \end{aligned} \tag{4.7}$$

Here, $\partial/\partial\gamma_{A^0}$ is the conormal derivative for A^0, which one should notice is not diagonal, but still symmetric, and with constant coefficients

$$A^0 = -\sum_{j,l} q_{jl} \frac{\partial^2}{\partial x_j\, \partial x_l}.$$

Here we have also used the compactness of the trace operator from $H^1(\Omega)$ into $L^2(\Gamma)$.

5. The limit problem and its constitutive laws. Making use of the results of § 4, one sees that $g^\varepsilon(t)$ converges weakly in $H^1(\Omega)$ to the solution $g^0(t)$ of

$$\begin{aligned} A^0 g^0(t) &= 0, \\ g^0(t) &= g^-(t) \quad \text{on } \Gamma^-, \\ \frac{\partial g^0(t)}{\partial \gamma_{A^0}} + P g^0(t) &= g^+ \quad \text{on } \Gamma^+. \end{aligned} \tag{5.1}$$

Also, by the estimates of (3.1) one can extract a sequence of values of ε [1] going to zero such that

$$\begin{aligned} u^\varepsilon &\rightharpoonup u^0 \quad \text{in } W^{1,2}(0, T; V^*) \cap L^\infty(0, T; L^2(\Omega)), \\ v^\varepsilon - g^\varepsilon &\rightharpoonup v^0 - g^0 \quad \text{in } W^{1,2}(0, T; L^2(\Omega)) \cap L^\infty(0, T; V). \end{aligned}$$

Hence, u converges strongly in $C([0, T]; V^*)$ and, for all $t \in [0, T]$, $u^\varepsilon(t)$ converges weakly to $u^0(t)$ in $L^2(\Omega)$.

Consequently, we can apply the result of (4.3) to (3.7) so that $V^\varepsilon(t)$, which obviously converges to $V^0(t) = \int_0^t (v^0(s) - g^0(s))\, ds$, satisfies

$$\begin{aligned} a^0(V^0(t), \varphi) &= \int_\Omega (F(t) + u_0 - u^0(t))\varphi, \\ V^0(t) &= 0 \quad \text{on } \Gamma^-. \end{aligned} \tag{5.2}$$

Using the equivalence with the weak formulation of type (2.14) we get

$$\begin{aligned} v^0 - g^0 &= 0 \quad \text{on } (0, T) \times \Gamma^-, \\ \int_Q -\varphi' u^0 + \int_0^t a^0(v^0 - g^0, \varphi) &= \int_Q f\varphi + \int_\Omega u_0 \varphi(0), \end{aligned} \tag{5.3}$$

for all φ in $W^{1,2}(0, T; V)$, $\varphi(T) = \varphi\big|_{(0, T)\times\Gamma^-} = 0$.

We now turn to (2.11), that is,

$$u^\varepsilon(t, x) \in \gamma\left(\frac{x}{\varepsilon}, v^\varepsilon(t, x)\right).$$

[1] The limit problem having a unique solution, it will be clear by the end of the proof that the whole sequence converges.

If ε is chosen so that g^ε converges to g^0 weakly in $H^1(\Omega)$, then v^ε converges to v^0 weakly in $W^{1,2}(0, T; L^2(\Omega) \cap L^\infty(0, T; H^1(\Omega))$, so that the convergence is uniform in $\mathscr{C}([0, T], L^2(\Omega))$, for example.

Let c be a real number, different from θ_i, and put $w_c^\varepsilon(x) = \gamma(x/\varepsilon, c)$. Clearly, w_c^ε converges weakly in L^2 to a constant:

$$\bar{\gamma}(c) = \frac{1}{\text{mes}\,(Y)} \int_Y \gamma(y, c)\, dy. \tag{5.4}$$

Using the monotonicity of $\gamma(x/\varepsilon, \cdot)$ we have

$$\mu(t, x) = (u^\varepsilon(t, x) - w_c^\varepsilon(x))(v^\varepsilon(t, x) - c) \geqq 0 \quad \text{a.e.}$$

Hence, using the proper convergences, we get that $\mu(t, x)$ converges weakly in the sense of measures on Ω for all t to $(u^0 - \bar{\gamma}(c))(v^0 - c)$ which has to be nonnegative. Hence, $u^0(t, x)$ belongs to the unique maximal extension of the monotone graph $\bar{\gamma}$, which we denote by $\tilde{\gamma}$. Consequently, (2.11) goes to

$$u^0(t, x) \in \tilde{\gamma}(v^0(t, x)) \quad \text{a.e. in } x \quad \text{for all } t. \tag{5.5}$$

6. Conclusion. We conclude that the limit equations correspond to the weak formulation of the following strong problem:

$$\begin{aligned} &\frac{du}{dt} + A^0 v = f, \\ &u(t, x) \in \tilde{\gamma}(v(t, x)), \\ &u(0, x) = u^0(\alpha), \\ &v(t, x) = g^-(t, x), \qquad x \in \Gamma^-, \\ &\frac{\partial v}{\partial \nu_{A^0}}(t, x) + pv(t, x) = g^+(t, x), \qquad x \in \Gamma^+, \end{aligned} \tag{6.1}$$

From the Stefan problem point of view, this is a nonisotropic Stefan problem.

Notice that we recover the heat diffusion operator of the linear case, that is, a homogeneous but anisotropic heat diffusion.

One also gets an "averaging" phenomenon for the graphs γ_i over Y; this is the only averaging consistent with the fact that both γ_i's are defined up to an additive constant and so is $\tilde{\gamma}$. Both temperatures of change of phases appear for discontinuities of $\tilde{\gamma}$, which is in agreement with daily experience (any other averaging of γ_1 and γ_2 would have yielded no discontinuity in the average, hence no change of phase). It is easy to see that the specific heat and latent heat of the limit medium are averages over Y of the corresponding terms.

Finally, on the theoretical side of things, it is of interest to realize that the isotropic diffusion laws are not stable under homogenization of Stefan problems, but anisotropic ones are stable.

It remains to prove that the above results can be extended to the case of temperature-dependent heat conductivity for each medium; this problem is more complicated but has been solved for the non-Stefan case (see Tartar [1]).

Acknowledgment. The author is very grateful to Professor J. M. Lasry for raising this interesting question, and to Professors L. Tartar and F. Murat for their helpful suggestions.

REFERENCES

A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU [1], *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

A. BOSSAVIT AND A. DAMLAMIAN [1], *Homogenization of the Stefan problem and application to magnetic composite media*, to appear.

H. BREZIS [1], *On some degenerate nonlinear parabolic equations*, in Nonlinear Functional Analysis, F. Browder, ed., Proc. Symp. in Pure Math., 18, American Mathematical Society, Providence, RI, 1970, pp. 28–38.

A. DAMLAMIAN [1] Thesis, University of Paris 6, 1976.

———, [2], *Some results on the multiphase Stefan problem*, Comm. Partial Differential Equations, 2 (1977), pp. 1017–1044.

G. DUVAUT [1], *Résolution d'un problème de Stefan*, CRAS Paris, 276 (1973), pp. 1461–1463.

J. L. LIONS [1], *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod-Gauthier Villars, Paris, 1969.

L. TARTAR [1], *Cours Peccot sur l'homogénéisation*, Collège de France, 1977, to appear. See also Cours Université Paris Sud ORSAY, 1980, to appear.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0003 $01.00/0

CONJUGATE AND FOCAL POINTS OF SECOND ORDER DIFFERENTIAL SYSTEMS*

E. C. TOMASTIK†

Abstract. Criteria are given to assure that b is not a conjugate or focal point to a for the second order nonlinear system $y''+g(t, y, y')=0$, where $|g(t, y, z)| \leqq k(t)|y|$. The results include the important linear case $y''+p(t)y=0$, where the matrix $p(t)$ is not assumed to be symmetric nor are the elements of $p(t)$ assumed to be of one sign. Nonoscillation and comparison theorems are also given. Most of the results seem to be new even in the scalar linear case.

1. Introduction. In this paper certain boundary value problems that are associated with conjugate and focal points of general systems of linear and nonlinear second order ordinary differential equations will be considered. The differential equation will be of the form

$$y''+g(t, y, y')=0, \tag{E}$$

where g and y are n-vectors. The boundary value problems considered are

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

$$y(a)=0=y'(b), \tag{2}$$

$$y'(a)=0=y(b). \tag{3}$$

For $i=1, 2, 3$, the equation (E) together with the boundary values (i) will be designated by (E_i).

Existence and uniqueness for the standard initial value problem is assumed for (E). It is also assumed that $g(t, y, z)$ is continuous on the set $\{(t, y, z): a \leqq t \leqq b\}$,

$$g(t, 0, 0)=0 \quad \text{for all } t \in [a, b], \tag{4}$$

$$|g(t, y, z)| \leqq k(t)|y| \quad \text{for all } t \in [a, b] \tag{5}$$

and all y, z where $k(t)$ is a continuous nonnegative scalar function, strictly positive at at least one point.

Certainly the most important special case will be the linear case

$$y''+p(t)y=0, \tag{6}$$

where $p(t)$ is a continuous $n \times n$ matrix. The matrix $p(t)$ will not be assumed to be symmetric, nor will the elements of $p(t)$ be assumed to be of one sign.

Conditions will be given on $k(t)$ and $p(t)$ to insure that (E_i) has no nontrivial solution. Nonoscillation and comparison theorems will also be given. Many of the results seem to be new even in the linear scalar case.

* Received by the editors December 3, 1979.
† Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268.

2. Conjugate and focal points. To begin, consider the following integral equations:

$$y(t)=(b-a)^{-1}\Big[(t-a)\int_t^b (b-s)g(s, y(s), y'(s))\,ds + (b-t)\int_a^t (s-a)g(s, y(s), y'(s))\,ds\Big], \tag{7}$$

$$y(t)=\int_a^t (s-a)g(s, y(s), y'(s))\,ds+(t-a)\int_t^b g(s, y(s), y'(s))\,ds, \tag{8}$$

$$y(t)=(b-t)\int_a^t g(s, y(s), y'(s))\,ds+\int_t^b (b-s)g(s, y(s), y'(s))\,ds. \tag{9}$$

The next lemma connects the differential equations (E_i) with the integral equations (7), (8), and (9).

LEMMA 1. *A function $y(t)$ continuous on $[a, b]$ satisfies* (E_1), (E_2), (E_3) *if and only if $y(t)$ satisfies* (7), (8), (9), *respectively.*

To establish the lemma, suppose that $y(t)$ satisfies (E_1). Then two integrations of (E_1), from a to t yield

$$y(t)=(t-a)y'(a)-\int_a^t (t-s)g(s, y(s), y'(s))\,ds.$$

Using the fact that $y(b)=0$ and solving for $y'(a)$ yields

$$y(t)=(b-a)^{-1}(t-a)\int_a^b (b-s)g(s, y(s), y'(s))\,ds-\int_a^t (t-s)g(s, y(s), y'(s))\,ds.$$

Now writing the first integral as $\int_a^t+\int_t^b$ and recombining yields (7). The proofs for (E_2) and (E_3) are similar and the converse is very easy.

Before continuing, it is necessary to make the following definitions. The scalar function $\varphi(t)$ will be termed an E_1 *admissible function on* $[a, b]$ if $\varphi(t)$ satisfies the following conditions:

(a) $\varphi(t)$ is continuous on $[a, b]$.

(b) $\varphi(t)>0$ on (a, b).

(c) If $\varphi(a)=0$, then near $t=a$, $\varphi(t)=(t-a)\varphi_a(t)$ where $\varphi_a(t)$ is continuous at and near $t=a$ and $\varphi_a(a)\neq 0$.

(d) If $\varphi(b)=0$, then near $t=b$, $\varphi(t)=(b-t)\varphi_b(t)$ where $\varphi_b(t)$ is continuous at and near $t=b$ and $\varphi_b(b)\neq 0$. The scalar function $\varphi(t)$ will be termed an $E_2(E_3)$ *admissible function on* $[a, b]$ if $\varphi(t)$ is an E_1 admissible function with $\varphi(b)>0$ ($\varphi(a)>0$).

LEMMA 2. *Let $i=1$, 2, or 3, and let the scalar function $\varphi(t)$ be an E_i admissible function on $[a, b]$. If $y(t)$ satisfies* (E_i), *then*

$$\sup\{|y(t)|/\varphi(t): t\in(a, b)\}$$

is finite.

Only the case $i=1$ and $\varphi(a)=0=\varphi(b)$ will be considered here, the other cases being easier.

To establish Lemma 2, notice that if y satisfies (E_1), then by well-known results $y(t)$ can be written as $y(t)=(t-a)y_a(t)$ near $t=a$ and $y(t)=(b-t)y_b(t)$ near $t=b$, where $y_a(t)$ is continuous at and near $t=a$ and $y_b(t)$ is continuous at and near $t=b$. Thus, $y(t)/\varphi(t)$ is continuous and, most importantly, bounded on (a, b). Thus, $\sup\{|y(t)|/\varphi(t): t\in(a, b)\}$ is a finite well-defined nonnegative number.

For any nonnegative continuous scalar function $\eta(t)$ on $[a, b]$, define the three maps T_k^i, i = 1, 2, 3 by:

$$(T_k^1\eta)(t) = (b-a)^{-1}\left[(t-a)\int_t^b (b-s)k(s)\eta(s)\,ds + (b-t)\int_a^t (s-a)k(s)\eta(s)\,ds\right],$$

$$(T_k^2\eta)(t) = \int_a^t (s-a)k(s)\eta(s)\,ds + (t-a)\int_t^b k(s)\eta(s)\,ds,$$

$$(T_k^3\eta)(t) = (b-t)\int_a^t k(s)\eta(s)\,ds + \int_t^b (b-s)k(s)\eta(s)\,ds.$$

Now notice that if φ is an E_i admissible function on $[a, b]$ for $i = 1, 2$ or 3, then $\varphi^{-1}(t)(T_k^1\varphi)(t)$ can be made continuous on $[a, b]$. This follows since it is easy to see that

$$\lim_{t=a} \varphi^{-1}(t)(T_k^i\varphi)(t) \quad \text{and} \quad \lim_{t=b} \varphi^{-1}(t)(T_k^i\varphi)(t)$$

exist and are finite. Thus the term $|T_k^i\varphi|_\varphi$ defined by

$$|T_k^i\varphi|_\varphi = \sup_{(a,\,b)} \varphi^{-1}(t)(T_k^i\varphi)(t)$$

is finite.

A basic theorem can now be established.

THEOREM 1. *Let $i = 1, 2$ or 3. If φ is an E_i admissible function on $[a, b]$ and*

$$|T_k^i\varphi|_\varphi < 1,$$

then (E_i) has no nontrivial solution.

To prove the theorem, let $i = 1$, 2 or 3, and let y be a solution of (E_i). Let $d = \sup\{|y(t)|/\varphi(t): t \in (a, b)\}$, a finite nonnegative number by Lemma 2. Then it follows readily from (7), (8), or (9) and (5) that

$$d \leqq |T_k^i\varphi|_\varphi d.$$

Since $|T_k^i\varphi|_\varphi < 1$, this last inequality implies that $d = 0$, which in turn implies that $y(t) = 0$ on $[a, b]$. Thus (E_i) has no nontrivial solution.

We continue with the following more specific theorems.

THEOREM 2. *If $\int_a^b k(s)(b-s)(s-a)\,ds < b-a$ or if $\int_a^t (s-a)^2(b-s)k(s)\,ds < (t-a)^2$ for all $t \in (a, b]$, then (E_1) has no nontrivial solution.*

To prove the first part of Theorem 2, take $\varphi(t) = t-a$ and let $f(t) = \varphi^{-1}(t)(T_k^1\varphi)(t)$ for $t \in (a, b)$ and $f(a) = \lim_{t=a} f(t)$ and $f(b) = \lim_{t=b} f(t)$. Then it is easy to see that $f(t)$ is continuous on $[a, b]$. Of course, $f(t) \geqq 0$ on $[a, b]$. It is easy to see that

$$f'(t) = -(t-a)^{-2}\int_a^t k(s)(s-a)^2\,ds,$$

which is less than zero if $t > a$, and therefore

$$\max f(t) = f(a) = (b-a)^{-1}\int_a^n (b-s)(s-a)k(s)\,ds,$$

which is less than one by hypothesis. Thus, $|T_k^1\varphi|_\varphi < 1$, and the first part of Theorem 2 follows from Theorem 1.

To prove the second part of Theorem 2, take $\varphi(t)=(t-a)(b-t)$ and let $f(t)=\varphi^{-1}(t)(T_k^1\varphi)(t)$. If $f(a)$ and $f(b)$ are defined as in the first part of the proof, $f(t)$ will be continuous and nonnegative on $[a, b]$. A computation yields

$$f'(t)=(t-a)^{-2}(b-t)^{-2}\Big[(t-a)^2\int_t^b (s-a)(b-s)^2k(s)\,ds - (b-t)^2\int_a^t (s-a)^2(b-s)k(s)\,ds\Big].$$

Notice that

$$\lim_{t=a} f'(t)=(b-a)^{-2}\int_a^b (s-a)(b-s)^2k(s)\,ds>0,$$

$$\lim_{t=b} f'(t)=-(b-a)^{-2}\int_a^b (s-a)^2(b-s)k(s)\,ds<0.$$

If $f'(t_0)=0$, then a computation yields

$$f(t_0)=(t_0-a)^{-2}\int_a^{t_0}(s-a)^2(b-s)k(s)\,ds,$$

which is less than one. Thus, $|T_k^1\varphi|_\varphi \leqq f(t_0)<1$ and the second part of Theorem 2 follows from Theorem 1.

The first part of Theorem 2 is due to Hartman [3, Chapt. XI, Thm. 5.1], in the linear scalar case (see also Reid [5], [6] for the linear self-adjoint case for systems).

It is interesting to compare the two parts of Theorem 2 in the trivial case $k(t)=1$. The first part readily yields $b^2<6$, but the second part yields $b^2<9$ which is a significant improvement. Later, it will be seen that calculating $\varphi^{-1}(t)(T_k^1\varphi)(t)$ directly when $\varphi(t)=(t-a)(b-t)$ and $k(t)=1$ yields $b^2<9.6$, which is very close to π^2.

THEOREM 3. *If* $\int_a^b k(s)(s-a)\,ds<1$ *or if* $\int_a^t k(s)(s-a)^2(2b-a-s)\,ds<(t-a)^2$ *for all* $t\in(a, b]$, *then* (E_2) *has no nontrivial solution.*

Proof. To prove the first part of Theorem 3 take $\varphi(t)=t-a$, and let $f(t)=\varphi^{-1}(t)(T_k^2\varphi)(t)$, and $f(a)=\lim f(t)$ and $f(b)=\lim f(t)$ as in the previous theorem. Then a computation yields

$$f'(t)=-(t-a)^{-2}\int_a^t k(s)(s-a)\,ds.$$

Thus,

$$|T_k^2\varphi|_\varphi=\max f(t)=f(a)=\int_a^b k(s)(s-a)\,ds<1,$$

and the first part of Theorem 3 follows from Theorem 1.

To prove the second part of Theorem 3 take $\varphi(t)=(t-a)(2b-a-t)$, and again let $f(t)$, $f(a)$, and $f(b)$ be defined as in the first part. Then

$$f'(t)=(t-a)^{-2}(2b-a-t)^{-2}\Big[-2(b-t)\int_a^t (s-a)^2(2b-a-s)k(s)\,ds + (t-a)^2\int_t^b k(s)(s-a)(2b-a-s)\,ds\Big],$$

and it follows readily that $\lim_{t=a} f'(t) > 0$. If $f'(t_0) = 0$, then

$$f(t_0) = (t_0 - a)^{-2} \int_a^{t_0} k(s)(s-a)^2(2b-a-s)\, ds < 1.$$

Thus, $|T_k^2\varphi|_\varphi < 1$ and the second part of Theorem 3 follows from Theorem 1.

THEOREM 4. *If* $\int_a^b k(s)(b-s)\, ds < 1$ *or if* $\int_a^t k(s)(b-s)(s-2a+b)\, ds < 2(t-a)$ *for all* $t \in (a, b]$, *then* (E_3) *has no nontrivial solution.*

The proof of Theorem 4 is similar to that of Theorem 3; take $\varphi(t) = b - t$ for the first part and $\varphi(t) = (b-t)(t-2a+b)$ for the second part.

THEOREM 5. *If* $k(t) < (\pi/(b-a))^2$ *on* $[a, b]$, *then* (E_1) *has no nontrivial solution.*

To prove this theorem, let the E_1 admissible function be $\varphi(t) = \sin \pi(b-a)^{-1}(t-a)$ and let M be such that $k(t) \leqq M < (\pi/(b-a))^2$ on $[a, b]$. Then

$$\varphi^{-1}(t)(T_k^1\varphi)(t) \leqq \varphi^{-1}(t)(T_M^1\varphi)(t) = M\varphi^{-1}(t)(T_1^1\varphi)(t) = M((b-a)/\pi)^2 < 1.$$

This implies that $|T_k^1\varphi|_\varphi < 1$ and thus from Theorem 1 that (E_1) has no nontrivial solution.

THEOREM 6. *If* $k(t) < (\pi/2(b-a))^2$ *on* $[a, b]$, *then neither* (E_2) *nor* (E_3) *has a nontrivial solution.*

The proof is similar to that of Theorem 5. Take the E_2 and E_3 admissible functions to be $\sin \pi(t-a)/2(b-a)$ and $\cos \pi(t-a)/2(b-a)$ respectively.

It will be seen later that Theorems 5 and 6 also follow as a corollary to a comparison theorem, Theorem 9.

3. Nonoscillation conditions on $[\alpha, \infty)$. For $i = 1$, 2 or 3, (E) is said to be E_i-nonoscillatory on $[\alpha, \infty)$ if, given any $b > a \geqq \alpha$, (E_i) has no nontrivial solution.

THEOREM 7. *If* $i = 1$ *or* 2 *and if* $\int_a^\infty sk(s)\, ds < \infty$, *then there exists* $\alpha \geqq a$, *such that* (E) *is* E_i*-nonoscillatory on* $[\alpha, \infty)$.

To prove the first part of the theorem, take the E_1 admissible function to be $\varphi(t) = 1$ and let $f(t) = (T_k^1\varphi)(t)$. Then of course $f(t) \geqq 0$, $f(a) = 0 = f(b)$. If $f'(t_0) = 0$, a computation shows that

$$\int_{t_0}^b (b-s)k(s)\, ds = \int_a^{t_0} (s-a)k(s)\, ds,$$

and

$$f(t_0) = \int_a^{t_0} (s-a)k(s)\, ds,$$

which for sufficiently large a will be less than one. Thus $|T_k^1\varphi|_\varphi < 1$, and Theorem 7 follows from Theorem 1.

To prove the second part of the theorem, notice that $\varphi(t) = 1$ is also an E_2 admissible function and that

$$(T_k^2 1)'(t) = \int_t^b k(s) \geqq 0.$$

Then

$$\sup_{(a,b)} (T_k^2 1)(t) = (T_k^2 1)(b) = \int_a^b (s-a)k(s)\, ds,$$

which for sufficiently large a will be less than one for any $b > a$. Thus $|T_k^2 1|_1 < 1$ and the second part of Theorem 7 follows from Theorem 1.

THEOREM 8. *If there exists* $\alpha \geqq a$ *such that* $\int_\alpha^t (t-s)k(s)\,ds < 1$ *for all* $t > \alpha$, *then* (E) *is* E_3*-nonoscillatory on* $[\alpha, \infty)$.

To prove the theorem take the E_3 admissible function to be $\varphi(t) = 1$. Then

$$(T_k^3 1)'(t) = -\int_a^t k(s)\,ds \leqq 0.$$

Then

$$|T_k^3 1|_1 = (T_k^3 1)(a) = \int_a^b (b-s)k(s)\,ds < 1,$$

and Theorem 8 follows from Theorem 1.

4. A comparison theorem.

THEOREM 9. *Suppose that* $P(t)$ *is a continuous nonnegative scalar function on* $[a, b]$. *Let* $i = 1, 2,$ *or* 3. *If* $\varphi(t)$ *is a (scalar) solution of the scalar equation* (E_i) *with* $g(t, y, y') = P(t)y$ *such that* $\varphi(t) > 0$ *on* (a, b), *and if* $|T_k^i \varphi|_\varphi < |T_P^i \varphi|_\varphi$ *then* (E_i) *has no nontrivial solution.*

To prove the theorem, notice that since φ is a solution to (E_i) with $g(t, y, y') = P(t)y$, $(T_P^i \varphi)(t) = \varphi(t)$. Thus

$$|T_k^i \varphi|_\varphi < |T_P^i \varphi|_\varphi = |\varphi|_\varphi = 1,$$

and the theorem follows from Theorem 1.

COROLLARY. *Let* $P(t)$ *be a nonnegative continuous scalar function on* $[a, b]$. *Let* $i = 1, 2$ *or* 3 *and suppose that* $\varphi(t)$ *is a scalar solution to* (E_i) *with* $g(t, y, y') = P(t)y$, *such that* $\varphi(t) > 0$ *on* (a, b). *If* $k(t) \leqq P(t)$ *on* $[a, b]$ *and there exists one point at which this inequality is a strict inequality, then* (E_i) *has no nontrivial solution.*

To prove the corollary for $i = 1$, first notice that

$$\varphi^{-1}(t)(T_k^1 \varphi)(t) < \varphi^{-1}(t)(T_P^1 \varphi)(t) \tag{10}$$

on (a, b). This inequality will now be established at both $t = a$ and $t = b$. Since the proofs at both these points are similar, only the proof for $t = a$ is given here. Toward this end recall that near $t = a$, $\varphi(t) = (t-a)\varphi_a(t)$ where $\varphi_a(t)$ is continuous and $\varphi_a(a) \neq 0$. It then readily follows that in the limit as t goes to a, the inequality (10) reduces to verifying that

$$\int_a^b (b-s)k(s)\varphi(s)\,ds < \int_a^b (b-s)P(s)\varphi(s)\,ds.$$

This is indeed true, and (10) holds on $[a, b]$. Thus $|T_k^1 \varphi|_\varphi < |T_P^1 \varphi|_\varphi$ and the corollary follows from Theorem 9. If $i = 2$ or 3, the proof is similar.

A stronger form of this corollary is given by Ahmad and Lazer [1] in the linear case where certain coefficients of the matrix $p(t)$ are of one sign, and $i = 1$.

5. A certain linear scalar case. This section will be concerned with the scalar equation

$$y'' + p(t)y = 0, \qquad y(0) = 0 = y(b), \tag{11}$$

where $0 \leqq p(t) \leqq t^n$. The results seem to be new.

THEOREM 10. *If* $0 \leqq p(t) \leqq t^n$ *on* $[0, b]$, *then* (11) *has no nontrivial solution for* $n = 1$ *if*

$$b^3 < 9^3(305 + 22\sqrt{22})^{-1}10 \cong 17.859,$$

for $n = 2$ if

$$b^4 < 30 \Big/ \Big[\sup_{0 \leqq x \leqq 1} (\tfrac{1}{2}(1+x+x^2+x^3) - x^4) \Big] \cong 28.002,$$

and for $n \geqq 3$ if

$$b^{n+2} < 2(n+2)(n+3)(2n+5)(3n+7)^{-1} = (n+2)(n+3)\frac{4+10/n}{3+7/n}.$$

To prove the theorem for the case $n = 1$ or 2, take the E_1 admissible function to be $\varphi(t) = t(b-t)$. After a tedious but straightforward computation, one obtains

$$\varphi^{-1}(t)(T_p^1\varphi)(t) \leqq (n+3)^{-1}(n+4)^{-1}\left[\frac{2}{n+2}\sum_{j=0}^{n+1} b^{n+2-j}t^j - t^{n+2}\right]$$

$$= \frac{b^{n+2}}{(n+3)(n+4)}\left[\frac{2}{n+2}(1+s+s^2+\cdots+s^{n+1}) - s^{n+2}\right], \tag{12}$$

if $t = bs$.

For $n = 2$, this implies the result stated in the theorem after using Theorem 1.

For $n = 1$, let the right side of (12) be $f(t)$. Then it is easy to see that $f(t)$ assumes its maximum at $t_0 = ((2+\sqrt{22})/9)b$ and $|T_p^1\varphi|_\varphi \leqq f(t_0) = \frac{1}{20}b^3(610 = +44\sqrt{22})9^{-3}$. Again the result follows from Theorem 1.

Now take the admissible function $\varphi(t)$ to be $\varphi(t) = t(b^{n+2} - t^{n+2})$. A tedious but easy computation yields

$$\varphi^{-1}(t)(T_p^1\varphi)(t) \leqq \frac{[(2n+4)(2n+5) - (n+2)(n+3)]b^{n+2} - (n+2)(n+3)t^{n+2}}{(n+2)(n+3)(2n+4)(2n+5)},$$

and thus

$$|T_p^1\varphi|_\varphi \leqq \frac{(2n+4)(2n+5) - (n+2)(n+3)}{(n+2)(n+3)(2n+4)(2n+5)}b^{n+2}, \tag{13}$$

which yields the result stated in the theorem in the case $n \geqq 3$, after applying Theorem 1 again.

For $n = 1$, (13) yields $b^3 < \frac{84}{5} = 16.8$ compared to approximately $b^3 < 17.8$ using (12). For $n = 2$, (13) yields $b^4 < \frac{360}{13} \cong 27.7$ compared to approximately $b^4 < 28$ using (12). However, for $n = 3$, (13) yields $b^5 < 41.25$, whereas (12) yields only about $b^5 < 40.0$. It appears that (12) yields $b^{n+2} < (n+3)(n+4)$ asymptotically, whereas it is obvious that (13) yields $b^{n+2} < (n+2)(n+3)\frac{4}{3}$. This result is an improvement over the result of Hartman [3], which yields in this case $b^{n+2} < (n+2)(n+3)$.

REFERENCES

[1] S. AHMAD AND A. LAZER, *An N-dimensional extension of the Sturm separation and comparison theory to a class of nonselfadjoint systems*, this Journal, 9 (1978), pp. 1137–1150.

[2] W. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics, 220, Springer-Verlag, Berlin and New York, 1971.

[3] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[4] M. LIAPOUNOFF, *Problème générale de la stabilité du mouvement*, Annals of Math. Stud. 17, Princeton University Press; Oxford University Press, London, 1947.

[5] W. REID, *Oscillation criteria for linear differential systems with complex coefficients*, Pacific J. Math., 6 (1956), pp. 733–751.

[6] ———, *A matrix Liapunov inequality*, J. Math. Anal. Appl., 32 (1970), pp. 424–434.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0004 $01.00/0

INTEGRATION OF INTERVAL FUNCTIONS*

OLE CAPRANI†, KAJ MADSEN‡ AND L. B. RALL§

Abstract. An interval function Y assigns an interval $Y(x) = (\underline{y}(x), \bar{y}(x)]$ in the extended real number system to each x in its interval $X = [a, b]$ of definition. The integral of Y over $[a, b]$ is taken to be the interval $\int_a^b Y(x)\,dx = [\int_a^b \underline{y}(x)\,dx, \int_a^b \bar{y}(x)\,dx]$, where $\int_a^b \underline{y}(x)\,dx$ is the lower Darboux integral of the lower endpoint function $\underline{y}$, and $\int_a^b \bar{y}(x)\,dx$ is the upper Darboux integral of the upper endpoint function $\bar{y}$. Since these Darboux integrals always exist in the extended real number system, it follows that all interval functions are integrable, no matter how nasty the endpoint functions $\underline{y}$, $\bar{y}$ are. The interval integral defined in this way includes the interval integral of R. E. Moore as the special case that $\underline{y}$, $\bar{y}$ are continuous, and hence Riemann integrable.

In addition to a construction of the interval integral in a form suitable for numerical approximation, some of its basic properties and other implications and applications of its definition are presented. The theory of interval integration given here supplies a previously lacking mathematical foundation for the numerical solution of integral equations by interval methods.

1. Intervals in the extended real number system.

In ordinary interval analysis [5], [6], the term *interval* refers to *closed* intervals of real numbers,

$$X = [a, b] = \{x \mid a \leqq x \leqq b\}, \tag{1.1}$$

with finite *endpoints* a, b. The *width*

$$w(X) = w([a, b]) = b - a, \tag{1.2}$$

of an interval with real endpoints is consequently finite. To develop the theory of integration of interval functions given below, it is convenient to use the *extended* real number system, which includes the values $\pm\infty$ [3]. Thus, in addition to *finite* intervals of the form (1.1) with a, b finite, there will be *infinite* intervals in the system of one of the following types:

(i) *semi-infinite* intervals

$$S_a = [a, +\infty], \qquad S^b = [-\infty, b], \qquad a, b \text{ finite}; \tag{1.3}$$

(ii) the *real line*

$$R = [-\infty, +\infty]; \tag{1.4}$$

and

(iii) the *indegenerate* intervals

$$S^{-\infty} = [-\infty, -\infty], \qquad S_{+\infty} = [+\infty, +\infty]. \tag{1.5}$$

(In what follows, "$+\infty$" will often be written simply as "∞".)

All the infinite intervals will be defined to be of *infinite width*, that is,

$$w(S_a) = w(S^b) = w(R) = w(S^{-\infty}) = w(S_{+\infty}) = +\infty, \tag{1.6}$$

* Received by the editors May 6, 1980.

† Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.

‡ Institute for Numerical Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark.

§ Mathematics Research Center, University of Wisconsin, Madison, Wisconsin. The work of this author was sponsored in part by the United States Army under contract DAAG29-80-C-0041 and the Danish Natural Science Research Council under grant 511-15849.

in the extended real number system. This definition is consistent with the type of limiting process used to define "improper" integrals, that is,

$$S_a = \lim_{b\to\infty} [a, b] = [a, \infty], \qquad S_\infty = \lim_{a\to\infty} [a, \infty] = [\infty, \infty], \tag{1.7}$$

and hence,

$$w(S_\infty) = \lim_{a\to\infty} w(S_a) = \lim_{a\to\infty} (\infty) = \infty, \tag{1.8}$$

so that it is reasonable in this sense to assign infinite widths to the indegenerate intervals $S^{-\infty}$ and S_∞.

In what follows, a closed interval in the extended real number system will be called simply an *interval.*

2. Interval arithmetic. Interval arithmetic [5], [6] as defined for finite intervals may also be performed in the system of intervals in the extended real number system defined in § 1 if suitable rules are adopted for computing with the values $\pm\infty$. In essence, these "rules" are a shorthand notation for the results of the types of limiting processes to be encountered in the theory of integration presented below. McShane [3, p. 21] gives the following rules:

$$\begin{aligned}
&\text{(i)} && -\infty < a < \infty \text{ for every real number } a;\\
&\text{(ii)} && \infty \cdot a = a \cdot \infty = \infty \text{ if } 0 < a \leqq \infty;\\
&\text{(iii)} && \infty \cdot a = a \cdot \infty = -\infty \text{ if } -\infty \leqq a < 0;\\
&\text{(iv)} && (-\infty) \cdot a = a \cdot (-\infty) = -\infty \text{ if } 0 < a \leqq \infty;\\
&\text{(v)} && (-\infty) \cdot a = a \cdot (-\infty) = \infty \text{ if } -\infty \leqq a < 0;\\
&\text{(vi)} && a/\infty = a/(-\infty) = 0 \text{ if } a \text{ is real};\\
&\text{(vii)} && \infty + a = a + \infty = \infty \text{ if } a > -\infty;\\
&\text{(viii)} && -\infty + a = a + (-\infty) = -\infty \text{ if } a < \infty;\\
&\text{(ix)} && \infty \cdot 0 = 0 \cdot \infty = (-\infty) \cdot 0 = 0 \cdot (-\infty) = 0.
\end{aligned} \tag{2.1}$$

Thus, rule (2.1ix) takes care of the "indeterminant" form "$0 \cdot \infty$" which can arise if one of the factors in a multiplication is an infinite interval. The product of two intervals will be *defined* to be

$$[a, b] \cdot [c, d] = [\min\{ac, ad, bc, bd\}, \max\{ac, ad, bc, bd\}] \tag{2.2}$$

in the extended real number system. In ordinary interval arithmetic [5, p. 9], (2.2) is a consequence of the definition $[a, b] \cdot [c, d] = \{z \mid z = x \cdot y, x \in [a, b], y \in [c, d]\}$ of multiplication of intervals. In the extended real number system, however, one has $\{z \mid z = x \cdot y, x \in [-1, 1], y \in [\infty, \infty]\} = \{-\infty, 0, \infty\}$ by (2.1ii, iii, ix), and the result is not an interval. Use of the rule (2.2) gives $[-1, 1] \cdot [\infty, \infty] = [-\infty, \infty]$, which circumvents this problem.

As in ordinary interval arithmetic, division by intervals containing 0 will not be defined. The *reciprocal* of an interval,

$$[c, d]^{-1} = \left[\frac{1}{d}, \frac{1}{c}\right], \qquad 0 \notin [c, d], \tag{2.3}$$

is defined for all zero-free intervals, with rule (2.1vi) used if $[c, d]$ is an infinite interval.

One has

$$S_a^{-1} = [a, \infty]^{-1} = \left[0, \frac{1}{a}\right], \quad a > 0; \qquad S_\infty^{-1} = [\infty, \infty]^{-1} = [0, 0];$$

$$(S^b)^{-1} = [-\infty, b]^{-1} = \left[\frac{1}{b}, 0\right], \quad b < 0; \qquad (S^{-\infty})^{-1} = [-\infty, -\infty]^{-1} = [0, 0]. \tag{2.4}$$

The indeterminant form "∞/∞" will thus not occur in the interval arithmetic under discussion, since division is defined by

$$\frac{[a, b]}{[c, d]} = [a, b] \cdot [c, d]^{-1}, \qquad 0 \notin [c, d], \tag{2.5}$$

and $[c, d]^{-1}$, if it exists, will have only finite endpoints by (2.3) and (2.4).

The indeterminant form "$\infty - \infty$" can appear in addition or subtraction according to the usual rules [5, pp. 8–9],

$$\begin{aligned} [a, b] + [c, d] &= [a + c, b + d], \\ [a, b] - [c, d] &= [a - d, b - c], \end{aligned} \tag{2.6}$$

but only if at least one of the terms is an indegenerate interval. Thus, an additional rule to augment the list (2.1) is needed, which is

$$\begin{aligned} [a, \infty] + [-\infty, -\infty] &= [a, \infty] - [\infty, \infty] = [-\infty, \infty], \\ [-\infty, b] + [\infty, \infty] &= [-\infty, b] - [-\infty, -\infty] = [-\infty, \infty], \end{aligned} \tag{2.7 (x)}$$

where a, b may be finite or infinite. Thus, rule (2.7x) assigns the value $+\infty$ to $\infty - \infty$ as an *upper* endpoint of an interval, and $-\infty$ as a *lower* endpoint.

Thus, the total collection of rules for interval arithmetic in the system of intervals defined over the extended real numbers consists of (2.1i–ix), (2.7x), (2.2), (2.3), (2.5) and (2.6). The interval arithmetic constructed in this way contains ordinary interval arithmetic on finite intervals [5], [6] in the sense that it gives the same results for finite intervals. The operations on infinite intervals are defined in such a way as to be convenient in the sequel for the construction of a theory of integration of interval functions. Other extensions of interval arithmetic to infinite intervals are possible, but will not be considered here.

3. Interval functions. Y is said to be an *interval function* of x on $[a, b]$ if it assigns a nonempty interval

$$Y(x) = [\underline{y}(x), \bar{y}(x)] = \{y \mid \underline{y}(x) \leqq y \leqq \bar{y}(x)\}, \tag{3.1}$$

to each $x \in [a, b]$. The (extended) real-valued functions $\underline{y}$, $\bar{y}$ are called the *endpoints* or *boundary functions* of Y, and the notation

$$Y = [\underline{y}, \bar{y}] \tag{3.2}$$

will be used, as well as the alternative notation

$$Y(x) = [\underline{y}, \bar{y}](x), \tag{3.3}$$

for the interval (3.1).

The interval function Y can also be identified with its graph, which is the set of points

$$Y = [a, b] \times Y(x) = \{(x, y) | x \in [a, b],\ y \in Y(x)\} \tag{3.4}$$

in the x, y-plane. Geometrically, the graph (3.4) extends from the "lines" $x = a$ on the left to $x = b$ on the right, and from the "curves" defined by $y = \underline{y}(x)$ below to $y = \bar{y}(x)$ above (recall that extended real values are permitted).

In the context of interval functions, a real-valued (or extended real-valued) function f is considered to be the *degenerate* interval function

$$f = [f, f]. \tag{3.5}$$

In the extended real number system, numbers $c \leqq d$ exist such that the graph (3.4) of Y is contained in the *rectangle* $R = [a, b] \times [c, d] = \{(x, y) | x \in [a, b],\ y \in [c, d]\}$ in the x, y-plane; that is,

$$Y \subset [a, b] \times [c, d] = R. \tag{3.6}$$

The set of all rectangles R for which (3.6) holds will be denoted by $R(Y)$ or by $R_{[a,b]}(Y)$ if it is desired to specify the *interval of definition* $[a, b]$ of Y.

If $[a, b]$ is a finite interval, then Y is said to be *finitely defined.* If (3.6) holds with c, d finite, then Y is called a *bounded* interval function. A bounded and finitely defined interval function is said to be finite; the graph of a finite interval function is obviously contained in a finite rectangle R with area $w([a, b]) \cdot w([c, d]) = (b - a) \cdot (d - c)$.

DEFINITION 3.1. For

$$c = \inf_{x \in [a,b]} \{\underline{y}(x)\}, \qquad d = \sup_{x \in [a,b]} \{\bar{y}(x)\}, \tag{3.7}$$

the interval

$$\nabla Y_{[a,b]} = [c, d] \tag{3.8}$$

is called the *vertical extent* of the interval function Y on $[a, b]$. If the interval of definition of Y is understood, then $\nabla Y_{[a,b]}$ may be abbreviated as ∇Y. The rectangle

$$R(\nabla Y) = R_{[a,b]}(\nabla Y) = [a, b] \times \nabla Y_{[a,b]} \tag{3.9}$$

is the "smallest" containing the graph of Y. One has

$$R(\nabla Y) = \bigcap_{R \in R(Y)} R; \tag{3.10}$$

that is, $R(\nabla Y)$ is the intersection of all rectangles (3.6) which contain the graph of Y.

Vertical extent of an interval function, as defined above, has the important property of being inclusion monotone with respect to the interval of definition of the interval function and inclusion of interval functions; $Y \subset Z$ means that the graph of Z contains the graph of Y considered to be point sets in the x, y-plane. More precisely, (3.7) and the definition (3.8) of vertical extent lead directly to the following result:

LEMMA 3.1. *If I, J are intervals on the x-axis with $I \subset J$, then*

$$\nabla Y_I \subset \nabla Y_J; \tag{3.11}$$

if Y, Z are interval functions on $X = [a, b]$ such that $Y \subset Z$, then

$$\nabla Y_{[a,b]} \subset \nabla Z_{[a,b]}. \tag{3.12}$$

4. Vertical measure and Darboux sums.

DEFINITION 4.1. The interval

$$W_{[a,b]}(Y) = w([a, b]) \cdot \nabla Y_{[a,b]} \tag{4.1}$$

is called the *vertical measure* of the interval function Y on $[a, b]$. Note that this quantity is interval-valued, and specifies the interval of definition of the interval function Y on which its vertical extent ∇Y is obtained.

(The goal in this paper is to construct a theory of Riemann-type integrals of interval functions. The *horizontal measure*

$$H_{[a,b]}(Y) = [a, b] \cdot w(\nabla Y_{[a,b]}) \tag{4.2}$$

of Y on $[a, b]$ may be useful in a Lebesgue-type integration theory, but this will not be pursued further here.)

Remark 4.1. Vertical measure is inclusion monotone with respect to inclusion of interval functions: If $Y \subset Z$, then $W_{[a,b]}(Y) \subset W_{[a,b]}(Z)$.

The assertion of Remark 4.1 follows immediately from Lemma 3.1.

As usual, a set of points $\{x_0, x_1, \cdots, x_n\}$ such that

$$a = x_0 \leqq x_1 \leqq \cdots \leqq x_{n-1} \leqq x_n = b \tag{4.3}$$

defines a *partition*,

$$\Delta_n = (X_1, X_2, \cdots, X_n), \tag{4.4}$$

of the interval $X = [a, b]$, where

$$X_i = [x_{i-1}, x_i], \qquad i = 1, 2, \cdots, n. \tag{4.5}$$

Obviously,

$$X = \bigcup_{i=1}^{n} X_i, \qquad w(X) = \sum_{i=1}^{n} w(X_i). \tag{4.6}$$

DEFINITION 4.2. The interval

$$\Sigma_{\Delta_n} Y = \sum_{i=1}^{n} w(X_i) \cdot \nabla Y_i = \sum_{i=1}^{n} W_{[x_{i-1}, x_i]}(Y) \tag{4.7}$$

is called the *Darboux sum* of the interval function Y corresponding to the partition Δ_n of $X = [a, b]$, where $\nabla Y_i = \nabla Y_{X_i} = \nabla Y_{[x_{i-1}, x_i]}$ has been written for brevity. For

$$\nabla Y_i = [c_i, d_i], \tag{4.8}$$

one has

$$c_i = \inf_{x \in X_i} \{\underline{y}(x)\}, \qquad d_i = \sup_{x \in X_i} \{\bar{y}(x)\} \tag{4.9}$$

and

$$\Sigma_{\Delta_n} Y = \left[\sum_{i=1}^{n} c_i \cdot w(X_i), \sum_{i=1}^{n} d_i \cdot w(X_i) \right]; \tag{4.10}$$

the endpoints of $\Sigma_{\Delta_n} Y$ are thus, respectively, the *lower Riemann sum* of the function y and the *upper Riemann sum* of the function $\bar{y}$ corresponding to the partition Δ_n of X [7].

The upper and lower limits of the interval (4.10) may also be interpreted as (elementary) integrals of *step-functions* [3], p. 54,

$$\sum_{i=1}^{n} c_i \cdot w(X_i) = \int_a^{'b} \underline{s}(x)\,dx = \int_X^{'} \underline{s}(x)\,dx \tag{4.11}$$

and

$$\sum_{i=1}^{n} d_i \cdot w(X_i) = \int_a^{'b} \bar{s}(x)\,dx = \int_X^{'} \bar{s}(x)\,dx. \tag{4.12}$$

In (4.11), the step function $\underline{s}(x)$ will have the values

$$\underline{s}(x) = c_i = \inf_{x \in X_i} \{\underline{y}(x)\}, \qquad x_{i-1} < x < x_i, \tag{4.13}$$

in all nondegenerate intervals X_i of the partition Δ_n. At each of the *partition points* x_i listed in (4.3), there will be a finite number of intervals X_{i-j_i}, $X_{i-j_i+1}, \cdots, X_i, X_{i+1}, \cdots, X_{i+k_i}$ which contain x_i. Define

$$\underline{s}(x_i) = \min \{c_j | x_i \in X_j\}, \qquad i = 0, 1, \cdots, n. \tag{4.14}$$

Similarly,

$$\bar{s}(x) = d_i = \sup_{x \in X_i} \{\bar{y}(x)\}, \qquad x_{i-1} < x < x_i \tag{4.15}$$

in nondegenerate intervals X_i of the partition Δ_n, and

$$\bar{s}(x_i) = \max \{d_j | x_i \in X_j\}, \qquad i = 0, 1, \cdots, n, \tag{4.16}$$

at the partition points $x_0, x_1, \cdots, x_n$. It follows that

$$\underline{s}(x) \leqq \underline{y}(x) \leqq \bar{y}(x) \leqq \bar{s}(x), \qquad a \leqq x \leqq b. \tag{4.17}$$

The properties of integrals of step-functions are well-documented [3, pp. 54–57]; for example, if s_1 and s_2 are step-functions on an interval X, and k is a finite constant, then

(4.18)

(a) $$\int_X^{'} k \cdot s_1(x)\,dx = k \cdot \int_X^{'} s_1(x)\,dx;$$

(b) $$\int_X^{'} (s_1(x) + s_2(x))\,dx = \int_X^{'} s_1(x)\,dx + \int_X^{'} s_2(x)\,dx;$$

(c) if $s_1(x) \leqq s_2(x)$ for all $x \in X$, then

$$\int_X^{'} s_1(x)\,dx \leqq \int_X^{'} s_2(x)\,dx.$$

Furthermore, if $s(x)$ is a step-function on X, then for each partition Δ_m of X

$$\sum_{j=1}^{m} \int_{X_j}^{'} s(x)\,dx = \int_X^{'} s(x)\,dx. \tag{4.19}$$

The integral of a step function is also invariant under translation [3, p. 57].

The above results may be used to prove corresponding assertions about the Darboux sums (4.7), taking into account the differences between real and interval arithmetic.

THEOREM 4.1. *If Y, Z are interval functions on $X=[a, b]$ and k is a constant, then*

$$\text{(a)} \quad \Sigma_{\Delta_n} k \cdot Y = k \cdot \Sigma_{\Delta_n} Y;$$

$$\text{(b)} \quad \Sigma_{\Delta_n}(Y+Z) \subset \Sigma_{\Delta_n} Y + \Sigma_{\Delta_n} Z; \tag{4.20}$$

(c) *if $Y \subset Z$ on X, then $\Sigma_{\Delta_n} Y \subset \Sigma_{\Delta_n} Z$ (inclusion monotonicity).*

Proof. For finite k, (4.20a) follows directly from (4.18a); rule (2.7x) allows one to drop the restriction of k to finite values. For $Y=[\underline{y}, \bar{y}]$, $Z=[\underline{z}, \bar{z}]$, the inequalities

$$\text{(a)} \quad \inf_{X_i}\{\underline{y}+\underline{z}\} \geqq \inf_{X_i}\{\underline{y}\} + \inf_{X_i}\{\underline{z}\},$$

$$\text{(b)} \quad \sup_{X_i}\{\bar{y}+\bar{z}\} \leqq \sup_{X_i}\{\bar{y}\} + \sup_{X_i}\{\bar{z}\} \tag{4.21}$$

on the intervals X_i, $i=1, 2, \cdots, n$, [3, p. 25] give

$$\nabla(Y+Z)_{X_i} \subset \nabla Y_{X_i} + \nabla Z_{X_i}, \tag{4.22}$$

from which

$$W_{X_i}(Y+Z) \subset W_{X_i}(Y) + W_{X_i}(Z), \tag{4.23}$$

$i=1, 2, \cdots, n$, and (4.20b) follows. Finally, the inclusion monotonocity of the vertical measure W (see Remark 4.1) with respect to inclusion of interval functions gives (4.20c). Q.E.D.

An analogue of (4.19) is also available immediately. For $m=2$, suppose that $a \leqq p \leqq b$, and that

$$\Delta_{n_1}^{(1)} = (X_{11}, X_{12}, \cdots, X_{1n_1}) \tag{4.24}$$

is a partition of $X_1=[a, p]$; similarly,

$$\Delta_{n_2}^{(2)} = (X_{21}, X_{22}, \cdots, X_{2n_2}) \tag{4.25}$$

is a partition of $X_2=[p, b]$. For $n=n_1+n_2$, it follows that

$$\Delta_n = (X_{11}, X_{12}, \cdots, X_{1n_1}, X_{21}, \cdots, X_{2n_2}), \tag{4.26}$$

will be a partition of $X=[a, b]$, and

$$\Sigma_{\Delta_n^{(1)}} Y + \Sigma_{\Delta_n^{(2)}} Y = \Sigma_{\Delta_n} Y. \tag{4.27}$$

This may be extended by induction to any positive integer $m>2$.

A type of mean value (or mean interval-value) theorem holds for the Darboux sums (4.7).

THEOREM 4.2. *If $X=[a, b]$ is a finite, nondegenerate interval, then an interval $\bar{Y}(\Delta_n) \subset \nabla Y_X$ exists for each partition Δ_n of X such that*

$$\Sigma_{\Delta_n} Y = w(X) \cdot \bar{Y}(\Delta_n). \tag{4.28}$$

Proof. By (4.7) and (4.20a),

$$\frac{1}{w(X)} \cdot \Sigma_{\Delta_n} Y = \sum_{i=1}^{n} \left(\frac{w(X_i)}{w(X)}\right) \cdot \nabla Y_i = [r, s], \tag{4.29}$$

say, where for $\alpha_i = w(X_i)/w(X)$, $i=1, 2, \cdots, n$,

$$r = \sum_{i=1}^{n} \alpha_i c_i, \qquad s = \sum_{i=1}^{n} \alpha_i d_i, \qquad \sum_{i=1}^{n} \alpha_i = 1. \tag{4.30}$$

Thus,

$$c = \min_{(i)} \{c_i\} \leq r \leq \max_{(i)} \{c_i\},$$
$$\min_{(i)} \{d_i\} \leqq s \leqq \max_{(i)} \{d_i\} = d, \tag{4.31}$$

and hence $[r, s] \subset \nabla Y_X$. Thus, by (4.29), (4.28) holds with $\bar{Y}(\Delta_n) = [r, s]$. Q.E.D.

5. Step-functions and Riemann sums. For each positive integer n, let S_n denote the set of all step-functions s_n on X having $n+1$ partition points $x_0, x_1, \cdots$, disposed according to (4.3). Furthermore, let

$$\underline{S}_n(Y) = \{\underline{s}_n | \underline{s}_n \in S_n, \underline{s}_n(x) \leqq \underline{y}(x), a \leqq x \leqq b\},$$
$$\bar{S}_n(Y) = \{\bar{s}_n | \bar{s}_n \in S_n, \bar{s}_n(x) \geqq \bar{y}(x), a \leqq x \leqq b\}. \tag{5.1}$$

The sets $\underline{S}_n(Y)$, $\bar{S}_n(Y)$ are nonempty, as $\underline{s}_n \equiv -\infty$ belongs to $\underline{S}_n(Y)$, and $\bar{\bar{s}}_n \equiv +\infty$ to The sets $\underline{S}_n(Y)$, $\bar{s}_n(Y)$ are nonempty, as $S_n \equiv -\infty$ belongs to $\underline{S}_n(Y)$ and $\bar{\bar{s}}_n \equiv +\infty$ to

$$\int_a^b \underline{s}_n(x)\, dx \leqq \int_a^b \bar{s}_n(x)\, dx, \tag{5.2}$$

and consequently

$$\bar{c}_n = \sup_{\underline{s}_n \in \underline{S}_n} \left\{ \int_a^b \underline{s}_n(x)\, dx \right\} \leqq \inf_{\bar{s}_n \in \bar{S}_n} \left\{ \int_a^b \bar{s}_n(x)\, dx \right\} = \underline{d}_n. \tag{5.3}$$

DEFINITION 5.1. For each positive integer n, let $\mathcal{D}_n$ denote the set of partitions (4.4). The interval

$$\Sigma_n Y = \bigcap_{\Delta_n \in \mathcal{D}_n} \Sigma_{\Delta_n} Y = [\bar{c}_n, \underline{d}_n], \tag{5.4}$$

is called the *Riemann sum of order n of the interval function Y over* $[a, b]$.

LEMMA 5.1. *The interval* $\Sigma_n Y$ *is nonempty; furthermore, if* $m > n$, *then*

$$\Sigma_m Y \subset \Sigma_n Y. \tag{5.5}$$

Proof. The assertion of the nonemptiness of the interval (5.4) is simply a restatement of (5.3). Denoting the set of Darboux sums (4.7) by $\mathcal{S}_n$, if $m > n$, then

$$\mathcal{S}_n \subset \mathcal{S}_m, \tag{5.6}$$

as if $\Sigma_{\Delta_n} \subset \mathcal{S}_n$; then one may take the partition Δ_m defined by

$$a = x_0 \leqq x_1 \leqq \cdots \leqq x_n = x_{n+1} = \cdots = x_m = b, \tag{5.7}$$

for which $\Sigma_{\Delta_n} Y = \Sigma_{\Delta_n} Y$, and thus $\Sigma_{\Delta_n} Y \in \mathcal{S}_m$ for $m > n$. The inclusion (5.5) then follows from (5.6) by the definition (5.4). Q.E.D.

The properties of Darboux sums listed in Theorem 4.1 survive the intersection (5.4) and thus become properties of Riemann sums, giving immediately the following result.

THEOREM 5.1. *If* Y, Z *are interval functions on* $X = [a, b]$ *and* k *is a constant, then*

(a) $\Sigma_n k \cdot Y = k \cdot \Sigma_n Y$;

(b) $\Sigma_n(Y \times Z) \subset \Sigma_n Y + \Sigma_n Z$; (5.8)

(c) *if* $Y \subset Z$ *on* X, *then* $\Sigma_n Y \subset \Sigma_n Z$ (*inclusion monotonicity*).

The additivity of Riemann sums with respect to the intervals over which they are defined will now be investigated. In order to be definite, the notation

$$\Sigma_n Y_{[a,b]} = \Sigma_n Y_X, \tag{5.9}$$

will be used to indicate the *interval of summation* $X = [a, b]$. Suppose that $a \leqq p \leqq b$ and $X_1 = [a, p]$, $X_2 = [p, b]$. The following results apply to the expression of the sum of an interval function Y over X in terms of its sums over X_1 and X_2.

LEMMA 5.2. *If* $X = [a, b]$ *is finite and nondegenerate, then*

$$W_{[a,p]}(Y) + W_{[p,b]}(Y) \subset W_{[a,b]}(Y), \tag{5.10}$$

that is,

$$w([a, p]) \cdot \nabla Y_{[a,p]} + w([p, b]) \cdot \nabla Y_{[p,b]} \subset w([a, b]) \cdot \nabla Y_{[a,b]}. \tag{5.11}$$

Proof. Let $\nabla Y_{[a,p]} = [c_1, d_1]$, $\nabla Y_{[p,b]} = [c_2, d_2]$, $\nabla Y_{[a,b]} = [c, d]$. Then, by the definition of vertical extent,

$$c = \min\{c_1, c_2\}, \qquad d = \max\{d_1, d_2\}. \tag{5.12}$$

For $\alpha = w([a, p])/w([a, b])$, one has $1 \geqq \alpha \geqq 0$ and $w([p, b])/w([a, b]) = 1 - \alpha \geqq 0$. Thus,

$$\begin{aligned} w([a, p]) \cdot \nabla Y_{[a,p]} + w([p, b]) \cdot \nabla Y_{[p,b]} \\ = w([a, b]) \cdot [\alpha c_1 + (1-\alpha)c_2, \alpha d_1 + (1-\alpha)d_2] \end{aligned} \tag{5.13}$$

and, as

$$c \leqq \alpha c_1 + (1-\alpha)c_2 \leqq \alpha d_1 + (1-\alpha)d_2 \leqq d, \tag{5.14}$$

by (5.12), (5.11) follows. Q.E.D.

THEOREM 5.2. *If* $X = [a, b]$ *is finite and nondegenerate, then for each positive integer* $n \geqq 2$

$$\Sigma_n Y_{[a,b]} \subset \bigcap_{n_1+n_2=n} \{\Sigma_{n_1} Y_{[a,p]} + \Sigma_{n_2} Y_{[p,b]}\} \subset \Sigma_{n-1} Y_{[a,b]}. \tag{5.15}$$

Proof. The set $\mathcal{S}_n$ of Darboux sums (4.7) may be decomposed into two disjoint subsets for each positive integer $n \geqq 2$: the set $\mathcal{S}_n^p$ of sums corresponding to partitions Δ_n^p which have p as a partition point, the set of which will be denoted by $\mathcal{D}_n^p$, and the complement of $\mathcal{S}_n^p$ relative to $\mathcal{S}_n$, $\mathcal{S}_n^{p'} = \mathcal{S}_n \backslash \mathcal{S}_n^p$, that is, the set of all Darboux sums corresponding to partitions Δ_n for which p is not a partition point. As $\mathcal{S}_n^p \subset \mathcal{S}_n$, one has

$$\Sigma_n Y_{[a,b]} \subset \bigcap_{\Delta_n^p \in \mathcal{D}_n^p} \Sigma_{\Delta_n^p} Y_{[a,b]}. \tag{5.16}$$

By (4.27), for $\Sigma_{\Delta_n^p} Y \in \mathcal{S}_n^p$ one can write

$$\Sigma_{\Delta_n} Y_{[a,b]} = \Sigma_{\Delta_{n_1}} Y_{[a,p]} + \Sigma_{\Delta_{n_2}} Y_{[p,b]}, \tag{5.17}$$

where $n_1 + n_2 = n$. Consequently, as

$$\bigcap_{\Delta_n^p \in \mathcal{D}_n^p} \Sigma_{\Delta_n} Y_{[a,b]} = \bigcap_{n_1+n_2=n} \{\Sigma_{n_1} Y_{[a,p]} + \Sigma_{n_2} Y_{[p,b]}\}, \tag{5.18}$$

the first inclusion of (5.15) follows.

Now, consider a partition Δ_{n-1} of $X = [a, b]$ for $n \geqq 2$, and let Δ_n^p denote the partition of X obtained by adding the point p to the set $\{x_0, x_1, \cdots, x_{n-1}\}$. Either $p = x_i$

for some i, $0 \leqq i \leqq n-1$, in which case

$$\Sigma_{\Delta_{n-1}} Y_{[a,b]} = \Sigma_{\Delta_n^p} Y_{[a,b]} = \Sigma_{\Delta_{n_1}} Y_{[a,p]} + \Sigma_{\Delta_{n_2}} Y_{[p,b]}, \tag{5.19}$$

$n_1 + n_2 = n$, or a nondegenerate interval $X_i = [x_{i-1}, x_i] \subset \Delta_{n-1}$ exists such that $x_{i-1} < p < x_i$. As

$$w([x_{i-1}, p]) \cdot \nabla Y_{[x_{i-1},p]} + w([p, x_i]) \cdot \nabla Y_{[p,x_i]} \subset w([x_{i-1}, x_i]) \cdot \nabla Y_{[x_{i-1},x_i]}, \tag{5.20}$$

by Lemma 5.2 one has

$$\Sigma_{\Delta_n^p} Y_{[a,b]} \subset \Sigma_{\Delta_{n-1}} Y_{[a,b]} \tag{5.21}$$

in this case. Thus, as $\{\Sigma_{\Delta_n^p} Y_{[a,b]}\} = \mathscr{S}_n^p$,

$$\bigcap_{\Delta_n^p \in \mathscr{D}_n^p} \Sigma_{\Delta_n^p} Y_{[a,b]} \subset \bigcap_{\Delta_{n-1} \in \mathscr{D}_{n-1}} \Sigma_{\Delta_{n-1}} Y_{[a,b]} = \Sigma_{n-1} Y_{[a,b]} \tag{5.22}$$

by (5.20) and (5.21), and the second inclusion in (5.15) now follows from (5.18). Q.E.D.

A mean interval-value theorem also holds for Riemann sums.

THEOREM 5.3. *If $X = [a, b]$ is finite and nondegenerate then, for each positive integer n, an interval $\bar{Y}_n \subset \nabla Y_X$ exists such that*

$$\Sigma_n Y = w(X) \cdot \bar{Y}_n. \tag{5.23}$$

Proof. As before, let $\mathscr{D}_n$ denote the set of all partitions Δ_n for each positive integer n. Then, by (4.28),

$$\Sigma_n Y = \bigcap_{\Delta_n \in \mathscr{D}_n} \Sigma_{\Delta_n} Y = w(X) \cdot \bigcap_{\Delta_n \in \mathscr{D}_n} \bar{Y}(\Delta_n), \tag{5.24}$$

so that (5.23) holds with

$$\bar{Y}_n = \bigcap_{\Delta_n \in \mathscr{D}_n} \bar{Y}(\Delta_n) \subset \nabla Y_X, \tag{5.25}$$

as each $\bar{Y}(\Delta_n) \subset \nabla Y_X$. Q.E.D.

6. Interval integrals.

DEFINITION 6.1. (The interval integral). If Y is an interval function defined on $X = [a, b]$, then the *interval integral of Y over $[a, b]$* is defined to be the interval

$$\int_a^b Y(x)\, dx = \int_X Y(x)\, dx = \bigcap_{n=1}^{\infty} \Sigma_n Y. \tag{6.1}$$

As usual, Y is said to be *integrable* over X if its interval integral (6.1) is defined.

Remark 6.1. By Lemma 5.1, the interval integral (6.1) is a nonempty closed interval, since it is the intersection of a (countable) collection of nested closed intervals.

Remark 6.2. An equivalent definition of the interval integral (6.1) is

$$\int_X Y(x)\, dx = \left[\int_X \underline{y}(x)\, dx, \int_X \bar{y}(x)\, dx\right], \tag{6.2}$$

where, for the sets of step-functions

$$\underline{S} = \bigcup_{n=1}^{\infty} \underline{S}_n, \qquad \bar{S} = \bigcup_{n=1}^{\infty} \bar{S}_n, \tag{6.3}$$

the *lower Darboux integral* of $\underline{y}$ over $X=[a, b]$ is defined to be [3, p. 57]

$$\underline{\int}_X \underline{y}(x)\,dx = \sup_{\underline{s}\in\underline{S}} \left\{ \int_X' \underline{s}(x)\,dx \right\}, \tag{6.4}$$

and similarly, the *upper Darboux integral* of $\bar{y}$ over $X=[a, b]$ is

$$\bar{\int}_X \bar{y}(x)\,dx = \inf_{\bar{s}\in\bar{S}} \left\{ \int_X' \bar{s}(x)\,dx \right\}. \tag{6.5}$$

The set $\underline{S}$ defined in (6.3) is the set of all step-functions bounded above by $\underline{y}$; similarly, $\bar{S}$ is the set of all step-functions which are greater than or equal to $\bar{y}$ at each point of X. As these sets are nonempty (recall the step-functions $\underline{s}\equiv-\infty$ and $\bar{s}\equiv+\infty$), the Darboux integrals (6.4) and (6.5) always exist, no matter how nasty the functions $\underline{y}$, $\bar{y}$ are from the standpoint of ordinary integration theory. This observation furnishes the following result.

THEOREM 6.1. (Theory of interval integration). *If Y is an interval function defined on $X=[a, b]$, then its interval integral* (6.1) *over $[a, b]$ exists.*

In other words, all interval functions are integrable (in the sense of interval integration). The simplicity of this theory is due to the fact that intervals are accepted as values of integrals, including the case that the integrand is degenerate (i.e., a single real function). The requirement that the integral of a real function be a real number rather than a possibly nondegenerate interval leads to numerous difficulties and correspondingly rich theories of integration (as elucidated in [3], for example), which constitute some of the most important chapters of real analysis. By the introduction of interval values for integrals, these difficulties are resolved, and the operation of integration is extended to all functions, interval or real. This is analogous to the way that the introduction of complex numbers extends the operation of root extraction to all numbers, complex or real. However, just as complex analysis does not supersede real analysis, it is to be expected that interval analysis will develop as a complementary, rather than a competitive, discipline to real analysis.

Some implications of the definitions of the interval integral given above, and some basic properties of interval integrals will now be investigated.

Remark 6.3. If $\underline{y}$, $\bar{y}$ are Riemann integrable on $[a, b]$, then

$$\int_a^b Y(x)\,dx = \left[(\mathrm{R})\int_a^b \underline{y}(x)\,dx,\ (\mathrm{R})\int_a^b \bar{y}(x)\,dx \right], \tag{6.6}$$

in terms of the Riemann integrals of the lower and upper endpoint functions.

This follows from (6.2) and the definition of a Riemann integrable function [3, p. 57] as one with equal upper and lower Darboux integrals; its Riemann integral is taken to be this common value, so that if y is a Riemann integrable function on $X=[a, b]$, then its Riemann integral is

$$(\mathrm{R})\int_a^b y(x)\,dx = \underline{\int}_X y(x)\,dx = \bar{\int}_X y(x)\,dx. \tag{6.7}$$

Remark 6.4. In case $\underline{y}$, $\bar{y}$ are continuous on $[a, b]$, then the construction of the interval integral of Y may be simplified by taking only the *equidistant partitions* $\bar{\Delta}_n$ defined by the points

$$x_k = a + \frac{k}{n}\cdot(b-a), \qquad k=0, 1, \cdots, n, \tag{6.8}$$

for each positive integer n, so that $w(X_k) = 1/n$, and hence, by (4.20a),

$$\Sigma_{\bar{\Delta}_n} Y = \frac{1}{n} \cdot \sum_{i=1}^{n} \nabla Y_i. \tag{6.9}$$

Here, the formation of the Riemann sums $\Sigma_n Y$ can be skipped, and the interval integral is given by

$$\int_a^b Y(x)\,dx = \bigcap_{n=1}^{\infty} \Sigma_{\bar{\Delta}_n} Y = \left[(\mathrm{R}) \int_a^b \underline{y}(x)\,dx, (\mathrm{R}) \int_a^b \bar{y}(x)\,dx \right] \tag{6.10}$$

[3, pp. 58–59], as continuous functions are Riemann integrable.

The interval integral (6.10) is the one proposed by R. E. Moore for continuous interval functions [5, Chapt. 8], [6, pp. 50–56], as the endpoint functions of a continuous interval function are necessarily continuous [5, p. 18], [6, p. 33]. Of course, even in the case $\underline{y}$, $\bar{y}$ are continuous, one may be able to find a partition Δ_n of $[a, b]$ other than $\bar{\Delta}_n$ such that $\Sigma_{\Delta_n} Y$ is properly contained in the Darboux sum (6.9), and hence provides a "more accurate" approximation to the interval integral than given by use of the equidistant partition. Some additional remarks about the numerical approximation of interval integrals will be made later.

Some basic properties of interval integrals come directly from the properties of the corresponding Riemann sums (5.4) which hold under the intersections in (6.1). Thus, from Theorem 5.1, one has the following result.

THEOREM 6.2. *If Y, Z are interval functions on $X = [a, b]$ and k is a constant, then*

$$\text{(a)} \quad \int_a^b k \cdot Y(x)\,dx = k \cdot \int_a^b Y(x)\,dx;$$

$$\text{(b)} \quad \int_a^b (Y(x) + Z(x))\,dx \subset \int_a^b Y(x)\,dx + \int_a^b Z(x)\,dx; \tag{6.11}$$

(c) *if $Y \subset Z$ on X, then*

$$\int_a^b Y(x)\,dx \subset \int_a^b Z(x)\,dx \qquad \text{(inclusion monotonicity).}$$

By taking intersections over all positive integers n of the expressions in (5.15), one gets immediately:

THEOREM 6.3. *If Y is an interval function defined on a finite, nondegenerate interval $X = [a, b]$, and p is such that $a \leqq p \leqq b$, then*

$$\int_a^p Y(x)\,dx + \int_p^b Y(x)\,dx = \int_a^b Y(x)\,dx. \tag{6.12}$$

Similarly, Theorem 5.3 furnishes the following mean interval-value theorem for interval integrals.

THEOREM 6.4. *If Y is defined on a finite, nondegenerate interval $X = [a, b]$, then an interval $\bar{Y} \subset \nabla Y_X$ exists such that*

$$\int_a^b Y(x)\,dx = w([a, b]) \cdot \bar{Y}. \tag{6.13}$$

Proof. Taking intersections of both sides of (5.23) over all positive integers n gives (6.13) with

$$\bar{Y} = \bigcap_{n=1}^{\infty} \bar{Y}_n \subset \nabla Y_X. \tag{6.14}$$

Q.E.D.

Theorem 6.4 is useful in connection with properties of indefinite integrals.

DEFINITION 6.2. The interval function

$$I_a Y(x) = \int_a^x Y(t)\, dt, \tag{6.15}$$

is called the *indefinite integral* of the interval function Y over $[a, x]$ for $x \geqq a$. ($I^b Y(x)$ is similarly defined over $[x, b]$ for $x \leqq b$.)

THEOREM 6.5. *If Y is a bounded interval function on $[a, b]$, then $I_a Y(x)$ is a continuous interval function at any $p \in [a, b]$.*

Proof. Suppose that $\nabla Y_{[a,b]} = [c, d]$ and take, as usual,

$$|\nabla Y_{[a,b]}| = \max\{|c|, |d|\}, \tag{6.16}$$

which is finite by hypothesis. For $a < p < b$, $a \leqq x \leqq p$, $\bar{Y}_{[x,p]} \subset \nabla Y_{[a,b]}$ exists such that

$$\int_a^p Y(x)\, dx = \int_a^x Y(t)\, dt + w([x, p]) \cdot \bar{Y}_{[x,p]}, \tag{6.17}$$

by Theorems 6.3 and 6.4; likewise, an interval $\bar{Y}_{[p,x]} \subset \nabla Y_{[a,b]}$ exists such that

$$\int_a^x Y(t)\, dt = \int_a^p Y(x)\, dx + w([x, p]) \cdot \bar{Y}_{[p,x]}, \tag{6.18}$$

for $p < x \leqq b$. Given any $\varepsilon > 0$, for $\delta = \varepsilon / |\nabla Y_{[a,b]}|$, the endpoints of $I_a Y(p)$ thus differ from the endpoints of $I_a Y(x)$ by less than ε for $|x - p| < \delta$. Continuity of $I_a Y(x)$ from the right at $x = a$ and from the left at $x = b$ is obtained from (6.18) and (6.17), respectively, as $I_a Y(a) = 0$ by Theorem 6.4. Q.E.D.

Indefinite integrals also exhibit a type of differentiability if the limits

$$I'_- Y(x) = \lim_{p \uparrow x} \frac{1}{w([p, x])} \cdot I_p Y(x) = \lim_{p \uparrow x} \bar{Y}_{[p,x]} \tag{6.19}$$

and

$$I'_+ Y(x) = \lim_{q \downarrow x} \frac{1}{w([x, q])} \cdot I^q Y(x) = \lim_{q \downarrow x} \bar{Y}_{[x,q]} \tag{6.20}$$

exist and are equal, where $\bar{Y}_{[p,x]}$ and $\bar{Y}_{[x,q]}$ are the intervals defined in Theorem 6.4 and x lies interior to the interval of definition $[a, b]$ of Y. (One-sided derivatives at $x = a$ and $x = b$ are defined by (6.20) and (6.19), respectively.)

DEFINITION 6.3. If the limits $I'_- Y(x)$ and $I'_+ Y(x)$ exist and are equal, then

$$I'_x Y = I'_- Y(x) = I'_+ Y(x) \tag{6.21}$$

is called the *derivative* of the indefinite integral of Y at x.

The following theorem gives a condition under which the derivative of an indefinite interval integral is equal to its integrand.

THEOREM 6.6. *If Y is a continuous interval function on $[a, b]$, then its indefinite integral is differentiable, and*

$$\begin{aligned} I'_x Y &= Y(x), \qquad a < x < b, \\ I'_+ Y(a) &= Y(a), \qquad I'_- Y(b) = Y(b). \end{aligned} \tag{6.22}$$

Proof. Let, for example, $\bar{Y}_{[p,x]}=[\underline{z}(p), \bar{z}(p)]$ for $p<x$. If the upper endpoint function $\bar{y}$ of Y is considered as an interval function, it follows from (4.31), (5.25), and (6.14),that $\bar{z} \in \nabla \bar{y}_{[p,x]}$. As $\bar{y}$ is continuous if Y is a continuous interval function,

$$\lim_{p \uparrow x} \nabla \bar{y}_{[p,x]} = \bar{y}(x) = \lim_{p \uparrow x} \bar{z}(p). \tag{6.23}$$

Similarly, $\lim_{p\uparrow x} \underline{z}(p) = \underline{y}(x)$, so that $I'_- Y(x)$ exists, and

$$I'_- Y(x) = Y(x), \qquad a < x \leqq b. \tag{6.24}$$

In the same way, one has

$$I'_+ Y(x) = Y(x), \qquad a \leqq x < b, \tag{6.25}$$

which establishes (6.22). Q.E.D.

7. Relationships between interval, Riemann, and Lebesgue integrals of real functions. Ordinarily, no distinction will be made between a real function y and the corresponding *degenerate* interval function $[y]=[y, y]$ having equal upper and lower endpoint functions. It is convenient, however, to distinguish between possible integrals of y over an interval $X=[a, b]$. The notation

$$\int_a^b y(x)\,dx, \qquad (\mathrm{L})\int_a^b y(x)\,dx, \qquad (\mathrm{R})\int_a^b y(x)\,dx \tag{7.1}$$

will be used to denote respectively the interval integral of y as a degenerate interval function (which integral always exists), the Lebesgue integral of y if y is Lebesgue integrable over $[a, b]$, and finally, the Riemann integral of y if it exists.

Remark 7.1. The integral of a degenerate interval function y is a degenerate interval, that is,

$$\int_a^b y(x)\,dx = [r, r], \tag{7.2}$$

if and only if the real function y is Riemann integrable over $[a, b]$, so that

$$r = (\mathrm{R})\int_a^b y(x)\,dx. \tag{7.3}$$

This follows directly from Remark 6.3 and the definition (6.7) of the Riemann integral.

Thus, one ordinarily expects an interval integration, even of a single function, to result in a nondegenerate interval. For example, if χ_ρ is the characteristic function of the rationals, that is,

$$\begin{aligned} &\chi_\rho(x)=1 \quad \text{for } x \text{ rational,} \\ &\chi_\rho(x)=0 \quad \text{for } x \text{ irrational,} \end{aligned} \tag{7.4}$$

then

$$\int_0^1 \chi_\rho(x)\,dx = [0, 1], \tag{7.5}$$

as is well known. On the other hand, some nondegenerate interval functions have

degenerate interval integrals. Consider the function Y defined by

$$Y(x)=\begin{cases} 0, & 0\leqq x<\frac{1}{3}, \\ [0,1], & x=\frac{1}{3}, \\ 1, & \frac{1}{3}<x<\frac{2}{3}, \\ [1,2], & x=\frac{2}{3}, \\ 2, & \frac{2}{3}<x\leqq 1; \end{cases} \tag{7.6}$$

i.e., Y is an *interval step-function*, which includes the "risers" as well as the "treads". For this function

$$\int_0^1 Y(x)\,dx=[1,1], \tag{7.7}$$

as the lower and upper boundary functions of Y have equal (Riemann) integrals.

Any interval function Y may be interpreted, of course, as a *set* of functions, that is,

$$Y=\{y\,|\,\underline{y}(x)\leqq y(x)\leqq \bar{y}(x),\ a\leqq x\leqq b\}. \tag{7.8}$$

If Y is degenerate, then the set (7.8) consists of only the single function $y=\underline{y}=\bar{y}$. Otherwise, Y will contain a number of functions, among which there may be subsets with certain distinguishing properties (continuity, differentiability, monotonicity, etc.). For the discussion of integration, the following subsets of functions will be singled out for special mention.

DEFINITION 7.1. If Y is an interval function on $[a, b]$, then the set of Lebesgue (Riemann) integrable functions $y\in Y$ will be called the *Lebesgue* (*Riemann*) *core* of Y, and will be denoted by $C_{\mathrm{L}}(Y)$ ($C_{\mathrm{R}}(Y)$).

One has $C_{\mathrm{R}}(Y)\subset C_{\mathrm{L}}(Y)$ always, but these sets may, of course, be empty. For example, if M is a subset of $[0, 1]$ which is not measurable in the sense of Lebesgue, then its characteristic function χ_M is a degenerate interval function with an empty Lebesgue (and hence Riemann) core. The characteristic function χ_P of the rationals considered earlier (see (7.4)) provides an example of a degenerate interval function with an empty Riemann core, but a nonempty Lebesgue core (the function χ_P itself).

DEFINITION 7.2. The *value* $v(C_{\mathrm{L}}(Y))$ ($v(C_{\mathrm{R}}(Y))$) of the Lebesgue (Riemann) core of Y on $[a, b]$ is defined by

$$\begin{aligned} v(C_{\mathrm{L}}(Y))&=\left\{r\,|\,r=(\mathrm{L})\int_a^b y(x)\,dx,\ y\in C_{\mathrm{L}}(Y)\right\}, \\ v(C_{\mathrm{R}}(Y))&=\left\{r\,|\,r=(\mathrm{R})\int_a^b y(x)\,dx,\ y\in C_{\mathrm{R}}(Y)\right\}, \end{aligned} \tag{7.9}$$

respectively, provided that the indicated cores of Y are nonempty.

Each set $v(C_{\mathrm{L}}(Y))$ and $v(C_{\mathrm{R}}(Y))$, when nonempty, is *convex*, that is, if one contains values r_1, r_2, with $r_1\leqq r_2$, then it contains the entire interval $[r_1, r_2]$. This is because if y_1 has integral r_1 and y_2 has integral r_2, then the functions $y_\theta=y_1+\theta(y_2-y_1)$ are all integrable for $0\leqq\theta\leqq 1$, and have integrals equal to $r_\theta=r_1+\theta(r_2-r_1)$, $0\leqq\theta\leqq 1$, which is just another expression for the interval $[r_1, r_2]$. As a matter of fact, the theory of Lebesgue integration [3] leads to the conclusion that

$$I_{\mathrm{L}}(Y)=v(C_{\mathrm{L}}(Y)), \tag{7.10}$$

if it exists, is a closed interval, which will be called the *Lebesgue subinterval* of the interval integral (6.1) of Y over $[a, b]$. The set $v(C_{\mathrm{R}}(Y))$, on the other hand, is not

necessarily closed. This is considered to be a defect of Riemann integration, and led to the construction of the theory of Lebesgue integration. However, as $v(C_R(Y))$ is convex, then its closure,

$$I_R(Y) = \overline{v(C_R(Y))}, \tag{7.11}$$

is a closed interval which, if it exists, will be called the *Riemann subinterval* of the interval integral of Y over $[a, b]$.

The purpose of the introduction of the intervals (7.10) and (7.11) is to provide some quantiative information about the Lebesgue and Riemann cores of an interval function Y which measures its "integrability" in a certain fashion. In the metric topology for intervals [5], [6], the *distance* between intervals $[a, b]$ and $[c, d]$ is defined to be

$$d([a, b], [c, d]) = \max\{|a - c|, |b - d|\}. \tag{7.12}$$

(In the extended real number system, rule (2.7x) is used to resolve any indeterminant forms entering into (7.12).)

DEFINITION 7.3. For

$$I(Y) = \int_a^b Y(x)\, dx, \tag{7.13}$$

if the Riemann core $C_R(Y)$ of Y is nonempty, then

$$\rho(Y) = d(I_R(Y), I(Y)) \tag{7.14}$$

is called the *Riemann gap* of the interval function Y on $[a, b]$; similarly, if $C_L(Y)$ is nonempty, then

$$\lambda(Y) = d(I_L(Y), I(Y)) \tag{7.15}$$

is called the *Lebesgue gap* of Y on $[a, b]$.

Remark 7.2. One has

$$\lambda(Y) \leqq \rho(Y), \tag{7.16}$$

in case both numbers are defined.

This follows from the inclusion $C_R(Y) \subset C_L(Y)$. If only one of the numbers $\lambda(Y)$, $\rho(Y)$ is defined, it will be $\lambda(Y)$ by the same token. For the example (7.4) of the degenerate interval function χ_ρ, one has $\lambda(\chi_\rho) = 1$, and $\rho(\chi_\rho)$ is not defined.

THEOREM 7.1. *If the endpoint functions* $\underline{y}$, $\bar{y}$ *are Riemann integrable over* $[a, b]$, *then* $\lambda(Y) = 0$; *if* $\lambda(Y) = 0$, *then* $\underline{y}$, $\bar{y}$ *are Lebesgue integrable, and*

$$\int_a^b Y(x)\, dx = \left[(\mathrm{L})\int_a^b \underline{y}(x)\, dx, (\mathrm{L})\int_a^b \bar{y}(x)\, dx\right]. \tag{7.17}$$

Proof. By Remark 6.3, the Riemann integrability of $\underline{y}$, $\bar{y}$ means that $\rho(Y) = 0$; hence, $\lambda(Y) = 0$ by (7.16). Conversely, if $\lambda(Y) = 0$, then the integral $I(Y)$ is finite, and bounded sequences $\{\underline{y}_n\}, \{\bar{y}_n\} \subset Y$ of Lebesgue integrable functions may be found which converge to $\underline{y}$ and $\bar{y}$, respectively. It follows [3, p. 81] that $\underline{y}$ and $\bar{y}$ are Lebesgue integrable on $[a, b]$ and, as

$$\lim_{n\to\infty} (\mathrm{L})\int_a^b \underline{y}_n(x)\, dx = \int_X \underline{y}(x)\, dx, \tag{7.18}$$

one has that

$$\int_X \underline{y}(x)\,dx = (\mathrm{L})\int_a^b \underline{y}(x)\,dx, \tag{7.19}$$

and similarly for $\bar{y}$, whence (7.17). Q.E.D.

Remark 7.3. If $\underline{y}$ and $\bar{y}$ are Lebesgue integrable on $[a, b]$, then

$$\lambda(Y) = \max\left\{(\mathrm{L})\int_a^b \underline{y}(x)\,dx - \int_X \underline{y}(x)\,dx,\ \int_X \bar{y}(x)\,dx - (\mathrm{L})\int_a^b \bar{y}(x)\,dx\right\}. \tag{7.20}$$

This is true because $\underline{y}$ is the "smallest" Lebesgue integrable function contained in the interval function Y, and $\bar{y}$ the "largest" in the sense that for each function $y \in C_{\mathrm{L}}(Y)$, one has $\underline{y}(x) \leqq y(x) \leqq \bar{y}(x)$, $a \leqq x \leqq b$. Thus,

$$v(C_{\mathrm{L}}(Y)) = [(\mathrm{L})\int_a^b \underline{y}(x)\,dx,\ (\mathrm{L})\int_a^b \bar{y}(x)\,dx], \tag{7.21}$$

from which (7.20) follows by (7.12).

8. Improper integrals. In ordinary integration theory, an integral

$$r = \int_a^b y(x)\,dx \tag{8.1}$$

is said to be *improper* if the interval of integration $[a, b]$ is infinite, or if its integrand is unbounded on $X = [a, b]$ in the sense that given any $M > 0$, there exists a nondegenerate subinterval X_M of X such that $|y(x)| \geqq M$ for $x \in X_M$. Supposing that y is unbounded on $X = [a, b]$ in the sense that given any $M > 0$ there exists a nonde-Riemann) on $[\alpha, b]$, one defines the *improper Riemann integral* of y over $[a, b]$ to be

$$(\mathrm{IR})\int_a^b y(x)\,dx = \lim_{\alpha \downarrow a} (\mathrm{R})\int_\alpha^b y(x)\,dx, \tag{8.2}$$

provided this limit exists (in the extended real number system; infinite values will be accepted here for improper integrals). Similarly, if y is Riemann integrable over $[a, b]$ for $b > a$ finite, then

$$(\mathrm{IR})\int_a^\infty y(x)\,dx = \lim_{b \to \infty} (\mathrm{R})\int_a^b y(x)\,dx \tag{8.3}$$

by definition, again if the indicated limit exists.

The definition of interval integrals given in § 6 yields values of certain improper Riemann integrals if the integrand y is interpreted to be the degenerate interval function $[y, y]$, for example,

$$\begin{aligned} &(\mathrm{a}) \quad \int_0^1 x^{-1/3}\,dx = [\tfrac{3}{2}, \infty], \\ &(\mathrm{b}) \quad \int_0^1 x^{-1}\,dx = [\infty, \infty], \\ &(\mathrm{c}) \quad \int_0^\infty (-e^{-x})\,dx = [-\infty, -1]. \end{aligned} \tag{8.4}$$

In the above, the value of the improper Riemann integral appears as the finite endpoint in each of the intervals (8.4a) and (8.4c). The indegenerate interval (8.4b) indicates correctly that the corresponding improper Riemann integral is *divergent.*

DEFINITION 8.1. An interval integral (6.1) is said to be *infinite* if its value is one of the indegenerate intervals $[-\infty, -\infty]$ or $[\infty, \infty]$, *indeterminant* if it is equal to $R = [-\infty, \infty]$, or *improper* if its value is a semi-infinite interval $[a, \infty]$ or $[-\infty, b]$; otherwise, it is said to be *finite.*

The relationship between improper interval and Riemann integrals will now be considered for the cases (8.2) and (8.3), as illustrated by (8.4a) and (8.4c), respectively.

Suppose that $y(x)$ is unbounded above at $x = a$. Thus, every Darboux sum (4.7) will contain a term of the form (after elimination of nondistinct partition points, if necessary)

$$w(X_1) \cdot \nabla y_1 = [w(X_1) \cdot c_1, \infty], \tag{8.5}$$

where $X_1 = [a, x_1]$ and

$$c_1 = \inf_{x \in X_1} \{y(x)\}. \tag{8.6}$$

The interval integral of y will hence be either improper or infinite. The following theorem is illustrated by (8.4a).

THEOREM 8.1. *Suppose that y is Riemann integrable over* $[\alpha, b]$ *for* $a < \alpha < b$, *and the indefinite interval integral* $I_a y(\alpha)$ *satisfies*

$$\lim_{\alpha \downarrow a} I_a y(\alpha) = \lim_{\alpha \downarrow a} \int_a^\alpha y(x)\, dx = [0, \infty]; \tag{8.7}$$

then the improper Riemann integral (8.2) *of y over* $[a, b]$ *exists, and*

$$\int_a^b y(x)\, dx = \left[(\mathrm{IR}) \int_a^b y(x)\, dx, \infty \right]. \tag{8.8}$$

Proof. One has

$$\int_a^b y(x)\, dx = \int_a^\alpha y(x)\, dx + \int_\alpha^b y(x)\, dx \tag{8.9}$$

by Theorem 6.3 and, by Remark 6.3,

$$\int_\alpha^b y(x)\, dx = \left[(\mathrm{R}) \int_\alpha^b y(x)\, dx, (\mathrm{R}) \int_\alpha^b y(x)\, dx \right] = (\mathrm{R}) \int_\alpha^b y(x)\, dx \tag{8.10}$$

as degenerate intervals may be identified with the corresponding real numbers. Taking the limit as $\alpha \downarrow a$ of both sides of (8.9) gives (8.8) Q.E.D.

In the case of integration over an infinite interval, say $[a, \infty]$, suppose, for example, that y is negative but that $y(x) \uparrow 0$ as $x \to \infty$, as in (8.4c). Then, each Darboux sum (4.7) will correspond to a partition Δ_n with x_{n-1} finite, $x_n = +\infty$, and as

$$\nabla y_n = [c, 0], \tag{8.11}$$

where $c = \inf_{x \in X_n} \{y(x)\} < 0$, $w(X_n) = w([x_{n-1}, \infty]) = \infty$, then each will contain a term equal to

$$w(X_n) \cdot \nabla y_n = [-\infty, 0], \tag{8.12}$$

by rules (2.1iii) and (2.1ix). The situation illustrated by the example (8.4c) is a case of the following result.

THEOREM 8.2. *Suppose that y is Riemann integrable over the finite interval* $[a, b]$ *for each* $b > a$, *and the indefinite interval integral* $I^{\infty}y(b)$ *satisfies*

$$\lim_{b\to\infty} I^{\infty}y(b) = [-\infty, 0]; \tag{8.13}$$

then, the improper Riemann integral (8.3) *of y over* $[a, \infty]$ *exists, and*

$$\int_a^{\infty} y(x)\,dx = \left[-\infty, (\mathrm{IR})\int_a^{\infty} y(x)\,dx\right]. \tag{8.14}$$

Proof. This follows exactly in the same way as Theorem 8.1 by writing

$$\int_a^{\infty} y(x)\,dx = \int_a^{b} y(x)\,dx + \int_b^{\infty} y(x)\,dx, \tag{8.15}$$

and noting that

$$\int_a^{b} y(x)\,dx = (\mathrm{R})\int_a^{b} y(x)\,dx \tag{8.16}$$

as a degenerate interval. Q.E.D.

Other cases of improper interval and Riemann integrals may be treated in a similar fashion..

9. Computational implications of the theory. One purpose of the theory of integration of interval functions developed above is to provide a theoretical framework for the investigation of the numerical solution of linear and nonlinear integral equations such as

$$u(x) = (\mathrm{R})\int_a^{b} g(x, t, u(x), u(t))\,dt, \tag{9.1}$$

by interval methods. One approach along these lines is to reformulate (9.1) as an interval equation,

$$U = T(U), \tag{9.2}$$

for an interval function U which contains the desired solution u of the integral (9.1). Under certain conditions, the operator T will be a contraction mapping [1], [2], and the iteration process

$$U_{n+1} = T(U_n), \qquad n = 0, 1, 2, \cdots \tag{9.3}$$

will converge to give a solution of (9.2). To implement this for the integral equation (9.1), one forms the interval functions $G_n = [\underline{g}_n, \bar{g}_n]$, $n = 0, 1, 2, \cdots$, where

$$\begin{aligned} \underline{g}_n(x, t) &= \inf\{g(x, t, U_n(x), U_n(t))\}, \\ \bar{g}_n(x, t) &= \sup\{g(x, t, U_n(x), U_n(t))\}, \end{aligned} \tag{9.4}$$

and then (9.3) becomes

$$U_{n+1}(x) = \int_a^{b} G_n(x, t, U_n(x), U_n(t))\,dt, \tag{9.5}$$

in terms of interval integration. Of course, if $\underline{g}_n(x, t)$ and $\bar{g}_n(x, t)$ are Riemann integrable in t, then the endpoint functions $\underline{u}_{n+1}$, $\bar{u}_{n+1}$ of U_{n+1} are obtained by Riemann integration. From a numerical standpoint, in this case approximations $u^*_{n+1} \leqq \underline{u}_{n+1}$, $u^{**}_{n+1} \geqq$

$\bar{u}_{n+1}$ may be obtained to prescribed accuracy by any one of a number of methods, including the use of Darboux sums as defined in § 4 [7], with higher order accuracy being obtainable from integration of Taylor polynomial approximations to the endpoint functions, or by other rules of numerical integration [4], [5], [6], [9], provided, of course, that the endpoint functions are smooth enough.

A particularly simple case occurs if g is *monotone* in the sense that

$$\begin{aligned} \underline{g}_n(x, t) &= g(x, t, \underline{u}_n(x), \underline{u}_n(t)), \\ \bar{g}_n(x, t) &= g(x, t, \bar{u}_n(x), \bar{u}_n(t)), \end{aligned} \tag{9.6}$$

that is, the endpoint functions of U_n transform into the endpoint functions of G_n, and if g further transforms Riemann integrable functions into Riemann integrable functions. Here, the iteration (9.5) can be carried out using only the endpoint functions if one starts with an interval $U_0 = [\underline{u}_0, \bar{u}_0]$ which has Riemann integrable endpoint functions. An example of this approach to the solution of a nonlinear integral equation was given by Rall [7], in which step-functions were used as endpoint functions (and T was approximated by a numerical operator S such that $T \subset S$). In many cases, continuous solutions u are sought for integral equations (9.1), which gives rise to the following concept.

DEFINITION 9.1. The *continuous core* $C_C(U)$ of an interval function U on $[a, b]$ is defined to be the set of continuous functions y contained in U, that is

$$C_C(U) = \{y \mid y \in U \cap C[a, b]\}. \tag{9.7}$$

Evidently, $C_C(U) \subset C_R(U)$, the Riemann core of U defined earlier.

If g is a continuous function of its arguments, and the interval operator T is such that the continuous function v defined by

$$v(x) = (\mathrm{R}) \int_a^b g(x, t, u(x), u(t))\, dt \tag{9.8}$$

belongs to $T(U)$ for $u \in C_C(U)$, then it follows that each continuous solution u of (9.1) will belong to $C_C(T(U))$ if it belongs to U and hence to $C_C(U)$. Thus, it is tempting to try to compute the sequence (9.3) using only $C_C(U_n)$, where U_0 is taken to have continuous endpoint functions. However, in general, the functions $\underline{g}_1(x, t)$ and $\bar{g}_1(x, t)$ obtained from (9.4) will be only *semi-continuous* if U_0 is replaced by $C_C(U_0)$, and these so-called L- and U-functions may not even be Riemann integrable [3]. The theory of interval integration developed in this paper resolves this difficulty by allowing computation with the interval functions U_n directly, regardless of the character of their endpoint functions.

Remark 9.1. If $u \in C_C(U_0)$ is a solution of (9.1), then for the sequence (9.3) constructed by the operations (9.4) and the interval integration (9.5), it follows from the condition (9.8) for continuous g that

$$u \in C_C(U_n), \qquad n = 0, 1, 2, \cdots; \tag{9.9}$$

furthermore, for

$$U = \bigcap_{n=1}^{\infty} U_n, \tag{9.10}$$

one has $u \in C_C(U)$.

Remark 9.2. In the favorable case that $U_{n+1} \subset U_n$, $n = 0, 1, 2, \cdots$, and

$$\lim_{n\to\infty} \sup_{[a,b]} \{w(U_n(x))\} = 0, \tag{9.11}$$

one has that $U = [u, u] = u$ defined by (9.10) satisfies the integral equation (9.1), since a degenerate interval integral of a degenerate interval function is necessarily a Riemann integral; furthermore, one has error bounds of the form

$$\underline{u}_n(x) \leqq u(x) \leqq \bar{u}_n(x), \qquad a \leqq x \leqq b, \tag{9.12}$$

for $n = 0, 1, 2, \cdots$.

Further applications of interval integration to the solution of integral equations will be investigated in subsequent papers.

REFERENCES

[1] OLE CAPRANI AND KAJ MADSEN, *Contraction mappings in interval analysis*, BIT, 15(1975), pp. 362–366.

[2] ———, *Mean value forms in interval analysis*, Computing, 25 (1980), pp. 147–154.

[3] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, NJ, 1944.

[4] R. E. MOORE, *The automatic analysis and control of error in digital computation based on the use of interval numbers*, in [8], pp. 61–130.

[5] ———, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[6] ———, *Methods and Applications of Interval Analysis*, SIAM Studies in Applied Mathematics 2, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[7] L. B. RALL, *Numerical integration and the solution of integral equations by the use of Riemann sums*, SIAM Rev., 7(1965), pp. 55–64.

[8] ——— ed., *Error in Digital Computation*, Vol. 1, John Wiley, New York, 1965.

[9] ———, *Applications of software for automatic differentiation in numerical computation*, Computing, Suppl. 2 (1980), pp. 141–156.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0005 $01.00/0

APPROXIMATION THEORY METHODS FOR LINEAR AND NONLINEAR DIFFERENTIAL EQUATIONS WITH DEVIATING ARGUMENTS*

M. S. HENRY† AND K. WIGGINS‡

Abstract. The authors develop a theory that unifies two existing theories devoted to finding approximate solutions to differential equations with deviating arguments. An algorithm emerges from this new theory, and it is compared computationally to the previously developed algorithms.

1. Introduction. In two recent papers, Allinger and Henry [1], Henry and Wiggins [8] have applied approximation theory techniques to construct approximate solutions to initial value problems with deviating arguments. The concepts discussed in these two papers are related to ideas discussed in a number of recent papers [2], [5], [6], [7], [9], [10], [11], [13], [16]. The methods of most of these latter references result in nonlinear best approximation problems; the methods of [1], [5], [7], [8], [16] lead to linear best approximation problems and thus are computationally feasible. Of the references identified thus far, only [1], [8] consider deviating arguments. The interested reader is referred to the bibliography of [1] for additional references involving differential equations with deviating arguments. In the present paper the authors compare the basically different approximation theory techniques in [1] and [8], and present a unifying theory.

2. The initial value problem. Consider the scalar value problem

$$\begin{aligned} &x^{(n)}(t)-\hat{f}(t, x(t), \cdots, x^{(n-1)}(t), x(h(t)), \cdots, x^{(n-1)}(h(t)))=0, \\ &x^{(i)}(0)=c_i, \qquad i=0, \cdots, n-1, \quad t \in J=[-\gamma, \tau], \end{aligned} \tag{1}$$

where γ and τ are nonnegative and $\gamma+\tau>0$. Initial value problem (1) is considered in both [1] and [8]. Reference [1] basically requires that (1) be a linear differential equation and that the deviating argument h in (1) satisfy $h(J) \subseteq J$. In the latter reference (1) is not required to be linear and $h(J)$ need not be contained in J. However, the techniques of [8] do not exploit any linear part that may occur in (1), and the results of [8] are local in nature.

A function x is a solution in (1) on J if the $x^{(i)}(t)$, $i=0, \cdots, n$ are defined and continuous for $t \in J \cup h(J)$ and if x satisfies (1) for $t \in J$. In (1) it is assumed that $\hat{f}: J \times R_{2n} \to R$ is continuous, and that $h \in C(J)$. Other hypotheses will be imposed later.

The initial value problem (1) may be rewritten as

$$\begin{aligned} &x^{(n)}(t)+\sum_{j=0}^{n-1} a_j(t)x^{(j)}(t)+\sum_{j=0}^{n-1} b_j(t)x^{(j)}(h(t)) \\ &\quad -f(t, x(t), \cdots, x^{(n-1)}(t), x(h(t)), \cdots, x^{(n-1)}(h(t)))=0, \\ &\quad x^{(i)}(0)=c_i, \qquad i=0, 1, \cdots, n-1, \quad t \in J, \end{aligned} \tag{2}$$

where $a_j, b_j \in C(J)$, and $f: J \times R_{2n} \to R$ is continuous. Thus the linear part of (1) is separated from the nonlinear part. For ease in notation the following designations are

* Received by the editors June 4, 1980.
† Department of Mathematics, Central Michigan University, Mount Pleasant, Michigan 48859.
‡ Department of Mathematics, Walla Walla College, College Place, Washington 99324.

employed:

$$A[x](t) = x^{(n)}(t) + \sum_{j=0}^{n-1} a_j(t)x^{(j)}(t), \tag{3}$$

$$B[X \circ h](t) = \sum_{j=0}^{n-1} b_j(t)x^{(j)}(h(t)) \tag{4}$$

and

$$N[X, h](t) = f(t, X(t), X(h(t))), \tag{5}$$

where

$$X(t) = (x(t), \dot{x}(t), \cdots, x^{(n-1)}(t))^T.$$

Then (2) becomes

$$\begin{aligned} &A[x](t) + B[X \circ h](t) = N[X, h](t), \\ &x^{(i)}(0) = c_i, \qquad i = 0, \cdots, n-1, \quad t \in J. \end{aligned} \tag{6}$$

The standard transformation

$$(x_1(t), \cdots, x_n(t))^T = (x(t), x(t), \cdots, x^{(n-1)}(t))^T = X(t)$$

converts (6) into

$$\begin{aligned} &\dot{X}(t) - \bar{A}(t)X(t) - \bar{B}(t)X(h(t)) = F(t, X(t), X(h(t))), \\ &X(0) = \Lambda, \qquad t \in J, \quad \text{where } \Lambda = (c_0, \cdots, c_{n-1})^T, \\ &F(t, X(t), X(h(t))) = [0, \cdots, 0, f(t, X(t), X(h(t)))]^T, \end{aligned} \tag{7}$$

$$\bar{A}(t) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0(t) & -a_1(t) & -a_2(t) & & -a_{n-2}(t) & -a_{n-1}(t) \end{pmatrix},$$

and where

$$\bar{B}(t) = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ -b_0(t) & -b_1(t) & & -b_{n-2}(t) & -b_{n-1}(t) \end{pmatrix}.$$

We note that [1, (2.2)] is a special case of (7), and that (7) clearly identifies the linear and nonlinear components of (6). In contrast, the companion matrix in [8, (1A)] is a constant matrix and the linear components are not identified.

The conditions on f in (2) imply that $F: J \times R_{2n} \to R_n$ may be assumed to be continuous. If $U = (u_1, u_2, \cdots, u_n)^T$ is an element of R_n, define

$$\|U\| = \max_{1 \leqq i \leqq n} (|u_i|);$$

if U is a continuous mapping from J to R_n, then

$$\|U\|_J = \max_J \|U(t)\|.$$

Now let $\bar{Z}(t)$, $t \in J$, be the fundamental matrix solution to the companion linear system to (7); that is, to

$$\dot{\bar{X}}(t) = \bar{A}(t)\bar{X}(t), \qquad \bar{X}(0) = I, \tag{8}$$

where I is the $n \times n$ identity matrix. If $\bar{Z}(t) = (z_{ij}(t))$, then

$$\|\bar{Z}\|_J = \max_J \max_{1 \leqq i \leqq n} \sum_{j=1}^{n} |z_{ij}(t)|.$$

Let $\bar{J}$ be a closed subinterval of J, and suppose that $0 \in \bar{J}$. Define $\Omega_{\bar{J}}$ to be the set of all continuously differentiable functions $U : \bar{J} \to R_n$ that satisfy $\|U(t) - \bar{Z}(t)\Lambda\| \leqq \theta|t|$ for some nonnegative constant θ and for all $t \in \bar{J}$. Then a standard argument establishes that the pair $(\Omega_{\bar{J}}, \rho)$ is a complete metric space, where

$$\rho(U, V) = \inf\{\theta : \|U(t) - V(t)\| \leqq \theta|t|, t \in \bar{J}\}.$$

We conclude the present section with a theorem that will be utilized throughout the remainder of the paper. The proof follows from [1, Thm. 1] and is omitted.

THEOREM 1. *Let $\bar{A}$ and $\bar{B}$ be the matrices in* (7), *and let $\bar{Z}$ be the fundamental solution to* (8). *Suppose that $h \in C(J)$, $h : \bar{J} \to \bar{J}$, and that* $\max_{\bar{J}}|h(t)| = m_{\bar{J}}$. *Assume* $\max(\|\bar{Z}\|_J, \|\bar{Z}^{-1}\|_J) = \alpha$, *and that $\|\bar{B}\|_J = \beta$. If $\alpha^2 \beta m_{\bar{J}} < 1$, then for fixed h the operator L_h defined by*

$$L_h[U](t) = \bar{Z}(t)\Lambda + \int_o^t \bar{Z}(t)\bar{Z}^{-1}(s)\bar{B}(s)U(h(s))\,ds, \qquad t \in \bar{J} \tag{9}$$

is a contraction operator on $\Omega_{\bar{J}}$.

3. Theory of approximation. Let Π_k denote the set of all scalar polynomials of degree k. Define

$$\begin{aligned} \mathscr{P}_k = \{p \in \Pi_k : p(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_{n-1}t^{n-1} + \beta_n t^n + \cdots + \beta_k t^k, \\ t \in J, \text{ where } \alpha_i = c_i/i!,\ i = 0, 1, \cdots, n-1, \text{ and } \beta_i \in R,\ i = n, n+1, \cdots, k\}. \end{aligned} \tag{10}$$

Thus $\mathscr{P}_k$ is the set of all scalar polynomials of degree at most k that satisfy the initial conditions in (6). Now let $\hat{\Pi}_k$ denote the set of functions $P : J \to R_n$ such that each component of P is an element of Π_k. Corresponding to (10), define

$$\hat{\mathscr{P}}_k = \{p \in \hat{\Pi}_k : P = [p, \dot{p}, \cdots, p^{(n-1)}]^T, p \in \mathscr{P}_k\}. \tag{11}$$

Hence for each $P \in \hat{\mathscr{P}}_k$, $P(0) = \Lambda$.

We are now in a position to pose the best approximation problem to be considered in the remainder of this paper. Let $\bar{J} \subseteq J$ be as described below (8), and suppose that $g : \bar{J} \to \bar{J}$ is continuous. For $x \in C^{n-1}(J)$ and for the F in (7), define

$$\hat{F}[X, g](t) = F(t, X(t), X(g(t))). \tag{12}$$

Now let P be an appropriate element of $\hat{\mathscr{P}}_k$ (to be specified later), and choose $V_b \in \hat{\mathscr{P}}_k$ satisfying

$$\begin{aligned} &\inf_{V \in \hat{\mathscr{P}}_k} \|\dot{V} - \bar{A}V - \bar{B}[V \circ g] - \hat{F}[P, g]\|_{\bar{J}} \\ &\quad = \|\dot{V}_b - \bar{A}V_b - \bar{B}[V_b \circ g] - \hat{F}[P, g]\|_{\bar{J}}. \end{aligned} \tag{13}$$

We note that V_b satisfies (13) if and only if the first component v_b of V_b satisfies

$$\begin{aligned}(14)\qquad &\inf_{v\in\mathscr{P}_k}\sup_{\bar{J}}|A[v](t)+B[V\circ g](t)-N[P,g](t)|\\ &\quad=\sup_{\bar{J}}|A[v_b](t)+B[V_b\circ g](t)-N[P,g](t)|.\end{aligned}$$

Although the best approximations V_b and v_b exist, uniqueness is not guaranteed.

Let

$$\{\phi_i(t)\}_{i=0}^{k}=\{A[\psi_i](t)+B[\Psi_i\circ g](t)\}_{i=0}^{k},$$

where

$$\psi_i(t)=t^i,\qquad i=0,\cdots,k,$$

and where

$$\Psi_i=[\psi_i,\psi_i',\cdots,\psi_i^{(n-1)}]^T.$$

Then the scalar minimization problem (14) can be expressed as

$$\begin{aligned}&\inf_{v\in\mathscr{P}_k}\sup_{\bar{J}}|A[v](t)+B[V\circ g]-N[P,g](t)|\\ (15)\qquad &=\inf_{\{\beta_i\}\in R_{k-n+1}}\sup_{\bar{J}}\left|\sum_{i=n}^{k}\beta_i\phi_i(t)+\sum_{i=0}^{n-1}\alpha_i\phi_i(t)-N[P,g](t)\right|\\ &=\inf_{\{\beta_i\}\in R_{k-n+1}}\sup_{\bar{J}}\left|\sum_{i=n}^{k}\beta_i\phi_i(t)-N^*[P,g](t)\right|,\end{aligned}$$

where $N^*[P,g]=N[P,g]-\sum_{i=0}^{n-1}\alpha_i\phi_i$. If $\{\phi_i\}_{i=n}^{k}$ satisfies the Haar condition [3, p. 74], then the best approximation problem (15) and consequently the minimization problems (13) and (14) result in unique best approximations.

A comparison of (14) with [1, (1.3) and (4.1)] reveals that (14) results in the *best approximate solution* (BAS) of degree k to [1, (4.1)] if $\bar{J}=J$ and if $N[X,g](t)\equiv r(t)$, that is, if N is independent of X and g. We also note that the best approximation problem (13) closely resembles the *substitute approximation solution* (SAS) minimization problem defined in [1, (5.3)]. The analysis in [1] requires that $\hat{F}$ in (12) involve only $X(g(t))$ and that the $a_i(t)\equiv 0$, $i=0,\cdots,n-1$, see [1, Thm. 5]. Finally (13) resembles [8, (14)]; however [8] requires that $\bar{A}$ be a constant matrix and no identification or utilization of the linear part of (6) or (7) occurs in [8]. Utilization of the linear part can be particularly important in the computations (see § 4). We will return to the above comparisons after further examination of (13) and (14). The remainder of the present section is devoted to establishing theoretical properties of the best approximations in (13) and (14) and to relating these best approximation problems to the initial value problem with deviating argument.

To this end we assume that h in (1) satisfies $h(0)=0$. This requirement and the condition that $h\in C(J)$ are standard assumptions in the theory of initial value problems with deviating arguments [12].

For any $\bar{J}\subseteq J$ define $l(\bar{J})$ to be the length of $\bar{J}$. Now construct a sequence of closed intervals $\{J_s\}_{s=1}^{\infty}$ as follows. Let $0\in J_s$, $J_{s+1}\subseteq J_s\subseteq J$, $h(J_{s+1})\subseteq J_s$, and require that $\lim_{s\to\infty}l(J_s)=0$. We know that $h(0)=0$ and $h\in C(J)$ insure that such a construction is possible. Define $g_s^*:J_s\to R$ by $g_s^*(t)=h(t)$, $t\in J_{s+1}$, $g_s^*(J_s)\subseteq J_s$, and $g_s^*\in C(J_s)$. Without loss of generality we can assume that

$$(16)\qquad \sup_{J_s}|g_s^*(t)|=m_s \text{ satisfies } \alpha^2\beta m_s=\mu_s<1$$

for all s, where α and β are defined in Theorem 1. Now construct a function $g_s \in C(J)$ satisfying $g_s(t) = g_s^*(t)$, $t \in J_s$, $g_s(J) \subseteq J$, and $\|g_s\|_J = \|g_s^*\|_{J_s}$.

If $m = \sup\{m_s\}$, then the construction of g_s^*, $s = 1, 2, \cdots$, and (16) insure that

$$\alpha^2 \beta m < 1. \tag{17}$$

Let

$$p_\Lambda(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_{n-1} t^{n-1}, \tag{18}$$

and define

$$P_\Lambda = [p_\Lambda, p_\Lambda, \cdots, p_\Lambda^{(n-1)}]^T. \tag{19}$$

Clearly $P_\Lambda(0) = \Lambda$ and hence $P_\Lambda \in \hat{\mathscr{P}}_k$. Let g be any element of $C(J)$ that satisfies

$$g(J) \subseteq J \quad \text{and} \quad \|g\|_J \leqq m. \tag{20}$$

We note that each g_s, $s = 1, 2, \cdots$ is such a g.

Let $K = \|P_\Lambda\|_J + 1$. Since $\hat{F}$ in (12) is a continuous function from $J \times R_{2n}$ into R_n, there exist $M(K)$ such that $\|\hat{F}(t, X, Y)\| \leqq M(K)$ whenever $\|X\| \leqq K$ and $\|Y\| \leqq K$. For any g satisfying (20) and $P \in \hat{\mathscr{P}}_k$, let

$$\dot{P}_\Lambda(t) - \bar{A}(t)P_\Lambda(t) - \bar{B}(t)[P_\Lambda \circ g](t) - \hat{F}[P, g](t) = \eta(P_\Lambda, P)(t). \tag{21}$$

For $\|P\| \leqq K$, (21) implies that

$$\|\eta(P_\Lambda, P)\|_J \leqq \hat{\eta}, \tag{22}$$

where

$$\hat{\eta} = \|\dot{P}_\Lambda\|_J + (\|\bar{A}\|_J + \|\bar{B}\|_J)\|P_\Lambda\|_J + M. \tag{23}$$

Now for any g satisfying (20) Theorem 1 implies that the operator L_g defined in (9) has a unique fixed point $Y \in \Omega_J$. Thus

$$Y(t) = \bar{Z}(t)\Lambda + \int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)\bar{B}(s)\,Y(g(s))\,ds. \tag{24}$$

Equality (21) implies that

$$\begin{aligned} P_\Lambda(t) = L_g[P_\Lambda](t) &+ \int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)\hat{F}[P, g](s)\,ds \\ &+ \int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)\eta(P_\Lambda, P)(s)\,ds. \end{aligned} \tag{25}$$

This equality implies that $P_\Lambda \in \Omega_J$. Consequently

$$\rho(P_\Lambda, Y) \leqq \rho(P_\Lambda, L_g[P_\Lambda]) + \rho(L_g[P_\Lambda], L_g[Y]). \tag{26}$$

Since L_g is a contraction operator on Ω_J, (26) implies that

$$\rho(P_\Lambda, Y) \leqq \frac{1}{1-\mu}\rho(P_\Lambda, L_g[P_\Lambda]), \quad \text{where } 0 < \mu < 1. \tag{27}$$

But (22) and (25) imply that $\|P_\Lambda(t) - L_g[P_\Lambda](t)\| \leqq \alpha^2(M + \hat{\eta})$.
Therefore by (27) we have that

$$\rho(P_\Lambda, Y) \leqq \frac{\alpha^2(M + \hat{\eta})}{1-\mu},$$

and hence

$$\|P_\Lambda(t) - Y(t)\| \leqq \frac{\alpha^2(M+\hat{\eta})}{1-\mu}|t|, \qquad t \in J. \tag{28}$$

We now select $J_{\bar{s}}$ to insure that

$$\frac{2\alpha^2(M+\hat{\eta})}{1-\eta}|t| \leqq 1, \qquad t \in J_{\bar{s}}. \tag{29}$$

We note that this is possible because $\lim_{s\to\infty} l(J_s) = 0$. Hereafter

$$g = g_{\bar{s}} \quad \text{and} \quad \bar{J} = J_{\bar{s}}; \tag{30}$$

with this designation for g we observe that $g(\bar{J}) \subseteq \bar{J}$ and that $g(J) \subseteq J$. Also $g = h$ on $J_{\bar{s}+1}$.

Now define

$$S_k = \left\{ P \in \hat{\mathscr{P}}_k : \|P(t) - P_\Lambda(t)\| \leqq \frac{2\alpha^2(M+\hat{\eta})}{1-\mu}|t|,\ t \in \bar{J} \right\}. \tag{31}$$

Clearly S_k is convex and compact. Let $P \in S_k$. Then from (31) and (29) we have that $\|P\|_{\bar{J}} \leqq \|P_\Lambda\|_J + 1 = K$. Thus, for $t \in \bar{J}$ and $P \in S_k$, $\|\hat{F}[P, g](t)\| \leqq M$. Assume the set $\{\phi_i(t)\}_{i=n}^{k}$, $t \in \bar{J}$, described below (14) satisfies the Haar condition. For any $P \in S_k$ define $T_k P$ as follows: if V_b is the (now unique) solution to (13), then

$$T_k P = V_b. \tag{32}$$

THEOREM 2. *Let g and $\bar{J}$ be as in* (30), *and assume* (29) *is valid for $t \in \bar{J}$. If T_k is defined by* (32), *then T_k maps S_k into S_k.*

Proof. Let $P \in S_k$. Then

$$\begin{aligned} &\|\dot{V}_b - \bar{A}V_b - \bar{B}[V_b \circ g] - \hat{F}[P, g]\|_{\bar{J}} \\ &\qquad \leqq \|\dot{P}_\Lambda - \bar{A}P_\Lambda - \bar{B}[P_\Lambda \circ g] - \hat{F}[P, g]\|_{\bar{J}} \leqq \hat{\eta}, \end{aligned} \tag{33}$$

where $\hat{\eta}$ is defined by (22). Inequality (33) implies that

$$\|\eta(V_b, P)\|_{\bar{J}} \leqq \hat{\eta}. \tag{34}$$

Now (21) with V_b in place of P_Λ holds if and only if

$$\begin{aligned} V_b(t) - \bar{Z}(t)\Lambda = &\int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)\hat{F}[P, g](s)\, ds \\ &+ \int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)\eta(V_p, P)(s)\, ds \\ &+ \int_0^t \bar{Z}(t)\bar{Z}^{-1}(s)B(s)[V_b \circ g](s)\, ds, \qquad t \in J. \end{aligned} \tag{35}$$

Thus

$$\|V_b(t) - L_g[V_b](t)\| \leqq \alpha^2(M+\hat{\eta})|t|. \tag{36}$$

We note that (35) and (9) imply that V_b and $L_g[V_b]$ are elements of $\Omega_{\bar{J}}$. Therefore (36) implies that

$$\rho(V_b, L_g[V_b]) \leqq \alpha^2(M+\hat{\eta}). \tag{37}$$

Due to the construction of g, L_g is a contraction operator on both $\Omega_{\bar{J}}$ and Ω_J, and Y in

(24) is a fixed point of L_g in $\Omega_{\bar{J}}$ and Ω_J. Consequently utilizing (37) and considering $Y \in \Omega_{\bar{J}}$, we obtain

$$\rho(V_b, Y) \leqq \rho(V_b, L_g[V_b]) + \rho(L_g[V_b], L_g[Y])$$
$$\leqq \alpha^2(M + \hat{\eta}) + \mu\rho(V_b, Y),$$

where μ is defined in (27). Thus $\rho(V_b, Y) \leqq \alpha^2(M + \hat{\eta})/(1 - \mu)$. This inequality implies that

$$\|V_b(t) - Y(t)\| \leqq \frac{\alpha^2(M + \hat{\eta})}{1 - \mu}|t|, \qquad t \in \bar{J}. \tag{38}$$

Thus, using (28) and (38), we have that

$$\|V_b(t) - P_\Lambda(t)\| \leqq \frac{2\alpha^2(M + \hat{\eta})}{1 - \mu}|t|, \qquad t \in \bar{J}.$$

Consequently $V_b \in S_k$, and $T_k : S_k \to S_k$. □

THEOREM 3. *Assume the hypotheses of Theorem* 2 *are valid. Then the mapping* T_k *defined in* (32) *is a continuous mapping.*

Proof. Let $\{\phi_i(t)\}_{i=n}^k$ be as defined below (14). By assumption $\{\phi_i\}_{i=n}^k$ satisfies the Haar condition. Suppose that $\hat{P} \in S_k$ and that $\bar{P}$ is any other element of S_k; denote $T_k[\hat{P}] = \hat{Q}$ and $T_k[\bar{P}] = \bar{Q}$. Suppose that

$$\hat{q}(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_{n-1}t^{n-1} + \hat{\beta}_n t^n + \cdots + \hat{\beta}_k t^k$$

and

$$\bar{q}(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_{n-1}t^{n-1} + \bar{\beta}_n t^n + \cdots + \bar{\beta}_k t^k$$

are the first elements of $\hat{Q}$ and $\bar{Q}$, respectively. Then, by (14) and (15),

$$\inf_{v \in \mathscr{P}_k} \sup_{\bar{J}} |A[v](t) + B[V \circ g](t) - N[\hat{P}, g](t)|$$
$$= \sup_{\bar{J}} |A[\hat{q}](t) + B[\hat{Q} \circ g](t) - N[\hat{P}, g](t)|$$
$$= \sup_{\bar{J}} \left| \sum_{i=n}^{k} \hat{\beta}_i \phi_i(t) - N^*[\hat{P}, g](t) \right|.$$

A similar chain of inequalities holds for $\bar{q}$. Now the classical Freud theorem [3, p. 82] implies that there exists a constant $\lambda_{\hat{p}}$ such that

$$\left\| \sum_{i=n}^{k} (\hat{\beta}_i - \bar{\beta}_i)\phi_i \right\|_{\bar{J}} \leqq \lambda_{\hat{p}} \|N^*[\hat{P}, g] - N^*[\bar{P}, g]\|_{\bar{J}}.$$

This inequality and an argument paralleling that given to prove [8, Thm. 1] now establish that T_k is continuous. □

COROLLARY 1. *Suppose that the hypotheses of Theorem* 2 *are satisfied. Then the mapping* T_k *defined in* (32) *has a fixed point* P_k *in* S_k.

Proof. Since S_k is a compact, convex subset of $\mathscr{P}_k$, the result follows from the Schauder fixed point theorem [18, p. 25]. □

We summarize the result of the corollary. There exists a $P \in \mathscr{P}_k$ such that if we seek the best approximation V_b to $\hat{F}[P, g]$ in the sense of (13), then this best approximation is $V_b = P$. In the scalar equivalent form (14) this best approximation is $v_b = p$, the first component of P. We note that the fixed point P of T_k guaranteed by Corollary 1 would not necessarily be the fixed point in [8, Cor. 1]. No fixed point analysis is given in [1].

The proofs of Theorems 2 and 3 do not depend on k, and consequently there exists a sequence of fixed points,

$$\{P_k(t)\}_{k=n}^{\infty}, \qquad t \in \bar{J}, \tag{39}$$

guaranteed by the corollary.

We now establish that any cluster point of (39) is a solution to (7) with $g(t)$ in place of $h(t)$, $t \in \bar{J}$. Since $g(t) = h(t)$, $t \in J_{\bar{s}+1}$, this will establish that any cluster point of (39) is a solution to (7) on $J_{\bar{s}+1}$. A similar result is established in [8, Thm. 2]; however, since ϕ_i, $i = n, \cdots, k$, is not necessarily a polynomial, a different approach is needed in the present paper.

THEOREM 4. *Assume the hypotheses of Theorem* 2 *are valid. Then the sequence* (39) *has a cluster point.*

Proof. Let P_k be a fixed point of the operator T_k, $k = n, n+1, \cdots$. Then $P_k \in S_k$. Thus $\{\|P_k\|_{\bar{J}}\}_{k=n}^{\infty}$ is a bounded sequence. Since P_k is a fixed point of T_k, inequality (33) now implies that $\{\|\dot{P}_k\|_{\bar{J}}\}_{k=n}^{\infty}$ is a bounded sequence. Therefore (39) is a uniformly bounded, equicontinuous family, and consequently (39) has a cluster point $W : \bar{J} \to R_n$. □

THEOREM 5. *Assume the hypotheses of Theorem* 2 *are satisfied. Let W be a cluster point of* (39). *Then W satisfies* (7) *with g in place of h, $t \in \bar{J}$; furthermore, W is a solution to* (7) *with deviating argument h on $J_{\bar{s}+1}$.*

Proof. Consider the initial value problem with deviating argument g,

$$\begin{aligned} &\dot{X}(t) - \bar{A}(t)X(t) - \bar{B}(t)X(g(t)) = F(t, W(t), W(g(t))), \\ &X(0) = \Lambda, \qquad t \in \bar{J}. \end{aligned} \tag{40}$$

Let $G(t, X) = F(t, W(t), W(g(t))) + \bar{B}(t)X$. Then G satisfies a Lipschitz condition in the second variable with Lipschitz constant β. Now [1, Thm. 1] and (17) imply that (40) has a unique solution on $\bar{J}$, say $\bar{W}(t)$. The Weierstrass theorem guarantees that there exists a sequence $\{Q_k\}_{k=n}^{\infty} \subseteq \hat{\mathcal{P}}_k$, such that $Q_k \to \bar{W}$ and $\dot{Q}_k \to \dot{\bar{W}}$ uniformly on $\bar{J}$. Since P_k is a fixed point of T_k,

$$\begin{aligned} &\|\dot{P}_{k(j)} - \bar{A}P_{k(j)} - \bar{B}[P_{k(j)} \circ g] - \hat{F}[P_{k(j)}, g]\| \\ &\qquad \leqq \|\dot{Q}_{k(j)} - \bar{A}Q_{k(j)} - \bar{B}[Q_{k(j)} \circ g] - \hat{F}[P_{k(j)}, g]\|_{\bar{J}}, \end{aligned}$$

where $P_{k(j)} \to W$ as $j \to +\infty$. This inequality now implies that

$$\lim_{j \to \infty} \|\dot{P}_{k(j)} - \bar{A}P_{k(j)} - \bar{B}[P_{k(j)} \circ g] - \hat{F}[P_{k(j)}, g]\|_{\bar{J}} = 0,$$

and consequently

$$\lim_{j \to \infty} \dot{P}_{k(j)} = \bar{A}W + \bar{B}[W \circ g] + \hat{F}[W, g]. \tag{41}$$

Thus the sequence $\{\dot{P}_{k(j)}(t)\}_{j=1}^{\infty}$ is uniformly convergent on $\bar{J}$. Integrating (41) results in

$$\lim_{j \to \infty} (P_{k(j)}(t) - P_{k(j)}(0)) = \int_0^t \{\bar{A}(s)W(s) + \bar{B}(s)[W \circ g](s) + \hat{F}[W, g](s)\}\, ds.$$

This equality now implies that

$$\dot{W}(t) = \bar{A}(t)W(t) + \bar{B}(t)[W \circ g](t) + \hat{F}[W, g](t), \qquad W(0) = \Lambda,$$

completing the proof. □

Hereafter we designate a kth degree fixed point of T_k as a *modified* SAS of degree k, abbreviated MSAS. We note that if $\{\phi_j(t)\}_{j=n}^{k}$ satisfies the Haar condition and if $t \in J_{\bar{s}+1}$, then the BAS of degree k considered in [1] is a special case of MSAS.

The more difficult theory developed for the MSAS does parallel that given in [8] for the SAS; however, the SAS development does not allow for the identification and subsequent exploitation of the linear component of (7). The next section demonstrates that the identification and utilization of the linear part of (7) frequently accelerates the calculations. Both the MSAS and SAS theories are local in the sense that the resulting cluster points satsify (7) on possibly smaller intervals contained in the original interval J.

4. Computations. The algorithm used to calculate the SAS approximate below is described in [8, § 4]. The MSAS algorithm now described is based on the theory of § 3. For k sufficiently large, initially choose $P_{k0}(t) = P_\Lambda(t)$. At the $l+1$ step solve via the second algorithm of Remes the best approximation problem

$$\begin{aligned}\inf_{V \in \mathscr{P}_k} &\|\dot{V} - \bar{A}V - \bar{B}[V \circ g] - \hat{F}[P_{k,l}, g]\|_{\bar{J}} \\ &= \inf_{\{\beta_i\} \in R_{k-n+1}} \sup_{\bar{J}} \left| \sum_{i=n}^{k} \beta_i \phi_i(t) + \sum_{i=0}^{n-1} \alpha_i \phi_i(t) - N[P_{k,l}, g](t) \right|,\end{aligned} \tag{42}$$

where the ϕ_i, $i = 0, \cdots, k$ are described below (14). Let the solution to this minimization problem be $P_{k,l+1}$ and continue. This procedure results in a sequence (uniquely determined if ϕ_i, $i = n, \cdots, k$ satisfies the Haar condition)

$$\{P_{k,l}\}_{l=0}^{\infty}, \tag{43}$$

with corresponding first elements

$$\{p_{k,l}\}_{l=0}^{\infty}. \tag{44}$$

Any cluster point of these sequences then represents the MSAS of degree k to (7) and (2), respectively, on $\bar{J}$.

All calculations for the tables below are computed in the scalar setting. Furthermore, the calculations are carried out on an interval $\bar{J} \subseteq J$ for which $h(\bar{J}) \subseteq \bar{J}$. If no such interval $\bar{J}$ exists, the construction of an appropriate g would be necessary for both the SAS and MSAS algorithms.

As in [8, pp. 438–439], a cluster point of (43) need not be a fixed point of T_k. An additional condition guaranteeing this is similar to that given in [8, p. 438].

Let p_k be the MSAS or SAS of degree k. The notation below is employed in the tables that follow:

$$\varepsilon_k = \max_J |A[p_k](t) + B[P_k \circ h](t) - N[P_k, h](t)|,$$

$$\Delta_k = \max_J |p_k(t) - x(t)|,$$

where $x(t)$ is the solution of (2), $t \in J$. The column designated by CYC (number of cycles) indicates the number of iterations required in the MSAS or SAS algorithms to achieve a prescribed tolerance. All calculations were effected on the IBM 370, in double precision arithmetic, at the College of Charleston. In each example the approximating class is $\mathscr{P}_6$.

EXAMPLE 1.

$$\ddot{x}(t) + 4x(t) = e^t,$$

$$\dot{x}(0) = 0, \qquad \dot{x}(0) = 1, \qquad t \in [-1, 1].$$

For this example $h(t) = t$. In the MSAS minimization problem (42),

$$A[x](t) = \ddot{x}(t) + 4x(t), \quad B[X \circ h](t) \equiv 0, \quad N[X, h](t) = e^t.$$

Thus $\phi_i(t) = i(i-1)t^{i-2} + 4t^i$, $i = 2, 3, \cdots, 6$, $\phi_0(t) = 4$, $\phi_1(t) = 4t$, and the minimization problem (42) is at the first iteration

$$\inf_{\{\beta_i\} \in R_5} \sup_J \left| \sum_{i=2}^{6} \beta_i [i(i-1)t^{i-2} + 4t^i] + 4t - e^t \right|.$$

Since N is actually independent of x for this example, the MSAS algorithm converges after just one iteration. This behavior is characteristic of the MSAS algorithm in those cases when (2) is linear, and thus the MSAS algorithm is preferable to the SAS algorithm in the strictly linear case. For Example 1, Table 1 demonstrates the preferability of the MSAS algorithm. Column 2 (CPU) of Table 1 and subsequent tables indicate the amount of time in seconds used by the central processor unit to compute the SAS and MSAS approximations of degree k. The column entitled EX represents the limiting extremal set of the SAS or MSAS algorithms (see [8, p. 438]).

TABLE 1

	CYC	CPU	ε_k	Δ_k	EX
SAS	12	4.50	2.0699 (10^{-2})	2.6895 (10^{-3})	−1., −.7973, −.2764, .3366, .8200, 1
MSAS	1	2.91	2.0699 (10^{-2})	2.6895 (10^{-3})	−1., −.7951, −.2772, .3374, .8200, 1

For Example 1, the actual solution is $x(t) = .4 \sin 2t - .2 \cos 2t + .2e^t$, and the SAS approximation of degree six is

$$P_6(t) = t + .498277t^2 - .482941t^3 - .120676t^4 + .084388t^5 + .014242t^6.$$

The coefficients of the MSAS approximation of degree six agree with those of the SAS approximation to six decimal places.

Example 2.

$$\ddot{x}(t) + 4x(t) + \frac{t}{5}\dot{x}\left(\frac{t}{2}\right) + \frac{2}{5}tx\left(\frac{t}{2}\right) = \frac{2t}{5}(\sin t + \cos t),$$

$$x(0) = 0, \qquad \dot{x}(0) = 2, \qquad t \in [-.6, .6].$$

TABLE 2

	CYC	CPU	ε_k	Δ_k
SAS	9	4.76	4.727 (10^{-3})	3.0796 (10^{-4})
MSAS	1	3.49	4.727 (10^{-3})	3.0797 (10^{-4})

The actual solution to Example 2 is $x(t) = \sin 2t$.

Remark. The MSAS part of Example 2 appears in [1, Ex. 1] for degrees five and seven.

Example 3.

$$\ddot{x}(t)+4x(t)+2x^2\left(\frac{t}{2}\right)=\cos 2t+1,$$

$$x(0)=1, \qquad \dot{x}(0)=0, \qquad t\in[-1,1].$$

The solution to Example 3 is $x(t)=\cos 2t$. The ϕ_i, $i=0,\cdots,6$ are described in Example 1, and the minimization problem (42) at the lth iteration is

$$\inf_{\{\beta_i\}\in R_5}\sup_J\left|\sum_{i=2}^{6}\beta_i[i(i-1)t^{i-2}+4t^i]+4-N[P_{6,l}\circ h](t)\right|,$$

where

$$N[X\circ h](t)=-2x^2\left(\frac{t}{2}\right)+\cos 2t+1.$$

The MSAS algorithm is initiated by selecting $P_{6,0}(t)\equiv 1$. Thus the first iteration is effected with $\hat{N}[P_{6,0}\circ h](t)=\cos 2t-1$.

TABLE 3

	CYC	CPU	ε_k	Δ_k
SAS	13	4.95	8.98 (10^{-3})	3.264 (10^{-4})
MSAS	6	4.06	8.98 (10^{-3})	3.264 (10^{-4})

Example 3 is interesting in that, even though the approximating functions ϕ_i in the MSAS algorithm are more complicated than the powers of t that always occur in the SAS algorithm, and in spite of the fact that (2) is nonlinear, the MSAS algorithm is slightly faster than the SAS algorithm.

Example 4.

$$\ddot{x}(t)+\tfrac{1}{2}\sin(2t)\dot{x}(t)+x^2(h(t))=1-\sin^3 t-\cos^2(h(t)),$$

$$x(0)=0, \qquad \dot{x}(0)=1, \quad t\in[-1,1],$$

where

$$h(t)\begin{cases}(t^3+t^2)\sin\dfrac{\pi}{t}, & t\neq 0,\\ 0, & t=0\end{cases}$$

The solution is $x(t)=\sin t$.

TABLE 4

	CYC	CPU	ε_k	Δ_k	EX
SAS	12	8.02	5.010(10^{-4})	9.965(10^{-4})	−1., −.802, −.340, −.290, .828, 1
MSAS	7	8.42	5.010(10^{-4})	9.963(10^{-4})	−1., −.800, −.343, .289, .829, 1

For Example 4, the SAS approximation is

$$P_6(t) = t + .310237(10^{-4})t^2 - .166268t^3 - .624052(10^{-4})t^4 + .783080(10^{-2})t^5 + .235912(10^{-4})t^6.$$

The MSAS approximation is

$$P_6(t) = t + .310617(10^{-4})t^2 - .166268t^3 - .624461(10^{-4})t^4 + .783080(10^{-2})t^5 + .236053(10^{-4})t^6.$$

5. Conclusions. This paper utilizes approximation theory techniques to obtain uniform approximate solutions to variants of initial value problem (2). An algorithm based on the theory of this paper is then constructed. This algorithm (MSAS) is compared to an algorithm (SAS) previously developed by the authors [8] to obtain approximate solutions to differential equations with deviating arguments. The MSAS algorithm is computationally superior to the SAS algorithm in those cases where (2) is linear. This superiority is not unexpected in that the MSAS fully utilizes the linear part of (2).

For nonlinear differential equations with deviating arguments, neither algorithm is consistently superior. In the nonlinear case the additional complexity of the MSAS algorithm might then suggest the preferability of the SAS algorithm.

The present paper does represent a unification of the theories presented in [1], [8] in the sense that if the differential equation with deviating argument is strictly nonlinear ($a_j(t) \equiv 0$ and $b_j(t) \equiv 0$, $j = 0, \cdots, n-1$, in (2)), then the MSAS and SAS theories produce the same results. The MSAS theory of the present paper does fully include the linear theory developed in [1].

REFERENCES

[1] G. ALLINGER AND M. S. HENRY, *Approximate solutions of differential equations with deviating arguments*, SIAM J. Numer. Anal., 13 (1976), pp. 412–426.

[2] A. BACOPOULOUS AND A. G. KARTSATOS, *On polynomials approximating solutions of nonlinear differential equations*, Pacific J. Math., 40 (1972), pp. 1–5.

[3] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[4] J. HALE, *Functional Differential Equations*, Applied Mathematical Sciences, vol. I, Springer-Verlag, New York, 1971.

[5] J. N. HENRY, M. S. HENRY AND D. E. TAYLOR, *An application of the Roberts minimal degree algorithm to initial value problems*, Int. J. Numer. Meth. Engng., 12 (1978), pp. 1347–1358.

[6] M. S. HENRY, *Best approximate solutions of nonlinear differential equations*, J. Approximation Theory, 3 (1970), pp. 59–65.

[7] M. S. HENRY AND K. WIGGINS, *Applications of approximation theory to the initial value problem*, J. Approximation Theory, 17 (1976), pp. 66–85.

[8] ———, *Applications of approximation theory to differential equations with deviating arguments*, Pacific J. Math., 76 (1978), pp. 431–441.

[9] R. G. HUFFSTUTLER AND F. M. STEIN, *The approximate solution of certain nonlinear differential equations*, Proc. Amer. Math. Soc., 19 (1968), pp. 998–1002.

[10] ———, *The approximate solution of $\dot{y} = F(x, y)$*, Pacific J. Math., 24 (1968), pp. 283–289.

[11] A. G. KARTSATOS AND E. B. SAFF, *Hyperpolynomial approximation of solutions of nonlinear integro-differential equations*, Pacific J. Math., 49 (1973), pp. 117–125.

[12] R. J. OBERG, *On the local existence of solutions of certain functional differential equations*, Proc. Amer. Math. Soc., 20 (1969), pp. 295–302.

[13] A. G. PETSOULAS, *The approximate solution of Volterra integral equations*, J. Approximation Theory, 14 (1975), pp. 152–159.
[14] T. J. RIVLIN, *An Introduction to the Approximation of Functions*, Blaisdell, Waltham, MA., 1969.
[15] G. H. RYDER, *Solutions of functional differential equations*, Amer. Math. Monthly, 76 (1969), pp. 1031–1033.
[16] D. SCHMIDT AND K. WIGGINS, *Minimax approximate solutions of linear boundary value problems*, Math. Comput., 33 (1979), pp. 139–148.
[17] K. SCHMITT, ed., *Delay and Functional Differential Equations and Their Applications*, Academic Press, New York, 1972.
[18] D. R. SMART, *Fixed Point Theorems*, Cambridge University Press, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0006 $01.00/0

FAMILIES OF BIORTHOGONAL RATIONAL FUNCTIONS IN A DISCRETE VARIABLE*

MIZAN RAHMAN†

Abstract. By using Toscano's [Boll. Un. Mat. Ital. (3), 4 (1949), pp. 398–409] finite difference representation for generalized hypergeometric functions, two families of biorthogonal rational functions in an integer-valued variable are obtained, one a ${}_3F_2$ series and the other a balanced ${}_4F_3$ series.

1. Introduction. There has been a growing interest in recent years in functions of discrete variables, especially the classical orthogonal polynomials which include Hahn, Krawtchouk, Meixner, Gottlieb, and Charlier polynomials (for a brief account of these polynomials see [11, pp. 221–227], [19, pp. 33–37]). Karlin and McGregor [15], [16] found applications of the Hahn polynomials in genetics; Delsarte [5], [6] and Sloane [19] used the dual Hahn and Krawtchouk polynomials in coding theory; Eagleson [10] exploited the properties of Krawtchouk polynomials to obtain some results in statistics; Cooper, Hoare and Rahman [4] found a probabilistic model in which all the polynomials mentioned above acquire a clear stochastic significance. Discrete polynomials also figure prominently in the group-theoretic works of Dunkl and Ramirez [7]–[9].

In trying to find a product formula and an addition theorem for the Hahn polynomials, the author [17] recently bumped into a very novel kind of discrete functions—a family of biorthogonal rational functions—which are expressible as balanced ${}_4F_3$ hypergeometric functions with unit argument. It is not unreasonable to expect that these rational functions will find applications in other areas and may enable us to solve problems that were hitherto considered difficult, if not intractable. The purpose of the present paper is to introduce such families of biorthogonal rational functions in a systematic manner and find some of their interesting properties.

Classical orthogonal systems can be defined in a number of equivalent ways, as solutions of certain differential or difference equations, as solutions of some recurrence relations, in terms of generating functions, or in terms of Rodrigues' formulas. For the purpose of generalizations, however, the definitions through Rodrigues' formulas seem to be very convenient. To illustrate the point, let us consider the Hahn polynomials $Q_n(x)$ which are represented as hypergeometric series [11], [13],

$$\begin{aligned} Q_n(x) &\equiv Q_n(x;\alpha,\beta,N) \\ &= {}_3F_2\left[\begin{matrix}-n, n+\alpha+\beta+1, -x \\ \alpha+1, -N\end{matrix}\right], \end{aligned} \tag{1.1}$$

where $x, n = 0, 1, \cdots, N$ and α, β are arbitrary real or complex parameters, except when they have such values as to render the above series meaningless. The above hypergeometric representation was first given by Erdélyi and Weber [11], [15] who also gave their Rodrigues' formula [22] (see also [3]),

(1.2)

$$\rho(x;\alpha,\beta,N)Q_n(x;\alpha,\beta,N) = \frac{\binom{n+\beta}{n}}{\binom{N}{n}\binom{N+\alpha+\beta+1}{N}}\Delta_x^n\left[\binom{x+\alpha}{\alpha+n}\binom{N-x+\beta+n}{\beta+n}\right],$$

* Received by the editors May 18, 1978, and in revised form September 8, 1980. This work was supported by the National Research Council of Canada under grant A6197.

† Department of Mathematics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

where

$$\rho(x;\alpha,\beta,N)=\binom{N}{x}\frac{(\alpha+1)_x(\beta+1)_{N-x}}{(\alpha+\beta+2)_N}, \tag{1.3}$$

is the weight function associated with the Hahn polynomials [13], [15], and

$$\Delta_x f(x)\equiv f(x+1)-f(x), \qquad \Delta_x^n f(x)=\Delta_x[\Delta_x^{n-1}f(x)], \qquad n=1,2,\cdots.$$

The factorial functions appearing in (1.3) are written in Pochhammer notation,

$$(a)_x=\begin{cases} a(a+1)\cdots(a+x-1), & x \text{ a positive integer},\\ \dfrac{\Gamma(a+x)}{\Gamma(a)}, & \text{arbitrary } x, \end{cases} \tag{1.4}$$

with $(a)_0=1$. These functions will henceforth be referred to as Pochhammer functions or products. In this notation, Rodrigues' formula (1.2) reads

$$\begin{aligned}&\rho(x;\alpha,\beta,N)Q_n(x;\alpha,\beta,N)\\ &=\frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)\Gamma(\alpha+\beta+2+N)}\frac{(N-n)!}{(\alpha+1)_n}\\ &\quad\times\Delta_x^n[(x-n+1)_{\alpha+n}(N-x+1)_{\beta+n}].\end{aligned} \tag{1.5}$$

The form of the right-hand side now suggests some generalizations. For example, one may try to compute the nth difference of the product of three or more Pochhammer functions. For the sake of generality, it is preferable to consider a Rodrigues' operator of the type

$$L_{n,k}(x)\equiv\Delta_x^n[(a_1+x)_{b_1}(a_2-x)_{b_2}\cdots(a_{2k-1}+x)_{b_{2k-1}}(a_{2k}-x)_{b_{2k}}], \tag{1.6}$$

where k is a fixed positive integer and the a's and b's are arbitrary complex parameters which may not be independent of n. The expression above can be computed in terms of a hypergeometric function by using the elementary formulas of finite difference calculus and some well-known transformation and summation theorems of the hypergeometric series [2], [18]. However, the hard work was done by Toscano [21, p. 403] when he obtained the finite difference representation for generalized hypergeometric functions:

$$\begin{aligned}&{}_{p+1}F_{q+1}\left[\begin{matrix}-n, x+\alpha_1,\cdots,x+\alpha_p\\ x, x+\beta_1,\cdots,x+\beta_q\end{matrix};t\right]\\ &=(-1)^n\frac{\Gamma(x)\Gamma(x+\beta_1)\cdots\Gamma(x+\beta_q)}{\Gamma(x+\alpha_1)\cdots\Gamma(x+\alpha_p)t^x}\Delta_x^n\left[\frac{\Gamma(x+\alpha_1)\cdots\Gamma(x+\alpha_p)}{\Gamma(x)\Gamma(x+\beta_1)\cdots\Gamma(x+\beta_q)}t^x\right].\end{aligned} \tag{1.7}$$

It follows immediately that

$$\begin{aligned}L_{n,k}(x)&=(-1)^n(a_1+x)_{b_1}(a_2-x)_{b_2}\cdots(a_{2k-1}+x)_{b_{2k-1}}(a_{2k}-x)_{b_{2k}}\\ &\quad\times{}_{2k+1}F_{2k}\left[\begin{matrix}-n, a_1+b_1+x, a_3+b_3+x,\cdots,a_{2k-1}+b_{2k-1}+x,\\ \qquad 1+x-a_2, 1+x-a_4,\cdots,1+x-a_{2k}\\ a_1+x, a_3+x,\cdots,a_{2k-1}+x, 1+x-a_2-b_2,\\ \qquad 1+x-a_4-b_4,\cdots,1+x-a_{2k}-b_{2k}\end{matrix};1\right].\end{aligned} \tag{1.8}$$

By reversing the series on the right and simplifying, we have an alternative form

$$
\begin{aligned}
L_{n,k}(x) &= (a_1+x+n)_{b_1}(a_3+x+n)_{b_3}\cdots(a_{2k-1}+x+n)_{b_{2k-1}} \\
&\quad \times (a_2-x-n)_{b_2}(a_4-x-n)_{b_4}\cdots(a_{2k}-x-n)_{b_{2k}} \\
&\quad \times {}_{2k+1}F_{2k}\left[\begin{matrix} -n,\ a_2+b_2-x-n,\ a_4+b_4-x-n,\ \cdots,\ a_{2k}+b_{2k}-x-n, \\ 1-a_1-x-n,\ \cdots,\ 1-a_{2k-1}-x-n \\ 1-a_1-b_1-x-n,\ 1-a_3-b_3-x-n,\ \cdots, \\ 1-a_{2k-1}-b_{2k-1}-x-n,\ a_2-x-n,\ \cdots,\ a_{2k}-x-n \end{matrix}\right]
\end{aligned} \tag{1.9}
$$

Our aim is to establish orthogonality or biorthogonality of certain hypergeometric functions by using these representations. It is clear that one cannot hope to accomplish much except in the simple cases $k=1$, $k=2$. The case of $k=1$ when $L_{n,k}(x)$ reduces to a ${}_3F_2$ series will be considered in §§ 2 and 3. When $k=2$ the hypergeometric function in (1.8) or (1.9) is a ${}_5F_4$ series which, in general, is not easy to manipulate. This ${}_5F_4$ series becomes a ${}_4F_3$ when one of the b parameters vanishes. Even a general ${}_4F_3$ does not have a known transformation formula unless it is balanced or nearly well-poised. In §§ 4 and 5, we shall deal with the case when $k=2$ and the resulting hypergeometric series is a balanced (or, more commonly, Saalschutzian) ${}_4F_3$.

In carrying out the calculations in the following sections, the most fundamental identity we shall need is Whipple's transformation formula [2, p. 56]

$$
{}_4F_3\left[\begin{matrix} x, y, z, -k \\ u, v, w \end{matrix}\right] = \frac{(v-z)_k(w-z)_k}{(v)_k(w)_k}\,{}_4F_3\left[\begin{matrix} u-x, u-y, z, -k \\ u, 1-v+z-k, 1-w+z-k \end{matrix}\right], \tag{1.10}
$$

provided the ${}_4F_3$ series is balanced, that is,

$$
u+v+w = x+y+z-k+1. \tag{1.11}
$$

Special cases of (1.10) are the Pfaff–Saalschutz theorem [1, p. 62] for a balanced ${}_3F_2$ series and the well-known transformation formulas for a general terminating ${}_3F_2$:

$$
{}_3F_2\left[\begin{matrix} -k, a, b \\ c, d \end{matrix}\right] = \frac{(d-b)_k}{(d)_k}\,{}_3F_2\left[\begin{matrix} -k, c-a, b \\ c, 1+b-d-k \end{matrix}\right] \tag{1.12}
$$

[2, pp. 17–20], [14] and

$$
{}_3F_2\left[\begin{matrix} -k, a, b \\ c, d \end{matrix}\right] = \frac{(c+d-a-b)_k}{(d)_k}\,{}_3F_2\left[\begin{matrix} -k, c-a, c-b \\ c, c+d-a-b \end{matrix}\right] \tag{1.13}
$$

[2, p. 98].

2. Rodrigues' formulas for ${}_3F_2$'s. If we set $k=1$ in (1.9) we get a ${}_3F_2$ series on the right-hand side that involves the parameters a_1, a_2, b_1 and b_2. However, it can be seen that one would get a ${}_3F_2$ from the general $L_{n,k}$ by setting all but two of the b-parameters equal to zero. For instance, in $L_{n,2}$ we may set b_2 and b_4 equal to zero and get a ${}_3F_2$ with the parameters a_1, a_3, b_1, b_3. It is easily seen that this ${}_3F_2$ is basically equivalent to the one for $k=1$. We may then, without any loss of generality, consider the general 4-parameter Rodrigues' formula

$$
\begin{aligned}
L_{n,1}(x) &\equiv L_n(x; a_1, a_3; b_1, b_3) \\
&\equiv \Delta_x^n[(a_1+x)_{b_1}(a_3+x)_{b_3}] \\
&= (a_1+x+n)_{b_1}(a_3+x+n)_{b_3}\,{}_3F_2\left[\begin{matrix} -n, 1-a_1-x-n, 1-a_3-x-n \\ 1-a_1-b_1-x-n, 1-a_3-b_3-x-n \end{matrix}\right].
\end{aligned}
$$

Using (1.12) and (1.13) and simplifying, we get

$$L_{n,1}(x)=\frac{(-1)^n(a_1+x)_{b_1}(a_3+x)_{b_3}(-b_1-b_3)_n}{(a_3+x)_n}\,{}_3F_2\left[\begin{matrix}-n,\,-b_1,\,a_1-a_3-b_3\\ -b_1-b_3,\,a_1+x\end{matrix}\right]$$

$$=\frac{(-1)^n(a_1+x)_{b_1}(a_3+x)_{b_3}(-b_1-b_3)_n}{(a_1+x)_n}\,{}_3F_2\left[\begin{matrix}-n,\,-b_3,\,a_3-a_1-b_1\\ -b_1-b_3,\,a_3+x\end{matrix}\right]. \tag{2.1}$$

On the other hand, if one applies (1.12) twice in a row and then simplifies, one obtains

$$L_{n,1}(x)=\frac{(a_1+x)_{b_1}(a_3+x)_{b_3}(-b_3)_n(a_3-a_1-b_1)_n}{(a_1+x)_n(a_3+x)_n}$$

$$\times\,{}_3F_2\left[\begin{matrix}-n,\,b_1+b_3-n+1,\,1-a_3-n-x\\ b_3+1-n,\,1-a_3+a_1+b_1-n\end{matrix}\right]. \tag{2.2}$$

Note that (2.2) reduces to the Rodrigues formula for Hahn polynomials if we choose $a_3=1-n$, $b_3=\alpha+n$, $b_1=\beta+n$, $a_1=-\beta-N-n$. Symmetry implies that $L_{n,1}(x)$ gives the same Hahn polynomials if a_1, b_1 are interchanged with a_3, b_3, respectively.

In order that we may obtain something that is essentially different from Hahn polynomials, we will assume that a_1 and a_3 are different from $1-n$.

Using (2.1), we are now able to write down two separate formulas:

$$\Delta_x^n[(a_1+x+1)_{b_1}(a_3+x)_{b_3+1}]=\frac{w^{(1)}(x)R_n^{(1)}(x)(-1)^n(-b_1-b_3-1)_n(a_3+x)}{(a_3+x)_n}, \tag{2.3}$$

and

$$\Delta_x^n[(a_1+x)_{b_1+1}(a_3+x+1)_{b_3}]=\frac{w^{(1)}(x)S_n^{(1)}(x)(-1)^n(-b_1-b_3-1)_n(a_1+x)}{(a_1+x)_n}, \tag{2.4}$$

where

$$w^{(1)}(x)=(a_1+x+1)_{b_1}(a_3+x+1)_{b_3}, \tag{2.5}$$

$$R_n^{(1)}(x)={}_3F_2\left[\begin{matrix}-n,\,-b_1,\,a_1-a_3-b_3\\ -b_1-b_3-1,\,a_1+x+1\end{matrix}\right], \tag{2.6}$$

$$S_n^{(1)}(x)={}_3F_2\left[\begin{matrix}-n,\,-b_3,\,a_3-a_1-b_1\\ -b_1-b_3-1,\,a_3+x+1\end{matrix}\right]. \tag{2.7}$$

We shall show in the following section that, under appropriate conditions, $\{R_n^{(1)}(x)\}$ and $\{S_n^{(1)}(x)\}$ for $n=0,1,\cdots$ form a biorthogonal system on a certain discrete set. Note that $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$ coincide when $a_1=a_3$.

3. Biorthogonality of $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$. We assume that

(i) $a_1,\,a_3\neq 0,\pm1,\pm2,\cdots$;

(ii) $a_1+b_1,\,a_3+b_3\neq 0,\pm1,\pm2,\cdots$;

(iii) $b_1,\,b_3,\,a_1+b_1-a_3,\,a_3+b_3-a_1$ are neither 0 nor positive integers; (3.1)

(iv) $b_1+b_3+1<0$;

(v) $x=0,\pm1,\pm2,\cdots$.

Then

$$
\begin{aligned}
&\sum_{x=-\infty}^{\infty} w^{(1)}(x) \\
&= \sum_{x=-\infty}^{\infty} \frac{\Gamma(a_1+b_1+1+x)\Gamma(a_3+b_3+1+x)}{\Gamma(a_1+1+x)\Gamma(a_3+1+x)} \\
&= \frac{\pi^2\Gamma(-b_1-b_3-1)}{\sin\pi(a_1+b_1+1)\sin\pi(a_3+b_3+1)\Gamma(-b_1)\Gamma(-b_3)\Gamma(a_3-a_1-b_1)\Gamma(a_1-a_3-b_3)} \\
&= \mu^{(1)}, \text{ say,}
\end{aligned} \tag{3.2}
$$

by virtue of Dougall's formula for the sum of a bilateral series [18, p. 180]. Because of the assumptions (3.1) the sum is a finite number.

Now let us consider the sum

$$
P_{m,n} \equiv \sum_{x=-\infty}^{\infty} w^{(1)}(x) R_n^{(1)}(x) S_m^{(1)}(x). \tag{3.3}
$$

Using the representations (2.6) and (2.7) and expanding the hypergeometric functions we obtain

$$
P_{m,n} = \sum_{k=0}^{n} \frac{(-n)_k(-b_1)_k(a_1-a_3-b_3)_k}{k!(-b_1-b_3-1)_k} \sum_{l=0}^{m} \frac{(-m)_l(-b_3)_l(a_3-a_1-b_1)_l}{l!(-b_1-b_3-1)_l} Q_{k,l}, \tag{3.4}
$$

where

$$
\begin{aligned}
Q_{k,l} &= \sum_{x=-\infty}^{\infty} \frac{(a_1+x+1)_{b_1}(a_3+x+1)_{b_3}}{(a_1+x+1)_k(a_3+x+1)_l} \\
&= \sum_{x=-\infty}^{\infty} \frac{\Gamma(a_1+b_1+1+x)\Gamma(a_3+b_3+1+x)}{\Gamma(a_1+k+1+x)\Gamma(a_3+l+1+x)} \\
&= \frac{\pi^2\Gamma(-b_1-b_3-1+k+l)}{\sin\pi(a_1+b_1+1)\sin\pi(a_3+b_3+1)\Gamma(-b_1+k)\Gamma(-b_3+l)} \\
&\quad \cdot \frac{1}{\Gamma(a_3-a_1-b_1+l)\Gamma(a_1-a_3-b_3+k)} \\
&= \mu^{(1)} \frac{(-b_1-b_3-1)_{k+l}}{(-b_1)_k(-b_3)_l(a_3-a_1-b_1)_l(a_1-a_3-b_3)_k},
\end{aligned} \tag{3.5}
$$

by virtue of the formula used to derive (3.2). Hence,

$$
\begin{aligned}
P_{m,n} &= \mu^{(1)} \sum_{k=0}^{n} \frac{(-n)_k}{k!} \sum_{l=0}^{m} \frac{(-m)_l(-b_1-b_3-1+k)_l}{l!(-b_1-b_3-1)_l} \\
&= \mu^{(1)} \sum_{k=0}^{n} \frac{(-n)_k}{k!} \frac{(-k)_m}{(-b_1-b_3-1)_m} \\
&= \begin{cases} \dfrac{\mu^{(1)} n!}{(n-m)!(-b_1-b_3-1)_m} \displaystyle\sum_{k=0}^{n-m} \frac{(-n+m)_k}{k!} & \text{if } n \geqq m, \\ 0 & \text{if } n < m. \end{cases}
\end{aligned}
$$

From the binomial theorem, it now follows that

$$P_{m,n} = \left\{ \mu^{(1)} \frac{n!}{(-b_1 - b_3 - 1)_n} \right\} \delta_{mn}. \tag{3.6}$$

The biorthogonality of $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$ also follow from the Rodrigues formulas (2.3) and (2.4). Suppose $n > m \geqq 0$. Then, by (3.5),

$$\begin{aligned} P_{m,n} &= \frac{(-1)^n}{(-b_1 - b_3 - 1)_n} \sum_{x=-\infty}^{\infty} \frac{(a_3 + x)_n}{a_3 + x} S_m^{(1)}(x)\, \Delta_x^n [(a_1 + x + 1)_{b_1} (a_3 + x)_{b_3+1}] \\ &= \frac{(-1)^n (-b_1 - b_3 - 1)_{n-1}}{(-b_1 - b_3 - 1)_n} \left[w^{(1)}(x) R_{n-1}^{(1)}(x) \frac{a_3 + x}{(a_3 + x)_{n-1}} S_m^{(1)}(x) \right]_{x=-\infty}^{\infty} \\ &\quad - \frac{(-1)^n}{(-b_1 - b_3 - 1)_n} \\ &\quad \cdot \sum_{x=-\infty}^{\infty} \Delta_x^{n-1} [(a_1 + x + 2)_{b_1} (a_3 + x + 1)_{b_3+1}]\, \Delta_x \left[\frac{(a_3 + x)_n}{(a_3 + x)} S_m^{(1)}(x) \right]. \end{aligned} \tag{3.7}$$

However,

$$\begin{aligned} \frac{w^{(1)}(x) R_{n-1}^{(1)}(x)(a_3 + x) S_m^{(1)}(x)}{(a_3 + x)_{n-1}} &= \frac{(a_1 + x + 1)_{b_1} (a_3 + x + 1)_{b_3} (a_3 + x) S_m^{(1)}(x)}{(a_3 + x)_{n-1}} \\ &= O(x^{b_1 + b_3 + 1 - (n-1)}) \quad \text{as } |x| \to \infty. \end{aligned}$$

Hence, by (iv) of assumptions (3.1) the first term on the r.h.s. of (3.7) vanishes for $n \geqq 1$. Summing by parts n times, we thus obtain

$$\begin{aligned} P_{m,n} = \frac{1}{(-b_1 - b_3 - 1)_n} \sum_{x=-\infty}^{\infty} &(a_1 + x + n + 1)_{b_1} (a_3 + x + n)_{b_3+1} \\ &\cdot \Delta_x^n \left\{ \frac{(a_3 + x)_n}{a_3 + x} {}_3F_2 \left[\begin{matrix} -m, -b_3, a_3 - a_1 - b_1 \\ -b_1 - b_3 - 1, a_3 + x + 1 \end{matrix} \right] \right\}. \end{aligned} \tag{3.8}$$

Since $m < n$, the expression within the curly brackets is a polynomial of degree $n - 1$ in x, and hence gives zero when it is acted on by the operator Δ_x^n. Thus, $P_{m,n} = 0$ for $m < n$. Similarly, for $0 \leqq n < m$ we use (2.4) and arrive at the same conclusion.

As we mentioned earlier, the biorthogonal rational functions $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$ coincide when $a_1 = a_3$. If, in addition, $b_1 = b_3$, then we obtain a system of orthogonal rational functions with the positive measure

$$w^{(1)}(x) = \{(a + x + 1)_b\}^2 \tag{3.9}$$

and total weight

$$\sum_{x=-\infty}^{\infty} w^{(1)}(x) = \Gamma(-2b - 1) \left\{ \frac{\pi \csc \pi(a + b + 1)}{\Gamma^2(-b)} \right\}^2, \tag{3.10}$$

subject to the restrictions

$$\begin{aligned} &\text{(i)} \quad a,\ a + b \neq 0, \pm 1, \pm 2, \cdots; \\ &\text{(ii)} \quad b < -\tfrac{1}{2}. \end{aligned} \tag{3.11}$$

The orthogonal functions in the integer-valued variable x, $-\infty < x < \infty$, have hypergeometric representation

$$R_n^{(1)}(x) = {}_3F_2\left[\begin{matrix} -n, -b, -b \\ -2b-1, a+x+1 \end{matrix}\right], \tag{3.12}$$

with the orthogonality relation

$$\begin{aligned} &\sum_{x=-\infty}^{\infty} \left\{\frac{\Gamma(a+b+1+x)}{\Gamma(a+1+x)}\right\}^2 {}_3F_2\left[\begin{matrix} -m, -b, -b \\ -2b-1, a+x+1 \end{matrix}\right] {}_3F_2\left[\begin{matrix} -n, -b, -b \\ -2b-1, a+x+1 \end{matrix}\right] \\ &= \left\{\frac{\pi \csc \pi(a+b+1)\Gamma(-2b-1)}{\Gamma^2(-b)}\right\}^2 \frac{\Gamma(n+1)}{\Gamma(-2b-1+n)} \delta_{m,n}. \end{aligned} \tag{3.13}$$

Like most orthogonal and biorthogonal systems in discrete variables, the systems defined in (2.6) and (2.7) also have continuous analogues. In fact, there exist two different continuous analogues in this particular instance, one where the parameters are purely real and the other where the parameters necessarily are complex.

First, let us consider the real case. Let $\alpha, \beta, \gamma, \delta$ be four real parameters such that

$$\begin{aligned} &\text{(i)} \quad \alpha+\beta>0, \quad \beta+\gamma>0, \quad \gamma+\delta>0, \quad \delta+\alpha>0; \\ &\text{(ii)} \quad \alpha+\beta+\gamma+\delta>1. \end{aligned} \tag{3.14}$$

Introduce the weight function

$$v(x) = [\Gamma(\alpha+x)\Gamma(\beta+1-x)\Gamma(\gamma+x)\Gamma(\delta+1-x)]^{-1}, \qquad -\infty < x < \infty, \tag{3.15}$$

with total mass

$$\nu = \int_{-\infty}^{\infty} v(x)\,dx = \Gamma(\alpha+\beta+\gamma+\delta-1)[\Gamma(\alpha+\beta)\Gamma(\beta+\gamma)\Gamma(\gamma+\delta)\Gamma(\delta+\alpha)]^{-1} \tag{3.16}$$

[12, p. 300].

It can be shown by application of the same integration formula that the rational functions

$$\begin{aligned} R_n(x) &= {}_3F_2\left[\begin{matrix} -n, \alpha+\beta, \alpha+\delta \\ \alpha+\beta+\gamma+\delta-1, \alpha+x \end{matrix}\right], \\ S_n(x) &= {}_3F_2\left[\begin{matrix} -n, \gamma+\beta, \gamma+\delta \\ \alpha+\beta+\gamma+\delta-1, \gamma+x \end{matrix}\right] \end{aligned} \tag{3.17}$$

are biorthogonal with respect to the weight $v(x)$. In particular,

$$\int_{-\infty}^{\infty} v(x) R_n(x) S_m(x)\,dx = \frac{\nu n!}{(\alpha+\beta+\gamma+\delta-1)_n} \delta_{m,n}. \tag{3.18}$$

One must realize, however, that even though the integral in (3.18) is well defined in view of the conditions (3.14), the functions $R_n(x)$ and $S_n(x)$ themselves have isolated singularities.

Now we shall consider the case of complex parameters. We assume that

$$\begin{aligned} &\text{(i)} \quad \operatorname{Im}(\alpha, \beta, \gamma, \delta) \neq 0, \\ &\text{(ii)} \quad \alpha+\beta, \beta+\gamma, \gamma+\delta, \delta+\alpha \neq 0, -1, -2, \cdots, \\ &\text{(iii)} \quad \operatorname{Im}(-\beta) \cdot \operatorname{Im}(-\delta) < 0, \\ &\text{(iv)} \quad \operatorname{Re}(\alpha+\beta+\gamma+\delta) > 1. \end{aligned} \tag{3.19}$$

The weight function to be considered in this case is

$$w(x)=\frac{\Gamma(x-\beta)\Gamma(x-\delta)}{\Gamma(\alpha+x)\Gamma(\gamma+x)}, \qquad -\infty<x<\infty, \tag{3.20}$$

which appears closer to $w^{(1)}(x)$ in (2.5) than $v(x)$ does. The total weight corresponding to $w(x)$ is

$$\mu=\int_{-\infty}^{\infty} w(x)\,dx=\frac{\pm 2\pi^2 i\Gamma(\alpha+\beta+\gamma+\delta-1)}{\sin[\pi(\beta-\delta)]\Gamma(\alpha+\beta)\Gamma(\beta+\gamma)\Gamma(\gamma+\delta)\Gamma(\delta+\alpha)}, \tag{3.21}$$

$\pm$, according as $\operatorname{Im}\beta \lesseqgtr \operatorname{Im}\delta$ [12, p. 300; note a misprint in the conditions following formula (19)].

It can be shown in an analogous manner that the rational functions $R_n(x)$, $S_n(x)$, defined in (3.17) but now with complex parameters satisfying the restrictions (3.19), have the biorthogonality relation

$$\int_{-\infty}^{\infty} w(x)R_n(x)S_m(x)\,dx=\frac{\mu n!}{(\alpha+\beta+\gamma+\delta-1)_n}\delta_{m,n}. \tag{3.22}$$

In the special case $\gamma=\bar{\alpha}$, $\delta=\bar{\beta}$, $S_n(x)$ becomes the complex conjugate of $R_n(x)$ with $w(x)=|\Gamma(x-\beta)/\Gamma(x+\alpha)|^2$ and

$$\mu=\frac{2\pi^2\Gamma(2\operatorname{Re}(\alpha+\beta)-1)}{\sinh(2\pi\operatorname{Im}\beta)}|\Gamma(\alpha+\beta)\Gamma(\alpha+\bar{\beta})|^{-2}. \tag{3.23}$$

4. The case of a balanced ${}_4F_3$. In this section, we shall assume that $k=2$ and

$$b_4=0, \qquad b_1+b_2+b_3+1=n. \tag{4.1}$$

The resulting ${}_4F_3$ series in (1.9) becomes balanced under this assumption. Then, by Whipple's transformation formula (1.10), we get

$$\begin{aligned}
&{}_4F_3\left[\begin{matrix}-n, 1-a_1-x-n, 1-a_3-x-n, a_2+b_2-x-n\\ 1-a_1-b_1-x-n, 1-a_3-b_3-x-n, a_2-x-n\end{matrix}\right]\\
&\quad={}_4F_3\left[\begin{matrix}1-a_3-x-n, a_2+b_2-x-n, 1-a_1-x-n, -n\\ 1-a_3-b_3-x-n, 1-a_1-b_1-x-n, a_2-x-n\end{matrix}\right]\\
&\quad=\frac{(-b_1)_n(a_1+a_2-1)_n}{(a_1+b_1+x)_n(1-a_2+x)_n}\\
&\qquad\cdot{}_4F_3\left[\begin{matrix}-n, 2-a_2-a_3+b_1-n, 1-a_1-x-n, b_1+b_2+1-n\\ b_1+1-n, 2-a_1-a_2-n, 2+b_1+b_2-a_3-x-2n\end{matrix}\right].
\end{aligned} \tag{4.2}$$

If we replace a_1, b_1, b_2 by a_1-n, b_1+n and b_2+n, respectively, we obtain

$$\begin{aligned}
L_n(x)&\equiv L_n(x; a_1-n, a_2, a_3; b_1+n, b_2+n, -b_1-b_2-1-n)\\
&=\Delta_x^n[(a_1-n+x)_{b_1+n}(a_2-x)_{b_2+n}(a_3+x)_{-b_1-b_2-1-n}]\\
&=\frac{w^{(2)}(x)(-1)^n(b_1+1)_n(2-a_1-a_2)_n(a_3+x)}{(a_3+x)_n}\\
&\qquad\cdot{}_4F_3\left[\begin{matrix}-n, n+b_1+b_2+1, 1-a_1-x, 2+b_1-a_2-a_3\\ b_1+1, 2-a_1-a_2, 2+b_1+b_2-a_3-x\end{matrix}\right],
\end{aligned} \tag{4.3}$$

where

$$w^{(2)}(x)=(a_1+x)_{b_1}(a_2-x)_{b_2}(a_3+x+1)_{-b_1-b_2-2}. \tag{4.4}$$

Replacing a_3 by a_3+1 in (4.3) and observing that

$$
\begin{aligned}
&{}_4F_3\left[\begin{matrix}-n,\ n+b_1+b_2+1,\ 1-a_1-x,\ 1+b_1-a_2-a_3\\ b_1+1,\ 2-a_1-a_2,\ 1+b_1+b_2-a_3-x\end{matrix}\right]\\
&\quad=\frac{(b_2+1)_n(a_3+x+1)_n}{(b_1+1)_n(1+b_1+b_2-a_3-x)_n}\\
&\qquad\cdot{}_4F_3\left[\begin{matrix}-n,\ n+b_1+b_2+1,\ 1-a_2+x,\ 1+a_3-a_1-b_1\\ b_2+1,\ 2-a_1-a_2,\ a_3+x+1\end{matrix}\right],
\end{aligned}
$$

we obtain a second formula,

$$
\begin{aligned}
L_n'(x)&\equiv L_n(x;\,a_1-n,\,a_2,\,a_3+1;\,b_1+n,\,b_2+n,\,-b_1-b_2-1-n)\\
&=\Delta_x^n[(a_1-n+x)_{b_1+n}(a_2-x)_{b_2+n}(a_3+x+1)_{-b_1-b_2-1-n}]\\
&=\{(a_1+x)_{b_1}(a_2-x)_{b_2}(a_3+x+1)_{-b-b_2-2}\}(-1)^{n+1}(b_2+1)_n(2-a_1-a_2)_n\\
&\quad\cdot\frac{(1+b_1+b_2-a_3-x)}{(1+b_1+b_2-a_3-x)_n}\,{}_4F_3\left[\begin{matrix}-n,\ n+b_1+b_2+1,\ 1-a_2+x,\ 1+a_3-a_1-b_1\\ b_2+1,\ 2-a_1-a_2,\ a_3+x+1\end{matrix}\right]
\end{aligned}
\tag{4.5}
$$

We have thus found the Rodrigues formula for two distinct families of rational functions $R_n^{(2)}(x)$, $S_n^{(2)}(x)$ defined by

$$
R_n^{(2)}(x)={}_4F_3\left[\begin{matrix}-n,\ n+b_1+b_2+1,\ 1-a_1-x,\ 2+b_1-a_2-a_3\\ b_1+1,\ 2-a_1-a_2,\ 2+b_1+b_2-a_3-x\end{matrix}\right],
\tag{4.6}
$$

$$
S_n^{(2)}(x)={}_4F_3\left[\begin{matrix}-n,\ n+b_1+b_2+1,\ 1-a_2+x,\ 1+a_3-a_1-b_1\\ b_2+1,\ 2-a_1-a_2,\ a_3+x+1\end{matrix}\right],
\tag{4.7}
$$

$n=0,1,2,\cdots$. Formulas (4.2) and (4.3) then read

$$
\begin{aligned}
&\Delta_x^n\left[(a_1-n+x)_{b_1+n}(a_2-x)_{b_2+n}(a_3+x)_{-b_1-b_2-1-n}\right]\\
&\quad=\frac{w^{(2)}(x)(-1)^n(b_1+1)_n(2-a_1-a_2)_nR_n^{(2)}(x)(a_3+x)}{(a_3+x)_n},
\end{aligned}
\tag{4.8}
$$

$$
\begin{aligned}
&\Delta_x^n\left[(a_1-n+x)_{b_1+n}(a_2-x)_{b_2+n}(a_3+x+1)_{-b_1-b_2-1-n}\right]\\
&\quad=\frac{w^{(2)}(x)(-1)^{n+1}(b_2+1)_n(2-a_1-a_2)_nS_n^{(2)}(x)(1+b_1+b_2-a_3-x)}{(1+b_1+b_2-a_3-x)_n}.
\end{aligned}
\tag{4.9}
$$

5. Biorthogonality of $R_n^{(2)}(x)$ and $S_n^{(2)}(x)$. In order to establish the conditions under which the rational functions $R_n^{(2)}(x)$ and $S_n^{(2)}(x)$ are biorthogonal with respect to the weight function $w^{(2)}(x)$, we first observe that for large $|x|$ the difference function

$$
\Delta_x^{n-k-1}[(a_1-n+x+k)_{b_1+n}(a_2-x-k)_{b_2+n}(a_3+x+k)_{-b_1-b_1-1-n}],
\tag{5.1}
$$

is of the order $|x|^k$, $k=0,1,\cdots,n-1$, and hence cannot vanish, in general, as $|x|\to\infty$. This implies that for arbitrary finite values of the parameters a_1, a_2, a_3, b_1 and b_2 the range of x in the summation $\Sigma w^{(2)}(x)R_n^{(2)}(x)S_m^{(2)}(x)$ cannot be from $-\infty$ to ∞. In order for this sum to lead to a constant multiple of $\delta_{m,n}$ one must be able to do the summation by parts through the use of Rodrigues' formulas (4.8) and (4.9), and at each successive summation one would require the vanishing of a difference function of the form (5.1) at both the lower and upper limits (to be more precise, if the upper limit is a finite number

N the vanishing is required at $x = N+1$). The only way this vanishing at either end can be achieved is to assume that at least two of the parameters a_1, a_2, a_3 are integers.

Let us assume that M and N are nonnegative integers, and

$$a_1 - 1 = M \quad \text{and} \quad a_2 - 1 = N. \tag{5.2}$$

Let us consider the sum

$$P_{m,n} = \sum_{x=-M}^{x=N} w^{(2)}(x) R_n^{(2)}(x) S_m^{(2)}(x).$$

By (4.4), (4.5) and (4.6) we have

$$P_{m,n} = \sum_{k=0}^{n} \frac{(-n)_k (n+b_1+b_2+1)_k (1+b_1-a_3-N)_k}{k!(b_1+1)_k(-M-N)_k} \sum_{l=0}^{m} \frac{(-m)_l (m+b_1+b_2+1)_l}{l!(b_2+1)_l} \frac{(a_3-b_1-M)_l}{(-M-N)_l} Q_{k,l}, \tag{5.3}$$

where

$$Q_{k,l} = \sum_{x=-M}^{N} (M+1+x)_{b_1} (N-x+1)_{b_2} \cdot (a_3+x+1)_{-b_1-b_2-2} \frac{(-M-x)_k (x-N)_l}{(2+b_1+b_2-a_3-x)_k (a_3+x+1)_l}. \tag{5.4}$$

Note that the sum on the r.h.s. of (5.4) vanishes unless $M+x \geqq k$ and $N-x \geqq l$. Transforming the summation variable x to y by setting $x+M = y+k$ and then simplifying the Pochhammer products, we obtain

$$Q_{k,l} = \frac{\Gamma(b_1+1)\Gamma(M+N+b_2+1)\Gamma(a_3-b_1-b_2-1-M)}{\Gamma(M+N+1)\Gamma(a_3-M+1)} \cdot \frac{(b_1+1)_k (-M-N)_{k+l}}{(-M-N-b_2)_k (a_3-M+1)_{k+l}} \cdot {}_3F_2\left[\begin{matrix} k+l-M-N, b_1+1+k, a_3-b_1-b_2-1-M \\ k-M-N-b_2, a_3-M+1+k+l \end{matrix}\right]. \tag{5.5}$$

The ${}_3F_2$ function on the right is a terminating balanced series with argument 1 and hence can be summed by Pfaff–Saalschutz theorem. Thus

$$\begin{aligned} &{}_3F_2\left[\begin{matrix} k+l-M-N, b_1+1+k, a_3-b_1-b_2-1-M \\ k-M-N-b_1, a_3-M+1+k+l \end{matrix}\right] \\ &= \frac{(-M-N-b_1-b_2-1)_{M+N-k-l} (b_1+1+k-N-a_3)_{M+N-k-l}}{(k-b_2-M-N)_{M+N-k-l} (-N-a_3)_{M+N-k-l}} \\ &= \frac{(b_1+b_2+2)_{M+N} (a_3-b_1-M)_{M+N}}{(a_3+1-M)_{M+N} (b_2+1)_{M+N}} \cdot \frac{(a_3+1-M)_{k+l} (b_2+1)_l (-M-N-b_2)_k}{(b_1+b_2+2)_{k+l} (a_3-b_1-M)_l (b_1+1-a_3-N)_k}. \end{aligned} \tag{5.6}$$

Using this in (5.5) we get

$$Q_{k,l}=\frac{\Gamma(b_1+1)\Gamma(b_2+1)}{\Gamma(b_1+b_2+2)}\cdot\frac{\Gamma(b_1+b_2+2+M+N)\Gamma(a_3-b_1+N)\Gamma(a_3-b_1-b_2-1-M)}{\Gamma(M+N+1)\Gamma(a_3-b_1-M)\Gamma(a_3+1+N)}$$
$$\cdot\frac{(b_1+1)_k(b_2+1)_l(-M-N)_{k+l}}{(b_1+b_2+2)_{k+l}(a_3-b_1-M)_l(b_1+1-a_3-N)_k}. \tag{5.7}$$

Hence, we obtain

$$P_{m,n}=\mu^{(2)}\sum_{k=0}^{n}\frac{(-n)_k(n+b_1+b_2+1)_k}{k!(b_1+b_2+2)_k}\,{}_3F_2\left[\begin{matrix}-m,\ m+b_1+b_2+1,\ k-M-N\\ b_1+b_2+2+k,\ -M-N\end{matrix}\right], \tag{5.8}$$

where

$$\mu^{(2)}=\frac{\Gamma(b_1+1)\Gamma(b_2+1)}{\Gamma(b_1+b_2+2)}$$
$$\cdot\frac{\Gamma(b_1+b_2+2+M+N)\Gamma(a_3-b_1+N)\Gamma(a_3-b_1-b_2-1-M)}{\Gamma(M+N+1)\Gamma(a_3+1+N)\Gamma(a_3-b_1-M)} \tag{5.9}$$
$$=\sum_{x=-M}^{N}w^{(2)}.$$

Observing that the ${}_3F_2$ series on the right of (5.8) is balanced, we apply (5.6) once again, and obtain

$$P_{m,n}=\mu^{(2)}\frac{(b_1+b_2+2+M+N)_m}{(-M-N)_m(b_1+b_2+2)_m}\sum_{k=0}^{n}\frac{(-n)_k(n+b_1+b_2+1)_k(-k)_m}{k!(2+b_1+b_2+m)_k}.$$

The series on the right obviously vanishes if $n<m$. Hence, for $n\geqq m$ we have

$$P_{m,n}=\mu^{(2)}\frac{(b_1+b_2+2+M+N)_m(-n)_m(n+b_1+b_2+1)_m(-1)^m}{(-M-N)_m(b_1+b_2+2)_{2m}}$$
$$\cdot\sum_{k=0}^{n-m}\frac{(m-n)_k(n+m+b_1+b_2+1)_k}{k!(2+b_1+b_2+2m)_k}$$
$$=\mu^{(2)}\frac{(b_1+b_2+2+M+N)_m(-n)_m(n+b_1+b_2+1)_m(-1)^m}{(-M-N)_m(b_1+b_2+2)_{2m}} \tag{5.10}$$
$$\cdot\frac{(1-n+m)_{n-m}}{(2+b_1+b_2+2m)_{n-m}},$$

by the Chu–Vandermonde theorem. But this vanishes unless $n=m$. Hence,

$$P_{m,n}=\mu^{(2)}\frac{(b_1+b_2+2+M+N)_n(n+b_1+b_2+1)_n(-n)_n(-1)^n}{(-M-N)_n(b_1+b_2+2)_{2n}}\delta_{m,n}, \tag{5.11}$$

which proves the biorthogonality of $R_n^{(2)}(x)$ and $S_n^{(2)}(x)$. The conditions on b_1, b_2, a_3 must be such that $\mu^{(2)}$ remains finite. In particular, if $b_1>-1$, $b_2>1$, then $\mu^{(2)}$ is finite and positive if $a_3>b_1+b_2+1+M$.

As in the case of $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$ it is instructive to see how the biorthogonality of $R_n^{(2)}(x)$ and $S_n^{(2)}(x)$ follows from their Rodrigues' formulas. Assuming that $n>m\geqq 0$

and using (4.4), (4.6), (4.8), we obtain

$$P_{m,n} = \frac{(-1)^n}{(b_1+1)_n(-M-N)_n} \sum_{x=-M}^{N} \frac{(a_3+x)_n}{a_3+x} S_m^{(2)}(x)$$
$$\cdot \Delta_x^n[(x+M-n+1)_{b_1+n}(N-x+1)_{b_2+n}(a_3+x)_{-b_1-b_2-1-n}]$$
$$= \frac{1}{(b_1+1)_n(-M-N)_n} \sum_{x=-M}^{N} (x+M+1)_{b_1+n}(N-x-n+1)_{b_2+n} \tag{5.12}$$
$$\cdot (a_3+x+n)_{-b_1-b_2-1-n}$$
$$\cdot \Delta_x^n\left\{\frac{(a_3+x)_n}{a_3+x} {}_4F_3\left[\begin{matrix}-m, m+b_1+b_2+1, x-N, a_3-b_1-M \\ b_2+1, -M-N, a_3+x+1\end{matrix}\right]\right\}.$$

The difference operator Δ_x^n is operating on a polynomial of degree $n-1$, and hence we get a zero on the r.h.s. Similarly, for $m > n \geqq 0$ we use (4.5), and (4.9) and find that $P_{m,n} = 0$. Thus, the Rodrigues' formulas (4.7) and (4.8) directly lead to the biorthogonality relation

$$\sum_{x=-M}^{N} w^{(2)}(x) R_n^{(2)}(x) S_m^{(2)}(x) = 0, \qquad m \neq n.$$

Unlike $R_n^{(1)}(x)$ and $S_n^{(1)}(x)$, the functions $R_n^{(2)}(x)$, $S_n^{(2)}(x)$ do not coincide for any finite choice of the parameters. However, if we let $a_3 \to \pm\infty$ then both $R_n^{(2)}(x)$ and $S_n^{(2)}(x)$ reduce to polynomials which can be shown to be linearly dependent. For, with $M = 0$, we get

$$\lim_{|a_3|\to\infty} R_n^{(2)}(x) = {}_3F_2\left[\begin{matrix}-n, n+b_1+b_2+1, -x \\ b_1+1, -N\end{matrix}\right] = Q_n(x; b_1, b_2, N)), \tag{5.13}$$

$$\lim_{|a_3|\to\infty} S_n^{(2)}(x) = {}_3F_2\left[\begin{matrix}-n, n+b_1+b_2+1, x-N \\ b_2+1, -N\end{matrix}\right]$$
$$= \frac{(-1)^n(b_1+1)_n}{(b_2+1)_n} {}_3F_2\left[\begin{matrix}-n, n+b_1+b_2+1, -x \\ b_1+1, -N\end{matrix}\right] \tag{5.14}$$
$$= \frac{(-1)^n(b_1+1)_n}{(b_2+1)_n} Q_n(x; b_1, b_2, N).$$

Thus, in the limit $|a_3| \to \infty$, the biorthogonal rational functions $R_n^{(2)}$ and $S_n^{(2)}$ reduce to the Hahn polynomials. These were the functions that we applied in [17] to obtain an addition theorem for Hahn polynomials.

Addendum. At the time of writing this paper it was brought to the author's attention that James Wilson [23] found a general system of biorthogonal rational functions of which $R_n^{(2)}(x)$, $S_n^{(2)}(x)$ are special cases. He also worked out the continuous analogues of these functions.

REFERENCES

[1] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] W. N. BAILEY, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[3] P. L. CHEBYSHEV, *Sur l'interpolation des valeurs equidistantes*, in Collected Works, Chelsea, New York, 1961, pp. 217–242.
[4] R. D. COOPER, M. R. HOARE AND MIZAN RAHMAN, *Stochastic processes and special functions: On the probabilistic origin of some positive kernels associated with classical orthogonal polynomials*, J. Math. Anal. Appl., 61 (1977), pp. 262–291.
[5] P. DELSARTE, *An Algebraic Approach to the Association Schemes of Coding Theory*, Philips Research Reports Supplements, 10, 1973.
[6] ———, *The association schemes of coding theory*, in Combinatorics, M. Hall, Jr., and J. H. Van Lint, eds., Math. Centre Tracts, 55, Math. Centre, Amsterdam, 1974, pp. 139–157.
[7] C. DUNKL, *A. Krawtchouk polynomial addition theorem and wreath products of symmetric groups*, Indiana Univ. Math. J., 25 (1976), pp. 335–358.
[8] ———, *Spherical functions on compact groups and applications to special functions*, Symposia Mathematica, 22 (1977), pp. 145–161.
[9] C. DUNKL AND D. RAMIREZ, *Krawtchouk polynomials and the symmetrization of hypergroups*, this Journal, 5 (1974), pp. 351–366.
[10] G. EAGLESON, *A characterization for positive definite sequences on the Krawtchouk polynomials*, Austral. J. Statist. 11 (1969), pp. 29–38.
[11] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, eds., *Higher Transcendental Functions*, Vol. II, Bateman Manuscript Project, McGraw-Hill, New York, 1953.
[12] ———, *Tables of Integral Transforms*, Vol. II, Bateman Manuscript Project, McGraw-Hill, New York, 1954.
[13] G. GASPER, *Nonnegativity of a discrete Poisson kernel for the Hahn polynomials*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.
[14] ———, *Projection formulas for orthogonal polynomials of a discrete variable*, J. Math. Anal. Appl., 45 (1974), pp. 176–198.
[15] S. KARLIN AND J. L. MCGREGOR, *The Hahn polynomials, formulas and an application*, Scripta Math., XXVI (1961), pp. 33–46.
[16] ———, *Linear growth models with many types and multidimensional Hahn polynomials*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975.
[17] MIZAN RAHMAN, *Product and addition theorems for Hahn polynomials*, submitted for publication.
[18] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.
[19] N. J. A. SLOANE, *An introduction to association schemes and coding theory*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975.
[20] G. SZEGO, *Orthogonal Polynomials*, Amer. Math. Soc. Colloquium Publ., Vol. XXIII, 4th edn., American Mathematical Society, Providence RI, 1975.
[21] L. TOSCANO, *I polinomi ipergeometrici nel calcolo delle differenze finite*, Boll. Un. Mat. Ital. (3), 4 (1949), pp. 398–409.
[22] M. WEBER AND A. ERDELYI, *On the finite difference analogue of Rodrigues' formula*, Amer. Math. Monthly, 59 (1952), pp. 163–168.
[23] JAMES WILSON, private communication.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0007 $01.00/0

ON INVERTIBILITY OF LINEAR ORDINARY DIFFERENTIAL BOUNDARY VALUE PROBLEMS*

JAMES S. MULDOWNEY†

Abstract. A criterion is given which is necessary and sufficient that certain homogeneous linear boundary value problems have only the trivial solution. The condition includes results of Pólya [Trans. Amer. Math. Soc. 24 (1922), pp. 312–324] on disconjugacy and Muldowney [Proc. Amer. Math. Soc., 74 (1979), pp. 49–55] on disfocality. A mean value theorem and a positivity result are also obtained. Analogues of the Sturm comparison principle are established including a generalization of a theorem of Hartman [Amer. J. Math., 91 (1969), pp. 306–362; 93 (1971), pp. 439–451] and Levin [Soviet Math. Dokl. 4 (1963), pp. 121–124] which gives a necessary and sufficient condition for invertibility in terms of the existence of solutions to a family of differential inequalities.

1. Introduction.

The nth order linear differential operator L defined by

$$Ly = y^{(n)} + a(t)y^{(n-1)} + \cdots + a_n(t)y,$$

in which the coefficients a_i are continuous functions, is said to be *disconjugate* on an interval I if the only solution of $Ly = 0$ having n zeros or more in I, counting multiplicities, is the zero solution. Further, L is *right disfocal* on I if it satisfies the more restrictive condition that the zero solution is the only one which satisfies $y^{(i-1)}(t_i) = 0$, $i = 1, \cdots, n$, for any set of points $t_i \in I$, $t_1 \leqq t_2 \leqq \cdots \leqq t_n$. L is *left disfocal* on I if $t_1 \geqq t_2 \geqq \cdots \geqq t_n$ in the preceding statement and is *disfocal* on I if no particular order needs to be imposed on the points t_i.

It was shown by Pólya [16] that a necessary and sufficient condition for the disconjugacy of L on a compact interval I is the existence of a family of solutions $u_1, \cdots, u_{n-1}$ of $Ly = 0$ satisfying

$$W(u_1, \cdots, u_k) > 0, \qquad k = 1, \cdots, n-1, \tag{1.1}$$

on I, where W is the Wronskian determinant $\det [u_i^{(j-1)}]\, i, j = 1, \cdots, k$. It was shown in [6] that the existence of $u_1, \cdots, u_{n-1}$ satisfying $Ly = 0$ and

$$W(u_1^{(j-1)}, \cdots, u_k^{(j-1)}) > 0, \qquad j = 1, \cdots, n-k+1, \quad k = 1, \cdots, n-1 \tag{1.2}$$

is necessary and sufficient for right disfocality of L on I.

The present paper considers conditions which are more restrictive than (1.1) and either more or less restrictive than (1.2). Necessary and sufficient conditions are obtained that $y = 0$ be the only solution of $Ly = 0$ which, counting multiplicities, has m_1 zeros in I followed by m_2 zeros of y', m_3 zeros of y'' and so forth, for certain sequences of nonnegative integers m_i. For example, it is proved that $y = 0$ is the only solution of $Ly = 0$ which has $n-1$ zeros in I followed by a single zero of y' if and only if there are solutions $u_1, \cdots, u_{n-1}$ satisfying (1.1) and $u_1' > 0$ on I. The conditions (1.1) and $u_1' > 0$, $W(u_1', u_2') > 0$ are necessary and sufficient that $y = 0$ is the only solution which has $n-2$ zeros on I followed by two zeros of y'. Operators with these properties are shown to have a positivity analogous to Čaplygin's inequality, and from this a generalization of the Pólya mean value theorem is deduced. The nature of solutions to these boundary value problems on minimal intervals of noninvertibility is also investigated. Finally, several comparison criteria for invertibility are obtained.

* Received by the editors July 19, 1979, and in revised form September 22, 1980. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A 7197.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

2. Notation and definitions. *Zeros.* Let $\tau = (t_1, \cdots, t_n)$; a function y has n zeros at τ if $y(t_i) = 0$, $i = 1, \cdots, n$, when the points t_i are all distinct, and $y(t_i) = y'(t_i) = \cdots = y^{(m-1)}(t_i) = 0$ if a point t_i occurs m times in τ.

Partitions. Let $t_1, \cdots, t_n$ be real numbers. A partition $(\tau_1; \tau_2; \cdots; \tau_l)$ of the ordered n-tuple $(t_1, \cdots, t_n)$ is generated by inserting $l-1$ semicolons instead of commas in this expression. The insertion of two or more adjacent (i.e., not separated by entries t_i) semicolons is not ruled out. The ordered set of entries $(t_j, \cdots, t_k)$ between the $(i-1)$th and ith semicolons is denoted τ_i, and $|\tau_i| = k - j + 1$ is the number of entries in τ_i. If the $(i-1)$th and ith semicolons are adjacent, then $\tau_i = \phi$, $|\tau_i| = 0$. For example, $(\tau_1; \tau_2; \tau_3; \tau_4) = (t_1, t_2; t_3; ; t_4, t_5, t_6)$ where $\tau_1 = (t_1, t_2)$, $\tau_2 = (t_3)$, $\tau_3 = \phi$, $\tau_4 = (t_4, t_5, t_6)$.

An entry t in τ_i has multiplicity m in τ_i if there are m entries equal to t in τ_i.

Finally, $(\tau_1; \cdots; \tau_l)$ is an increasing partition if $t_1 \leqq t_2 \leqq \cdots \leqq t_n$ and $t \in \tau_i$, $s \in \tau_j$ and $i < j$ implies either $t < s$ or $t = s$ and $i + m \leqq j$, where m is the multiplicity of t in τ_i. For example, the partition $(1, 2, 2; 2, 3)$ is not increasing, while $(1, 2; 2, 2, 3)$ and $(1, 2, 2; ; 2, 3)$ are increasing. The partition is decreasing if the partition obtained by replacing each entry by its negative is increasing.

Invertibility. Let $m_1, \cdots, m_l \geqq 0$ be integers such that $m_1 + \cdots + m_l = n$. The operator L is right-$(m_1, \cdots, m_l)$-invertible on an interval I if, for each increasing partition $(\tau_1; \cdots; \tau_l)$ of n points in I, the only solution of $Ly = 0$ such that $y^{(j-1)}$ has m_j zeros at τ_j, $j = 1, \cdots, l$, is the zero solution. The operator is left-$(m_1, \cdots, m_l)$-invertible if "increasing" may be replaced by "decreasing" in the preceding statement. It is $(m_1, \cdots, m_l)$-invertible if the qualification "increasing" or "decreasing" may be omitted.

Thus, for example, L is (n)-invertible means that it is disconjugate, and L is right-$(1, \cdots, 1)$-invertible means that it is right disfocal.

The functions Ω, Ω_j. Let $u_1, \cdots, u_n$ be real-valued functions containing the points $t_1, \cdots, t_n$ in their domains and let $(\tau_1; \cdots; \tau_l)$ be a partition of $(t_1, \cdots, t_n)$. The function $\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l)$ is defined to be the determinant of the $n \times n$ matrix A described as follows. The rows of A numbered $|\tau_1| + \cdots + |\tau_{i-1}| + 1$ through $|\tau_1| + \cdots + |\tau_{i-1}| + |\tau_i|$ form a block of dimension $r \times n$, $r = |\tau_i|$, the jth column of which is

$$\operatorname{col}\,[u_j^{(i-1)}(s_1), u_j^{(i-1)}(s_2), \cdots, u_j^{(i-1)}(s_r)]$$

if $\tau_i = (s_1, s_2, \cdots, s_r)$ and s_i are all distinct. If an element s of τ_i has multiplicilty m, then the corresponding m entries $u_j^{(i-1)}(s), \cdots, u_j^{(i-1)}(s)$ are replaced by $u_j^{(i-1)}(s)$, $u_j^{(i)}(s), \cdots, u_j^{(i+m-2)}(s)$.

For example, the Wronskian determinant $W(u_1, \cdots, u_n)(t) = \det\,[u_i^{(j-1)}(t)]$, $i, j = 1, \cdots, n$, may be denoted $\Omega(u_1, \cdots, u_n)(t, \cdots, t)$ or $\Omega(u_1, \cdots, u_n)(t; \cdots; t)$. As a further example, $\Omega(u_1, u_2, u_3)(t_1, t_2; t_3)$ denotes

$$\begin{vmatrix} u_1(t_1) & u_2(t_1) & u_3(t_1) \\ u_1(t_2) & u_2(t_2) & u_3(t_2) \\ u_1'(t_3) & u_2'(t_3) & u_3'(t_3) \end{vmatrix}, \qquad \begin{vmatrix} u_1(t_1) & u_2(t_1) & u_3(t_1) \\ u_1'(t_1) & u_2'(t_1) & u_3'(t_2) \\ u_1'(t_3) & u_2'(t_3) & u_3'(t_3) \end{vmatrix},$$

when $t_1 \neq t_2$, $t_1 = t_2$, respectively.

The operator L is right(left)-$(m_1, \cdots, m_l)$-invertible on I if and only if $\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) \neq 0$ for every increasing (decreasing) partition $(\tau_1, \cdots; \tau_l)$ of n points in I with $|\tau_i| = m_i$ where $u_1, \cdots, u_n$ is a fundamental solution set of $Ly = 0$ and is invertible if the qualifications "right" "left", "increasing", "decreasing" are omitted.

The functions Ω_j are defined by

$$\Omega_1(u_1, \cdots, u_n)(\tau: t) = \Omega(u_1, \cdots, u_n)(\tau, t, \cdots, t),$$

where $|(\tau, t, \cdots, t)| = n$, and

$$\Omega_j(u_1, \cdots, u_n)(\tau: t) = \Omega(u_1, \cdots, u_n)(\tau_1: \tau_2; \cdots; \tau_j),$$

where $\tau_1 = \tau$, $\tau_i = \phi$, $1 < i < j$ and $\tau_j = (t, \cdots, t)$, $|\tau_j| = n - |\tau|$, if $j > 1$. Thus $\Omega_j(u_1, \cdots, u_n)(\tau : t)$ is the determinant of the matrix for which the kth column is

$$\begin{aligned}
&\operatorname{col}[u_k^{(j-1)}(t), \cdots, u_k^{(j+n-2)}(t)] && \text{if } \tau = \phi,\\
&\operatorname{col}[u_k(t_1), u_k^{(j-1)}(t), \cdots, u_k^{(j+n-3)}(t)] && \text{if } \tau = (t_1),\\
&\operatorname{col}[u_k(t_1), u_k'(t_1), u_k^{(j-1)}(t), \cdots, u_k^{(j+n-4)}(t)] && \text{if } \tau = (t_1, t_1),\\
&\operatorname{col}[u_k(t_1), u_k(t_2), u_k^{(j-1)}(t), \cdots, u_k^{(j+n-4)}(t)] && \text{if } \tau = (t_1, t_2), \quad t_1 \neq t_2,
\end{aligned}$$

and so forth.

Descartes systems. A system of functions $(u_1, \cdots, u_n)$ is Descartes on an interval if the Wronskians $W(u_{i_1}, \cdots, u_{i_k})$ are positive on the interval for each increasing subsequence $\{i_j\}$ of $\{1, \cdots, n\}$.

Properties I *and* I'. Let $\{n_j\}$ be integers such that

$$n = n_1 = n_2 = \cdots = n_r > n_{r+1} > \cdots > n_l > n_{l+1} = 0, \qquad l \geqq r \geqq 1. \tag{2.1}$$

A system of functions $(u_1, \cdots, u_n)$ has Strict Property I on an interval if

$$W(u_1^{(j-1)}, \cdots, u_k^{(j-1)}) > 0, \qquad k = 1, \cdots, n_j, \quad j = 1, \cdots, l. \tag{2.2}$$

The system has Property I if (2.2) holds on the interval except in the cases $k = n_{j+1} + 1, \cdots, n_j, j = r+1, \cdots, l$ when the Wronskians may be nonnegative. The system has (Strict) Property I' if $(u_n, -u_{n-1}, \cdots, (-1)^{n-1}u_1)$ has (Strict) Property I; that is, for the values of k and j above, the Wronskians $W(u_{n-k+1}^{(j-1)}, \cdots, u_n^{(j-1)})$ are positive or nonnegative as before.

For example, if $n = n_1 = 4$, $n_2 = 3$, $n_3 = 1$ $(r = 1, l = 3)$, then the system (u_1, u_2, u_3, u_4) has Property I if

$$\begin{aligned}
&u_1 > 0, \quad W(u_1, u_2) > 0, \quad W(u_1, u_2, u_3) > 0, \quad W(u_1, u_2, u_3, u_4) > 0,\\
&u_1' > 0, \quad W(u_1', u_2') \geqq 0, \quad W(u_1', u_2', u_3') \geqq 0,\\
&u_1'' \geqq 0,
\end{aligned}$$

and it has Strict Property I if all inequalities are strict.

3. Invertibility conditions, positivity results and extremal intervals. As an alternative proof of his mean value theorem, Pólya states and sketches a proof of a result [16, Thm. V] which is equivalent to Lemma 1. The details of Pólya's proof are in [6, pp. 376–378].

LEMMA 1. *Let $\tau = (t_1, \cdots, t_n)$ be a nondecreasing sequence of points in an interval I and $u_1, \cdots, u_n$ be functions having $n-1$ derivatives existing on I. If $W(u_1, \cdots, u_k) > 0$, $k = 1, \cdots, n-1$, $W(u_1, \cdots, u_n) \geqq 0$ on I and $W(u_1, \cdots u_n)(s) \neq 0$ for some $s \in [t_1, t_n]$,* then $\Omega(u_1, \cdots, u_n)(\tau) > 0$.

LEMMA 2. *Let* $\tau = (t_1, \cdots, t_q)$ *and* $\tau' = (t_1, \cdots, t_{q-1})$ *if* $q > 1$, $\tau' = \phi$ *if* $q = 1$. *Then*

$$\begin{aligned}\Omega_j(u_1, \cdots, u_k)(\tau\colon t)\Omega_{j-1}(u_1, \cdots, u_{k-1})(\tau'\colon t)\\ = \Omega_{j-1}(u_1, \cdots, u_k)(\tau'\colon t)\Omega_j(u, \cdots, u_{k-1})(\tau\colon t)\\ + \Omega_{j-1}(u_1, \cdots, u_k)(\tau\colon t)\Omega_j(u, \cdots, u_{k-1})(\tau'\colon t)\end{aligned}$$

for any functions $u_1, \cdots, u_k$ *for which these expressions exist.*

In the case $\tau = (t_1)$, this identity is [14, Lemma 2] with j replaced by $j-1$.

The proof of Lemma 2 is based on the identities (3.1), (3.2). If A_1, B_1, A, B, C, D are any real numbers, then

$$\begin{vmatrix} A_1 & B_1 \\ C & D \end{vmatrix} A = \begin{vmatrix} A & B \\ C & D \end{vmatrix} A_1 + \begin{vmatrix} A_1 & B_1 \\ A & B \end{vmatrix} C, \tag{3.1}$$

which is equivalent to the identity

$$\begin{vmatrix} A_1 & A_1 & B_1 \\ A & A & B \\ C & C & D \end{vmatrix} = 0.$$

If $a_{r_1 \cdots r_m}^{s_1 \cdots s_m}$ denotes the minor of the $k \times k$ matrix $[a_i^j]$ determined by the rows $r_1, \cdots, r_m$ and the columns $s_1, \cdots, s_m$ and if $b_i^j = a_{12\cdots p,p+i}^{12\cdots p,p+j}$, then

$$a_{12\cdots k}^{12\cdots k}(a_{12\cdots p}^{12\cdots p})^{k-p-1} = b_{12\cdots k-p}^{12\cdots k-p}, \qquad p = 1, \cdots, k-1. \tag{3.2}$$

This is Sylvester's identity; cf. [3, p. 32].

Choosing $a_{12\cdots k}^{12\cdots k} = \Omega_j(u_1, \cdots, u_k)(\tau\colon t)$ and $a_{12\cdots k-2}^{12\cdots k-2} = \Omega_j(u_1, \cdots, u_{k-2})(\tau'\colon t)$ with $p = k-2$, we obtain from (3.2) that

$$\Omega_j(u_1, \cdots, u_k)(\tau\colon t)\Omega_j(u_1, \cdots, u_{k-2})(\tau'\colon t) = \begin{vmatrix} A_1 & B_1 \\ C & D \end{vmatrix}, \tag{3.3}$$

where

$$\begin{aligned} A_1 &= \Omega_j(u_1, \cdots, u_{k-1})(\tau\colon t), & B_1 &= \Omega_j(u_1, \cdots, u_{k-2}, u_k)(\tau\colon t),\\ C &= \Omega_j(u_1, \cdots, u_{k-1})(\tau'\colon t), & D &= \Omega_j(u_1, \cdots, u_{k-2}, u_k)(\tau'\colon t).\end{aligned}$$

Now, if

$$A = \Omega_{j-1}(u_1, \cdots, u_{k-1})(\tau'\colon t), \qquad B = \Omega_{j-1}(u_1, \cdots, u_{k-2}, u_k)(\tau'\colon t),$$

then, from (3.2),

$$\Omega_{j-1}(u_1, \cdots, u_k)(\tau'\colon t)\Omega_j(u_1, \cdots, u_{k-2})(\tau'\colon t) = \begin{vmatrix} A & B \\ C & D \end{vmatrix}, \tag{3.4}$$

$$\Omega_{j-1}(u_1, \cdots, u_k)(\tau\colon t)\Omega_j(u_1, \cdots, u_{k-2})(\tau'\colon t) = \begin{vmatrix} A_1 & B_1 \\ A & B \end{vmatrix}. \tag{3.5}$$

Lemma 2 follows from (3.1), (3.3), (3.4), (3.5) when $\Omega_j(u_1, \cdots, u_{k-2})(\tau'\colon t) \neq 0$. In the case that this determinant equals zero the result is obtained as the limiting behavior of slight perturbations of $u_1, \cdots, u_k$ for which the determinant is nonzero.

LEMMA 3. *A necessary and sufficient condition that the set of functions* $(u_1, \cdots, u_k)$ *be a Descartes system on an interval* I *is that all Wronskians of consecutive functions* $W(u_p, u_{p+1}, \cdots, u_q)$, $1 \leqq p \leqq q \leqq n$, *be positive on I.*

This is [2, Proposition 4, p. 88].

PROPOSITION 1. *Suppose* $(u_1, \cdots, u_n)$ *has Property* I (*Property* I′) *on an interval* $[a, b]$ *and* $(\tau_1; \cdots; \tau_l)$ *is an increasing partition of n points* $(t_1, \cdots, t_n)$ *in* $[a, b]$ *such that* $\sum_{j \geqq 1} |\tau_j| \leqq n_i$, $i = 1, \cdots, l$. *Then*

$$\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) > 0.$$

Proof. For $l = 1$, Proposition 1 is true for all n, from Lemma 1. The proof is by induction on l. It may be assumed without loss of generality that $|\tau_1| = p$, $0 < p < n$; the induction hypothesis is that the proposition holds for all systems with Property I and increasing partitions $(\tau_1; \cdots; \tau_m)(m \leqq l-1)$. If $s_1 = \max\{s \in \tau_1\}$, it is asserted that the functions $(v_1, \cdots, v_{n-p})$ defined by

$$v_i(t) = \Omega_1(u_1, \cdots, u_p, u_{p+i})(\tau_1 : t), \qquad i = 1, \cdots, n-p$$

are such that Wronskians $W(v_1^{(j-1)}, \cdots, v_k^{(j-1)})$ are all positive on $(s_1, b]$ for $k = 1, \cdots, N_j = \min\{n_j, n-p\}$, $j = 2, \cdots, l$ except in the cases $k = N_{j+1}+1, \cdots, N_j$, $j = s+1, \cdots, l$ (where $s = \max\{j : n_j \geqq n-p\}$) when the Wronskians are nonnegative. Furthermore these Wronskians satisfy the same inequalities on $[s_1, b]$ for $j = m_1+1, \cdots, l$ where m_1 is the multiplicity of s_1 in τ_1. Thus the system $(v_1', \cdots, v_{n-p}')$ has Property I on $(s_1, b]$ with respect to the sequence $N_2, \cdots, N_l$ and $(v_1^{(m_1)}, \cdots, v_{n-p}^{(m_1)})$ has Property I on $[s_1, b]$ with respect to the sequence $N_{m_1+1}, \cdots, N_l$. From the induction hypothesis it now follows that $\Omega(v_1', \cdots, v_{n-p}')(\tau_2; \cdots; \tau_l) > 0$, if $(\tau_2; \cdots; \tau_l)$ is any partition of $n-p$ points in $[s_1, b]$ such that $(\tau_1; \tau_2; \cdots; \tau_l)$ is increasing. But

$$\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l)[\Omega(u_1, \cdots, u_p)(\tau_1)]^{n-p-1} = \Omega(v_1', \cdots, v_{n-p}')(\tau_2; \cdots; \tau_l) > 0,$$

from Sylvester's identity (3.2), and therefore

$$\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) > 0,$$

since $\Omega(u_1, \cdots, u_p)(\tau_1) > 0$ by Lemma 1, proving Proposition 1.

The assertion about the sign of the Wronskians $W(v_1^{(j-1)}, \cdots, v_k^{(j-1)})$ in the preceding proof is based on the identity (from (3.2))

$$W(v_1^{(j-1)}, \cdots, v_k^{(j-1)}) = \Omega_j(u_1, \cdots, u_{p+k})(\tau_1 : t)[\Omega(u_1, \cdots, u_p)(\tau_1)]^{k-1}.$$

Since $\Omega(u_1, \cdots, u_p)(\tau_1) > 0$, by Lemma 1, the Wronskian has the same sign as $\Omega_j(u_1, \cdots, u_{p+k})(\tau_1 : t)$. The assertion now follows from Lemma 2 by induction on j, k, if we recall that

$$\Omega_j(u_1, \cdots, u_k)(\tau : t) = \Omega(u_1, \cdots, u_k)(\tau) > 0 \quad \text{if } k = |\tau|,$$

$$\Omega_1(u_1, \cdots, u_k)(\tau : t) = \Omega(u_1, \cdots, u_k)(\tau, t, \cdots, t) > 0,$$

both from Lemma 1, and

$$\Omega_j(u_1, \cdots, u_k)(\tau : t) = W(u_1^{(j-1)}, \cdots, u_k^{(j-1)})(t) \quad \text{if } \tau = \phi.$$

Let n_j, $j = 1, \cdots, l$ be natural numbers satisfying (2.1) and let m_j, $j = 1, \cdots, l$ be nonnegative integers such that

$$\sum_{j=1}^{l} m_j = n \sum_{j=1}^{l} m_j \leqq n_i, \qquad i = 2, \cdots, l. \tag{3.6}$$

Remark. It is useful to observe that, if $r = 1$ in (2.1), then an operator L is right-$(m_1, \cdots, m_l)$-invertible on an interval, for all $\{m_j\}$ satisfying (3.6), if and only if L is right-$(p_1, \cdots, p_l)$-invertible, where $p_j = n_j - n_{j+1}$, $j = 1, \cdots, l$. This follows from Rolle's theorem.

THEOREM 1. (a) *A sufficient condition that L be right-$(m_1, \cdots, m_l)$-invertible on $[a, b]$ is that there exist solutions $(u_1, \cdots, u_n)$ of $Ly = 0$ having Property* I *(or Property* I′*) on $[a, b]$ with respect to the sequence $\{n_j\}$, $j = 1, \cdots, l$, if $\{m_j\}$ satisfies* (3.6).

(b) *If $r = 1$ in* (2.1) *and L is right-$(m_1, \cdots, m_l)$-invertible on $[a, b]$ for each sequence $\{m_j\}$ satisfying* (3.6), *then there exist solutions $(u_1, \cdots, u_n)$ of $Ly = 0$ such that each of the systems*

$$(u_{n-n_j+1}^{(j-1)}, \cdots, u_n^{(j-1)}), \qquad j = 1, \cdots, l$$

is Descartes on $[a, b]$.

In the case $r = l = 1$, this is Pólya's criterion [16] for disconjugacy (i.e., (n)-invertibility). If $n_j = n - j + 1$, $j = 1, \cdots, n$, this is essentially the criterion given in [14] for right disfocality (i.e., right-$(1, \cdots, 1)$-invertibility) where the condition of (a) was given in terms of Strict Property I; this was also shown to be necessary.

It is of interest to consider operators L in which the coefficients a_i, $i = 1, \cdots, n$ are real constants in the context of Theorem 1. Let $p(\lambda) = e^{-\lambda t}L(e^{\lambda t})$ be the characteristic polynomial with $\lambda_1, \cdots, \lambda_n$ the roots of $p(\lambda) = 0$. A necessary and sufficient condition for the disconjugacy of L on every interval is that λ_i be all real. The necessity is obvious since, if there is a complex root, then there is an oscillatory solution. The sufficiency follows from considering the system $(u_1, \cdots, u_n)$, where $u_i(t) = e^{\lambda_i t}$, if $\lambda_1 < \cdots < \lambda_n$ and, if λ_i is a multiple root of multiplicity m, $u_{i+j}(t) = t^j e^{\lambda_i t}$, $j = 0, \cdots, m-1$, for which $W(u_1, \cdots, u_k) > 0$, $k = 1, \cdots, n$ on every interval. For $k = 1, \cdots, n-1$, L is right-$(n-k, 1, \cdots, 1)$-invertible on every interval if and only if all of the roots are real and there are at least k nonnegative roots. Thus, from Remark 3.1, if $p(\lambda) = 0$ has at least k positive roots, then L is right-$(m_1, \cdots, m_{k+1})$-invertible for any sequence $\{m_j\}$ satisfying

$$\sum_{j=1}^{k+1} m_j = n, \qquad \sum_{j=i}^{k+1} m_j \leqq k + 2 - i, \qquad i = 2, \cdots, k+1.$$

In particular, L is right disfocal on every interval if and only if there are at least $n-1$ nonnegative roots. Finally, all n roots positive is necessary and sufficient for L to be right-$(m_1, \cdots, m_l)$-invertible on every interval for every sequence of nonnegative integers $\{m_j\}$ satisfying $\sum_{j=1}^{l} m_j = n$. To illustrate, consider the operators

$$L = (D+1)(D+2), \quad M = (D+1)(D-1), \quad N = (D-1)(D-2).$$

All three are disconjugate on every interval, M and N are disfocal on every interval, while N has the property that $u = 0$ is the only solution u of $Ny = 0$ for which

$$u^{(i-1)}(t_i) = 0, \qquad u^{(j-1)}(t_j) = 0, \qquad t_i < t_j, \quad i \leqq j$$

or

$$u^{(i-1)}(t_i) = u^{(j-1)}(t_i) = 0, \qquad i \neq j.$$

To see the sufficiency assertions about constant coefficient operators observe that the functions $(u_1, \cdots, u_n)$ satisfy the appropriate Property I′ in Theorem 1(a) for every interval. The necessity assertions follow by considering the determinants $\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2)$ with $\tau_1 = \{t_1, \cdots, t_1\}$, $\tau_2 = \{t_2, \cdots, t_2\}$, $|\tau_1| = n-k$, $|\tau_2| = k$, $t_1 < t_2$. From Rolle's theorem and $W(u_1, \cdots, u_n) > 0$, if t_2 is sufficiently close to t_1, then

$$\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2) > 0, \tag{3.7}$$

and, if $t_2 \to \infty$, then

$$\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2) = W(u_1, \cdots, u_{n-k})(t_1) W(u'_{n-k+1}, \cdots, u'_n)(t_2)[1 + o(1)]. \tag{3.8}$$

Also $W(u_1, \cdots, u_{n-k})(t_1) > 0$ and $W(u'_{n-k+1}, \cdots, u'_n)(t_2) < 0$ if t_2 is sufficiently large and $\lambda_{n-k+1} < 0 \leqq \lambda_{n-k+2}$, so that $\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2) < 0$ if t_2 is large, from (3.8), and therefore $\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2) = 0$ for some $t_2 > t_1$, from (3.7). Therefore L is right-$(n-k+1, 1, \cdots, 1)$-invertible but not right-$(n-k, 1, \cdots . 1)$-invertible on sufficiently long intervals.

Proof of Theorem 1. To prove part (a), observe that if $(\tau_1; \cdots; \tau_l)$ is an increasing partition of n points in $[a, b]$ and $|\tau_j| = m_j$, $j = 1, \cdots, l$, then it follows from Proposition 1(a) that $\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) > 0$. Thus the only solution of $Ly = 0$ such that $y^{(j-1)}$ has m_j zeros at τ_j, $j = 1, \cdots, l$, is the zero solution.

To prove part (b), suppose that L is right-$(m_1, \cdots, m_l)$-invertible on $[a, b]$ if $\{m_j\}$ satisfies (3.6). Notice that this implies L is right-$(m_1, \cdots, m_l)$-invertible also on $[c, d]$, for some $c < a$, $d > b$, since $\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) = 0$ for some $(\tau_1; \cdots; \tau_l) \subset [c, d]$ for each $c < a$, $d > b$ implies, by continuity, that this also holds for some $(\tau_1; \cdots; \tau_l) \subset [a, b]$, contradicting the invertibility assumption on $[a, b]$. Let $(u_1, \cdots, u_n)$ be such that $Lu_i = 0$,

$$u_i^{(k-1)}(c) = 0, \quad k = 1, \cdots, i-1, \qquad u_i^{(i-1)}(c) > 0, \quad i = 1, \cdots, n, \tag{3.9}$$

$$u_i^{(k-1)}(d) = 0, \qquad k = j, \cdots, n-i+j-1, \quad i = n-n_j+1, \cdots, n-n_{j+1}, \quad j = 1, \cdots, l, \tag{3.10}$$

where $n_1 = n$ and $n_{l+1} = 0$. It is asserted that, for each $j = 1, \cdots, l$, all Wronskians of consecutive functions from the set $(u_{n-n_j+1}^{(j-1)}, \cdots, u_n^{(j-1)})$ are positive on (c, d) and therefore, from Lemma 3, each of these is a Descartes system on $[a, b]$, completing the proof.

To prove the assertion of positivity of the consecutive Wronskians made in the preceding paragraph, suppose that $n - n_j + 1 \leqq p \leqq q \leqq n$ and that $W(u_p^{(j-1)}, u_{p+1}^{(j-1)}, \cdots, u_q^{(j-1)})(t_0) = 0$ for some $t_0 \in (c, d)$. This implies that there is a nontrivial solution u of the form

$$u = c_p u_p + c_{p+1} u_{p+1} + \cdots + c_q u_q$$

such that $u^{(j-1)}$ has a zero of multiplicity $m_j = q - p + 1$ at t_0. From (3.9), u has a zero of multiplicity $m_1 = p - 1$ at c; from (3.10), $u^{(h-1)}$ has a zero multiplicity $m_h = n - q$ at d, if h is determined by $n - n_h + 1 \leqq q \leqq n - n_{h+1}$. Thus, if $m_k = 0$, $k \neq 1, j, h$, it follows that $\{m_k\}$ satisfies (3.6), contradicting the invertibility assumption on L, and therefore none of the Wronskians can vanish on (c, d). The positivity of the Wronskians on (c, d) follows from (3.9), which implies that, near $t = c$, for some $\mu_i > 0$

$$u_i^{(k-1)}(t) = [\mu_i + o(1)](t-c)^{i-k}, \qquad k = 1, \cdots, i, \quad i = 1, \cdots, n,$$

so that the Wronskians are positive near c and thus throughout (c, d).

THEOREM 2. *Suppose* $r \geqq 2$, $l \leqq n$ *in* (2.1) *and the equation* $Ly = 0$ *has a system* $(u_1, \cdots, u_n)$ *of solutions with Property* I [*Property* I'] *on* (a, b). *Let* f *be* n *times differentiable on* (a, b). *Then the conditions*

(i) $Lf \geqq 0$ *on* (a, b),

(ii) $f^{(j-1)}$ *has* m_j *zeros at* τ_j, $j = 1, \cdots, l$,

where $(\tau_1; \cdots; \tau_l)$ *is an increasing partition of* n *points* $(t_1, \cdots, t_n)$ *in* (a, b) *and* $\{m_j\}$ *satisfies* (3.6), *imply*

$$P(t)f(t) \geqq 0, \qquad t \in (r_1, s_1),$$

and, if $\sum_{j=p}^{l} m_j < n_p$ *for any* $p = 2, \cdots, l$, *then*

$$P(t)f^{(p-1)}(t) \geqq 0, \qquad t \in (r_p, s_p).$$

Moreover, the inequalities are strict at any point $t \neq t_i$ *such that* $Lf \not\equiv 0$ *on some interval of the form* $[t, t_i]$ *or* $[t_i, t]$. *Here* $P(t) = \prod_{i=1}^{n} (t - t_i)$, $r_p = \max\{t \in \tau_i,\ i < p\}$, $p = 2, \cdots, l$, $s_p = \min\{t \in \tau_i,\ i > p\}$, $p = 1, \cdots, l-1$, $r_1 = a$, $s_l = b$.

Note that in the case $r = 1$, L satisfies the conditions of Theorem 2 if and only if it is right-$(m_1, \cdots, m_l)$-invertible, by Theorem 1(b).

The proof of Theorem 2 closely parallels that of Proposition 1. The proof of Proposition 1 used Lemma 1 with the stronger assumption $W(u_1, \cdots, u_n) > 0$. Under the conditions of Theorem 2, the same proof may be carried out for the system $(u_1, \cdots, u_{n+1})$, $u_{n+1} = f$, using the full generality of Lemma 1. The conditions of Theorem 2 thus imply $\Omega(u_1, \cdots, u_n, f)(\tau_1; \cdots; \tau_p^t; \cdots; \tau_l) \geqq 0$, where τ_p^t is obtained by inserting t in the ordered set τ_p, so that $(\tau_1; \cdots; \tau_p^t; \cdots; \tau_l)$ is increasing. It is not difficult to verify that

$$\Omega(u_1, \cdots, u_n, f)(\tau_1; \cdots; \tau_p^t; \cdots; \tau_l) = f^{(p-1)}(t)\Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_l) \operatorname{sgn} P(t),$$

and the theorem follows.

Theorem 2 is a generalization of Čaplygin's inequality which was first proved for $l = p = 1$ when $t_1 = t_2 = \cdots t_n$. The general formulation for disconjugate operators is an immediate consequence of the work of Pólya [16, Thm. V], i.e., Lemma 1 of this paper.

It was also seen in [11], [13] that if $Lf \geqq 0$ and f has fewer than n zeros in an interval (a, b), then inequalities can still be derived for certain differential expressions Mf. A similar generalization can be given for Theorem 2 if the partition $(\tau_1, \cdots; \tau_l)$ contains less than n points.

COROLLARY 1 (Mean value theorem). *Suppose* L *satisfies the conditions of Theorem 2 and* f *is an* n *times differentiable function on* (a, b) *such that* $f^{(j-1)}$ *has* m_j *zeros at* τ_j, $j = 1, \cdots, l$, *where* $(\tau_1; \cdots; \tau_l)$ *is an increasing partition of* $n+1$ *points* $t_1, \cdots, t_{n+1}$ $(t_1 \neq t_{n+1})$ *in* (a, b). *If*

$$\sum_{j=1}^{l} m_j = n + 1, \qquad \sum_{j=i}^{l} m_j \leqq n_i, \qquad i = 2, \cdots, l,$$

then $Lf(\xi) = 0$ *for some* $\xi \in (t_1, t_{n+1})$. *If* f *is not identically zero in* (t_1, t_{n+1}) *then* Lf *changes sign in* (t_1, t_{n+1}).

Proof. This follows from alternately omitting t_1 and t_{n+1} from the partition to obtain, by Theorem 2,

$$f(t) \prod_{i>1} (t - t_i) \geqq 0, \quad f(t) \prod_{i<n+1} (t - t_i) \geqq 0,$$

if $Lf \geqq 0$. If f is identically zero then $Lf \equiv 0$; *if* $f \not\equiv 0$ these inequalities contradict each other and a similar contradiction is obtained if $Lf \leqq 0$.

COROLLARY 2 (Cauchy mean value theorem). *Suppose* L *satisfies the conditions of Theorem 2, and* f, g *are any* n *times differentiable functions on* (a, b). *If* ϕ, ψ *are chosen so that*

(i) $L\phi = 0$, $L\psi = 0$,

(ii) $(f - \phi)^{(j-1)}$, $(g - \psi)^{(j-1)}$ *have* m_j *zeros at* τ_j, *with* τ_j *as in Theorem 2,*

then for $t \in (r_1, s_1)$ *there exists* $\xi \in (a, b)$ *such that*

$$[f(t) - \phi(t)]Lg(\xi) = [g(t) - \psi(t)]Lf(\xi).$$

If $\sum_{j=p}^{l} m_j < n_p$ *for any* $p = 2, \cdots, l$, *and* $t \in (r_p, s_p)$, *there exists* $\xi \in (a, b)$ *such that*

$$[f(t)-\phi(t)]^{(p-1)}Lg(\xi) = [g(t)-\psi(t)]^{(p-1)}Lf(\xi).$$

Proof. If $t \in \tau_p$, the result is true for any ξ. If $t \notin \tau_p$, $t \in (r_p, s_p)$ and $[g(t)-\psi(t)]^{(p-1)} = 0$, then by Corollary 1 there exists ξ such that $Lg(\xi) = 0$ also. Finally, if $t \in (r_p, s_p)$ and $[g(t)-\psi(t)]^{(p-1)} \neq 0$, then a constant c may be chosen so that

$$[f(t)-\phi(t)]^{(p-1)} - c[g(t)-\psi(t)]^{(p-1)} = 0. \tag{3.11}$$

Applying Corollary 1 to the function $(f-\phi)-c(g-\psi)$ shows that $Lf(\xi) = cLg(\xi)$ for some ξ, which yields the result on multiplying (3.11) by $Lg(\xi)$.

If $\{m_j\}$ is any sequence satisfying (3.6) when $\{n_j\}$ satisfies (2.1) with $r = 1$, then L is right-$(m_1, \cdots, m_l)$-invertible on a neighborhood of any point a in the domain of its coefficients. This follows from the continuity of Ω and Rolle's theorem. Let $\beta(a)$ be the supremum of all b such that L is right-$(m_1, \cdots, m_l)$-invertible on $[a, b]$ for all such sequences $\{m_j\}$. Then there is a sequence $\{m_j\}$ for which L is not right-$(m_1, \cdots, m_l)$-invertible on $[a, \beta(a)]$, again by the continuity of Ω.

Let the solutions $u_i(c, t)$, $i = 1, \cdots, n$, of $Ly = 0$ be defined by

$$u_i^{(j-1)}(c) = \delta_{ij} \qquad (\delta_{ij} = 0, i \neq j, \delta_{ii} = 1), \tag{3.12}$$

and let $W_k^i(c, t) = W(u_{n-k+1}^{(j-1)}, \cdots, u_n^{(j-1)})(c, t)$.

LEMMA 4. (a) *If* c *is in the domain of* β, *then* $(u_1(c, t), \cdots, u_n(c, t))$ *has Strict Property* I′ *on* $(c, \beta(c))$; *that is,*

$$W_k^i(c, t) > 0, \qquad t \in (c, \beta(c)), \quad k = 1, \cdots, n_j, \quad j = 1, \cdots, l.$$

(b) *For any* a *there exists a* $\delta > 0$ *such that, if* $c \in (a-\delta, a+\delta)$, *then* $\beta(c) > a+\delta$.

None of the Wronskians $W_k^i(c, t)$ can vanish at $t_0 \in (c, \beta(c))$ since this, together with (3.12), contradicts the invertibility of L on $[c, t_0]$ as in the proof of Theorem 1(b). The positivity of the Wronskians also follows as in Theorem 1(b) from the asymptotic behavior of $u_i^{(j-1)}(c, t)$, $t \to c+$. Part (b) of the lemma follows from (a) since L is invertible on a neighborhood of a so that $\beta(c) > a+\delta$ if $c \in (a-\delta, a+\delta)$ and δ is small.

THEOREM 3. *If* $b = \beta(a)$, *then* $W_k^h(a, b) = 0$ *and there exists a solution* u *of* $Ly = 0$ *such that*

$$u^{(j-1)}(a) = 0, \qquad j = 1, \cdots, n-k,$$

$$u^{(j-1)}(b) = 0, \qquad j = h, \cdots, h+k-1$$

for some k, $n_{h+1} < k \leqq n_h$, *and some* h, $1 \leqq h \leqq l$. *Also* $u^{(j-1)} > 0$ *on* (a, b), $j = 1, \cdots, h$.

Proof. First, not all $W_k^i(a, b)$ can be positive $k = 1, \cdots, n_j$, $j = 1, \cdots, l$ because in that case $W_k^i(a, t) > 0$ for all $t \in (a, b]$, by Lemma 4. Since $W_k^i(c, t)$ is continuous in (c, t), it follows that $\lim_{c \to a} W_k^i(c, t) = W_k^i(a, t)$, uniformly with respect to $t \in [a, b]$. Thus, if $\delta > 0$ satisfies Lemma 4(b) and $c < a$ is sufficiently close to a, $W_k^i(c, t) > 0$, $t \in [a+\delta, b]$ and $W_k^i(c, t) > 0$, $t \in (c, a+\delta)$. Therefore $(u_1(c, t), \cdots, u_n(c, t))$ has Strict Property I′ on $[a, b]$ contradicting the noninvertibility of L on $[a, b]$, by Theorem 1(a).

Thus, $W_k^h(a, b) = 0$ for some $k \leqq n_h$; choose k to be the smallest number for which it occurs for a particular h. Then

$$0 = W_k^h(a, b) = W(u_{n-k+1}^{(h-1)}, \cdots, u_n^{(h-1)})(a, b)$$

implies there is a nontrivial solution

$$u = c_{n-k+1}u_{n-k+1} + \cdots + c_n u_n$$

such that $u^{(h-1)}$ has k zeros at b and, from (3.12) u has $n-k$ zeros at a. Moreover, u is

unique to within a constant multiple since $W_{k-1}^h(a,b) \neq 0$ implies that the system of k equations determining the k constants c_i has rank $k-1$. This also implies $c_{n-k+1} \neq 0$. The function $u^{(h-1)}$ cannot vanish in (a,b) since $W_i^h(a,t) > 0$, $t \in (a,b]$, $i = 1, \cdots, k-1$, and $u^{(h-1)}$ has k zeros at b; thus, if $u^{(h-1)}(c) = 0$, $c \in (a,b)$, then, by Pólya's mean value theorem on the interval $[c,b]$,

$$\begin{aligned} 0 &= W(u^{(h-1)}, u_{n-k+2}^{(h-1)}, \cdots, u_n^{(h-1)})(a,\xi) \\ &= c_{n-k+1} W(u_{n-k+1}^{(h-1)}, \cdots, u_n^{(h-1)})(a,\xi) = c_{n-k+1} W_k^h(a,\xi) \end{aligned}$$

for some $\xi \in (a,b)$, contradicting Lemma 4(a). Thus the constants c_i may be chosen so that $u^{(h-1)} > 0$ on (a,b). Now u has $n-k$ zeros at a, and $k \leqq n_h \leqq n-h+1$ implies $n-k \geqq h-1$, which together with $u^{(h-1)} > 0$ on (a,b) give $u^{(j-1)} > 0$ on (a,b), $j = 1, \cdots, h$.

The only assertion of Theorem 3 which remains to be proved is that $n_{h+1} < k$ for some extremal solution u of the type discussed in the preceding paragraph. If this is not the case then $n_{h+1} \geqq k$ and choose h maximal such that $W_k^h(a,b) = 0$. Then the solution u has $n-k$ zeros at a, $u^{(h-1)}$ has k zeros at b and $n-k \geqq h$ since $k \leqq n_{h+1} \leqq n-h$. Therefore $u^{(h-1)}$ has at least 1 zero at a and k zeros at b and, from Rolle's theorem, $u^{(h)}$ has a zero at $c \in (a,b)$ as well as $(k-1)$ zeros at b. Now, $W_i^{h+1}(a,t) > 0$, $t \in (a,b]$, $i = 1, \cdots, k-1$ and Pólya's mean value theorem implies

$$\begin{aligned} 0 &= W(u^{(h)}, u_{n-k+2}^{(h)}, \cdots, u_n^{(h)})(a,\xi) \\ &= c_{n-k+1} W(u_{n-k+1}^{(h)}, \cdots, u_n^{(h)})(a,\xi) = c_{n-k+1} W_k^{h+1}(a,\xi) \end{aligned}$$

for some $\xi \in (a,b)$, contradicting Lemma 4(a).

COROLLARY 3. *The function β is increasing on its domain.*

Proof. It is clear that β is nondecreasing. Suppose $b = \beta(a)$ and $c \in (a,b)$; then $\beta(c) \geqq b$ and it suffices to show that $\beta(c) > b$. If $\beta(c) = b$, then $W_k^h(c,b) = 0$ for some k and h as in Theorem 3. If $h > 1$, then $\Omega(u_1, \cdots, u_n)(\tau_1; \tau_2; \cdots; \tau_h) = 0$, where $\tau_1 = (c, \cdots, c)$, $\tau_h = (b, \cdots, b)$, $\tau_i = \phi$, $i \neq 1, h$, $|\tau_1| = n-k$, $|\tau_h| = k$ and let $n-k$ be maximal such that this holds. Since (as shown in the proof of Theorem 3) the extremal solution u is unique to within a multiple, it follows that this solution is given by

$$u(t) = \Omega(u_1, \cdots, u_n)(\tau_1', t; \tau_2; \cdots; \tau_h)$$

where $\tau_1' = (c, \cdots, c)$, $|\tau_1'| = n-k-1$. In particular, $u(c) = \cdots = u^{(n-k-1)}(c) = 0$, $u^{(n-k)}(c) \neq 0$. If

$$\begin{aligned} &Y(t,s) = \Omega(u_1, \cdots, u_n)(\sigma_1; \tau_2; \cdots; \tau_{h-1}; \sigma_h), \\ &\sigma_1 = (t, \cdots, t), \quad \sigma_h = (s, \cdots, s), \quad |\sigma_1| = n-k, \quad |\sigma_h| = k, \end{aligned} \tag{3.13}$$

then

$$Y(c,b) = \Omega(u_1, \cdots, u_n)(\tau_1; \cdots; \tau_h) = 0,$$

$$\frac{\partial}{\partial t} Y(c,b) = u^{(n-k)}(c) \neq 0.$$

By the implicit function theorem, there exists a continuous function T defined in a neighborhood of b such that $T(b) = c$ and $Y(T(s), s) = 0$. Therefore, from (3.13), there is a nontrivial solution u which has $n-k$ zeros at $T(s)$ and $u^{(h-1)}$ has k zeros at s. If $s < b$ is sufficiently close to b then $T(s) > a$, contradicting $b = \beta(a)$.

In the case $h=1$, $\Omega(u_1,\cdots,u_n)(\tau)=0$ where $\tau=(c,\cdots,c,b,\cdots,b)$, c has multiplicity $n-k$ in τ and b has multiplicity k. This may also be shown to lead to a contradiction in the same way as was done in the case $h>1$, completing the proof.

It is well known (cf. Coppel [2, Thm. 7, p. 102]) that, when $l=1$ β is continuous on its domain. An example given in [14] shows that this is not necessarily the case when $l>1$.

4. Comparison criteria. For the remainder of the paper it will be assumed that $r=1$ in the expression (2.1). It will also be assumed that the coefficients a_k in L are of class C^{n-k} on their domain. The adjoint L^* of L is defined by

$$L^*y=(-1)^n(a_0y)^{(n)}+(-1)^{(n-1)}(a_1y)^{(n-1)}+\cdots+(a_ny)$$

where $a_0=1$. From Theorem 3, it is of interest to consider boundary conditions of the form, for each $k=1,\cdots,n-1$,

$$(4.1)\qquad [a,b]\qquad \begin{aligned} &y^{(j-1)}(a)=0, && j=1,\cdots,n-k,\\ &y^{(j-1)}(b)=0, && j=k,\cdots,h+k-1, \end{aligned}$$

where h is determined by $n_{h+1}<k\leqq n_h$. Here $m_1=n-k$, $m_h=k$, $m_j=0$ if $j=1,h$ so that (3.6) holds. We will also consider adjoint boundary conditions of the form

$$(4.2)\qquad [a,b]\qquad \begin{aligned} &l_jy(a)=0, && j=1,\cdots,k,\\ &l_jy(b)=0, && j=1,\cdots,n-h-k+1,\quad n-h+2,\cdots,n, \end{aligned}$$

where

$$l_jy=(-1)^{j-1}(a_0y)^{(j-1)}+(-1)^{(j-2)}(a_1y)^{(j-2)}+\cdots+(a_{j-1}y).$$

Lagrange's identity (cf. [5, p. 67]) states that if u, v are functions in the domains of L, L^* respectively, then

$$(4.3)\qquad vLu-uL^*v=[uv]',$$

where $[uv]=ul_nv+u'l_{n-1}v+\cdots+u^{(n-1)}l_1u^{(n-1)}l_iv$. Also, if $y=u$ satisfies and $y=v$ satisfies (4.2) $[a,b]$, then

$$(4.4)\qquad [uv](b)-[uv](a)=0.$$

In [12, Thm. 3.1], it was shown that a boundary value problem $Ly=0$, $Uy=0$ on an interval $[a,b]$ has no nontrivial nonnegative solution if there exists a function ψ on $[a,b]$ such that $L^*\psi\geqq 0$, $U^*\psi=0$, where U^* is a boundary operator adjoint to U, and $L^*\psi>0$ on a set of positive measure. If we choose U and U^* to be the boundary operators in (4.1) $[a,b]$ and (4.2) $[a,b]$ respectively, it follows from Theorem 3 and its corollary that invertibility conditions for an operator L may be given in terms of the existence of solutions to $L^*\psi\geqq 0$, $U^*\psi=0$. However U^* depends on the operator L in this case, so it is somewhat difficult to construct functions ψ with general applicability. Lemma 5 helps us to avoid this difficulty.

LEMMA 5. *If $b=\beta(a)$, then there is a nontrivial solution v of $L^*y=0$ satisfying the boundary conditions* (4.2) $[a,b]$ *for some $k=1,\cdots,n-1$ and $v>0$ on (a,b). The boundary value problems $L^*y=0$,* (4.2) $[c,d]$, *have no nontrivial solution if $[c,d]$ is a proper subinterval of $[a,b]$.*

Proof. From Corollary 3, the problems $Ly = 0$, (4.1) $[c, d]$, have no nontrivial solution if $[c, d]$ is a proper subinterval of $[a, b]$. Thus, the Green's functions $G(t, s)$ exist for these problems. Now $G(s, t)$ are the Green's functions associated with the adjoint problems $L^*y = 0$, (4.2) $[c, d]$, so these problems have no nontrivial solutions. The same argument shows that $L^*y = 0$, (4.2) $[a, b]$, have a nontrivial solution for some $k = 1, \cdots, n-1$. It remains to show that some such solution v is positive on (a, b). Let the nontrivial solutions $v_1, \cdots, v_{n-1}$ of $L^*y = 0$ be such that each v_k satisfies the $n-k$ boundary conditions

$$l_j y(b) = 0, \qquad j = 1, \cdots, n-h-k+1, n-h+2, \cdots, n. \tag{4.5}$$

Then $W(v_1, \cdots, v_k) \neq 0$, $k = 1, \cdots, n-1$, on (a, b) and the solutions may be chosen so that $W(v_1, \cdots, v_k) > 0$ on (a, b). Otherwise, if $W(v_1, \cdots, v_k)(c) = 0$ for some $c \in (a, b)$, one finds a nontrivial solution to $L^*y = 0$, (4.2) $[c, b]$ (since $v^{(j-1)}(a) = 0$, $j = 1, \cdots, k$, is equivalent to $l_j v(a) = 0, j = 1, \cdots, k$) and it has already been shown that no such solution exists. For each $k = 1, \cdots, n-1$, the solutions $(v_1, \cdots, v_k)$ form a basis for the set of solutions satisfying (4.5); also $W(v_1, \cdots, v_k)(a) = 0$ for each k for which $L^*y = 0$, (4.2) $[a, b]$, has a nontrivial solution. Let k be the smallest number such that $W(v_1, \cdots, v_k)(a) = 0$. Since $W(v_1, \cdots, v_{k-1})(a) \neq 0$, if $k > 1$, there is a nontrivial solution $v = c_1 v_1 + \cdots + c_k v_k$ $(c_k \neq 0)$, unique to within a constant multiple, of the problem $L^*y = 0$, (4.2) $[a, b]$, for this k. Moreover v does not vanish in (a, b) because $v(c) = 0$, $c \in (a, b)$, and $v(a) = v'(a) = \cdots = v^{(k-1)}(a) = 0$ (equivalent to $l_1 v(a) = l_2 v(a) = \cdots = l_k v(a) = 0$), together with $W(v_1, \cdots, v_j) > 0$, $j = 1, \cdots, k-1$, on $[a, b)$, implies

$$0 = W(v_1, \cdots, v_{k-1}, v)(\xi) = c_k W(v_1, \cdots, v_k)(\xi)$$

for some $\xi \in (a, c)$, by Pólya's mean value theorem [16]. Since $c_k \neq 0$, this contradicts $W(v_1, \cdots, v_k) > 0$ on (a, b). In the case $k = 1$, $v = c_1 v_1$ $(c_1 \neq 0)$ and this function does not vanish in (a, b) either. Thus the constants c_i may be chosen so that $v > 0$ on (a, b).

LEMMA 6. *Suppose there exists a function ψ which has an absolutely continuous derivative of order $n-1$ on $[a, b]$, satisfies the boundary conditions* (4.1) $[a, b]$ *and $L\psi \geqq 0$ on $[a, b]$ with strict inequality on a set of positive measure. Then the boundary value problem $L^*y = 0$,* (4.2) $[a, b]$, *has no solution which is positive on (a, b).*

Proof. This is a special case of [12, Thm. 3.1.]. It follows from the fact that if $y = v$ is a positive solution of $L^*y = 0$, (4.2) $[a, b]$, then from (4.3) and (4.4)

$$\begin{aligned} 0 &= [\psi v](b) - [\psi v](a) \\ &= \int_a^b (vL\psi - \psi L^*v) = \int_a^b vL\psi > 0, \end{aligned}$$

a contradiction.

The following theorem, which may be considered a Sturm comparison criterion for invertibility, may be deduced from Lemmas 5 and 6 and Corollary 3.

THEOREM 4. *A sufficient condition that L be right-$(m_1, \cdots, m_l)$-invertible on $[a, b]$ if $\{m_j\}$ satisfies* (3.6) $(r = 1)$, *is that on each interval $[a, c]$, $c \in (a, b]$ and for each $k = 1, \cdots, n-1$, there exists a function $\psi_k \in AC^{n-1}[a, c]$ such that*

(i) $y = \psi_k$ *satisfies* (4.1) $[a, c]$,

(ii) $L\psi_k \geqq 0$, *with strict inequality on a set of positive measure in* $[a, c]$.

A necessary condition is that such functions exist and satisfy

$$(-1)^k \psi_k^{(j-1)} > 0 \quad \text{on } (a, c), \qquad j = 1, \cdots, h.$$

The necessity of the existence of such functions follows from the existence of the appropriate Green's functions, so that $Ly = 1$, (4.1) $[a, c]$, has a solution ψ_k for each $k = 1, \cdots, n-1$. The sign of the derivatives of ψ_k follows from Theorem 2.

The following corollary reduces, in the case of disconjugacy, to a comparison principle of Levin [7] and Nehari [15].

COROLLARY 4. *Suppose that* $L_i y = Ly + q_i y$, $i = 1, 2$, *where* $q_1 \leqq 0 \leqq q_2$, *on an interval* J. *If* L_1 *and* L_2 *are both right-*$(m_1, \cdots, m_l)$*-invertible on* J *for all sequences* $\{m_j\}$ *satisfying* (3.6) $(r = 1)$ *then so also is* L.

Proof. If k is even, let ψ_k be the solution of $L_1 y = 1$, (4.1) $[a, b]$; then $L\psi_k = 1 - q_1\psi_k > 0$, since $(-1)^k\psi_k > 0$ and $q_1 \leqq 0$. Similarly, if k is odd, let ψ_k be the solution of $L_2 y = 1$, (4.1) $[a, b]$, so that $L\psi_k = 1 - q_2\psi_k > 0$. This applies to any subinterval [a, b] of J.

Note that, for a second order operator L only the case $k = 1$ occurs, and in this case the result reduces to: L is disconjugate (right-disfocal) if L_2 is disconjugate (right-disfocal).

When $l = 1$, Corollary 5 is a disconjugacy criterion of Hartman [4] and Levin [9]. A proof for this case may also be found in Coppel [2].

The following notation will be used. The symbol $(u_i, \cdots, \hat{u}_j, \cdots, u_k)$ denotes the $(k-i)$-tuple obtained by omitting u_j from $(u_i, \cdots, u_k)$ if $i \leqq j \leqq k$ and it denotes the $(k-i+1)$-tuple $(u_i, \cdots, u_k)$ if $j < i$ or $j > k$. Given $k \in \{1, \cdots, n-1\}$, let h be determined by $n_{h+1} < k \leqq n_h$ and let

$$\Omega(u_1, \cdots, u_{n+1})(k: a, b, t) = \Omega(u_1, \cdots, u_{n+1})(\tau_1; \cdots; \tau_h),$$

where

$$\tau_1 = (a, \cdots, a, t), \quad \tau_h = (b, \cdots, b), \quad |\tau_1| = n - k + 1, |\tau_h| = k$$

and $|\tau_j| = 0$, $j \neq 1, h$. Also let

$$\Omega(u_1, \cdots, u_n)(k: a, b) = \Omega(u_1, \cdots, u_n)(\sigma_1; \cdots; \sigma_h),$$

where

$$\sigma_1 = (a, \cdots, a), \quad \sigma_h = (b, \cdots, b), \quad |\sigma_1| = n - k, \quad |\sigma_h| = k$$

and $|\sigma_j| = 0$, $j \neq 1, h$.

Remark. Proposition 1 shows that $\Omega(u_1, \cdots, u_n)(k: a, b) > 0$ for each $k = 1, \cdots, n-1$, if $(u_1, \cdots, u_n)$ has Strict Property I′ on $[a, b]$.

COROLLARY 5. *A necessary and sufficient condition that* L *be right-*$(m_1, \cdots, m_l)$*-invertible on* $[a, b]$ *for all* $\{m_j\}$ *satisfying* (3.6) $(r = 1)$ *is that there exist functions* $u_2, \cdots, u_n \in C^n[a, b]$ *such that*

(i) $(-1)^{n-j+1} Lu_j \geqq 0$, $j = 2, \cdots, n$, *on* $[a, b)$,

(ii) $W(u_{n-k+1}^{(j-1)}, \cdots, \hat{u}_i^{(j-1)}, \cdots, u_n^{(j-1)}) > 0$ *for all* $i = 1, \cdots, n$, *if* $k = 1, \cdots, n-1$ *when* $j = 1$ *and* $k = 1, \cdots, n_j$ *when* $j = 2, \cdots, l$.

Proof. The condition is necessary. Indeed, as was shown in Theorem 1 for the invertibility specified, it is necessary that there exist solutions $(u_1, \cdots, u_n)$ of $Ly = 0$ such that each of the systems $(u_{n-n_j+1}^{(j-1)}, \cdots, u_n^{(j-1)})$ is a Descartes system on $[a, b]$, which is more restrictive than the necessary condition claimed here.

To prove sufficiency, let $u_1 = e^{-\lambda t}$, $u_{n+1} = e^{\lambda t}$. The constant λ may be chosen large enough to ensure that $(-1)^{n-j+1} Lu_j > 0$ on $[a, b]$ for $j = 1, n+1$, and that each of the systems $(u_1, \cdots, \hat{u}_i, \cdots, u_{n+1})$, $i = 1, \cdots, n+1$ has Strict Property I′ on any pre-assigned closed subinterval of $[a, b)$. Thus, from the remark preceding the statement of

Corollary 5, it follows that

$$\Omega(u_1, \cdots, \hat{u}_i, \cdots, u_{n+1})(k: a, c) > 0, \qquad i = 1, \cdots, n+1, \tag{4.4}$$

if $c \in (a, b)$, $k = 1, \cdots, n-1$, and λ is sufficiently large. Now consider the functions ψ_k defined on (a, c) by

$$\psi_k(t) = \Omega(u_1, \cdots, u_{n+1})(k: a, c, t), \qquad k = 1, \cdots, n-1$$

and on $[a, c]$ by continuity. Since ψ_k satisfies (4.1) $[a, c]$ and

$$L\psi_k = \sum_{i=1}^{n+1} \Omega(u_1, \cdots, \hat{u}_i, \cdots, u_{n+1})(k: a, c)(-1)^{n-i+1} Lu_i > 0$$

if λ is sufficiently large, by (4.4) and condition (i) it follows from Theorem 4 that $\beta(a) \geqq b$. The functions $u_2, \cdots, u_n$ may be extended to the left of a with the conditions of the corollary still holding, thus showing $\beta(d) \geqq b$ for some $d < a$. Since β is increasing (Corollary 3), it follows that $\beta(a) > b$ and L is invertible as asserted on $[a, b]$.

As an example, a second order operator L is disconjugate on an interval $[a, b]$ if and only if there exists a function u_2 such that

$$u_2 > 0, \qquad Lu_2 \leqq 0 \quad \text{on } [a, b),$$

and L is right disfocal (i.e., right-$(1, 1)$-invertible) on $[a, b]$ if and only if u_2 exists such that

$$u_2 > 0, \quad u_2' > 0, \quad Lu_2 \leqq 0 \quad \text{on } [a, b).$$

The invertibility of boundary value problems associated with constant coefficient operators was discussed in detail in § 3 following the statement on Theorem 1. This discussion can now be extended to nonconstant coefficient operators. Hartman [4] and Levin [9] (cf. Coppel [2]) show that a nonconstant coefficient operator L is disconjugate on every interval if there exist constants $\mu_2, \cdots, \mu_n$ such that

$$\lambda_1(t) \leqq \mu_2 \leqq \lambda_2(t) \leqq \mu_3 \leqq \cdots \leqq \mu_n \leqq \lambda_n(t),$$

where $\lambda_i(t)$ are the roots of the characteristic equation $p(t, \lambda) = 0$, and $p(t, \lambda) = e^{-\lambda t} L(e^{\lambda t})$. This is accomplished by choosing $u_i(t) = e^{\mu_i t}$, $i = 2, \cdots, n$ if the constants μ_i are all distinct and by multiplying these functions by appropriate polynomials in the case of repeated values of μ_i. A similar observation holds in the more general context considered here. Corollary 5 shows that L is right-$(n-k, 1, \cdots, 1)$-invertible if at least k of the numbers μ_i are nonnegative and, in particular, L is right disfocal if all $n-1$ of the numbers $\mu_2, \cdots, \mu_n$ are nonnegative.

Results for disconjugacy related to Corollary 5 are proved in [11], [12], where, essentially, the condition that $(-1)^{n-j+1} Lu_j \geqq 0$ is replaced by a condition that $|Lu_j|$ be small. Similar results may be proved for invertibility. In particular, the result of [11] may be proved almost verbatim for invertibility, provided Pólya's mean value theorem [16] in the proof is replaced by the mean value theorem of the present paper, Corollary 1.

Perhaps the simplest sets of functions satisfying the conditions (4.1) $[a, b]$ for use in Theorem 4 are polynomials of degree n constructed as follows. Consider the functions

$$\psi(n, k, 1; t) = \frac{1}{n!}(t-a)^{n-k}(b-t^k), \qquad k = 1, \cdots, n$$

and

$$\psi(n, k, m; t) = \int_a^t \psi(n-1, k, m-1; s)\, ds, \qquad k = 1, \cdots, n-m+1, \quad m = 2, \cdots, n.$$

From

$$\psi(n, k, 1; t) = \frac{1}{n!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (b-a)^j (t-a)^{n-j}$$

it follows that

$$\psi(n, k, m; t) = \frac{1}{(n-m+1)!} \sum_{j=0}^{k} \frac{(-1)^{k-j}}{(n-j)(n-j-1)\cdots(n-j-m+2)} \binom{k}{j} (b-a)^j (t-a)^{n-j}.$$

These functions satisfy $\psi^{(n)}(n, k, m; t) = (-1)^k$ and, for $k = 1, \cdots, n-1$ (cf. [12]),

$$\begin{aligned} |\psi(n, k, 1; t)| &\leqq \frac{(n-1)^{n-1}}{n! n^n} (b-a)^n, \\ |\psi^{(j)}(n, k, 1; t)| &\leqq \frac{j}{(n-j)! n} (b-a)^{n-j}, \qquad j = 1, \cdots, n-1, \end{aligned} \tag{4.6}$$

while, for $k = 1, \cdots, n$,

$$|\psi^{(j)}(n, k, 1; t)| \leqq \frac{(b-a)^{n-j}}{(n-j)!}, \qquad j = 0, \cdots, n-1. \tag{4.7}$$

Clearly, if $m > 1$, $\psi(n, k, m; t)$ is positive on $(a, b]$ and has its maximum at $t = b$ only. This maximum may be shown by induction on k to be

$$\psi(n, k, m; b) = \frac{(b-a)^n}{n!} \binom{n-1}{m-2} \Big/ \binom{n-1}{k+m-2}.$$

Therefore

$$0 \leqq \psi(n, k, m; t) \leqq \frac{(b-a)^n}{n[(n-1)/2]![n/2]!}, \qquad \begin{matrix} k = 1, \cdots, n-m+1, \\ m = 2, \cdots, n, \end{matrix} \tag{4.8}$$

where $[x]$ denotes the greatest integer not exceeding x. Since

$$\begin{aligned} \psi^{(j)}(n, k, m; t) &= \psi(n-j, k, m-j; t), \qquad j = 0, \cdots, m-1, \\ \psi^{(j)}(n, k, m; t) &= \psi^{(j-m+1)}(n-m+1, k, 1; t), \qquad j = m, \cdots, n-1, \end{aligned}$$

(4.7) and (4.8) imply that

$$|\psi^{(j)}(m, k, m; t)| \leqq \frac{(b-a)^{n-j}}{(n-j)[(n-j-1)/2]![(n-j)/2]!}, \qquad j = 1, \cdots, m-2,$$

$$|\psi^{(j)}(n, k, m; t)| \leqq \frac{(b-a)^{n-j}}{(n-j)!}, \qquad j = m-1, \cdots, n-1.$$

All of the preceding inequalities are strict with equality holding at most at one point.

COROLLARY 6. (a) *A sufficient condition for L to be disconjugate on $[a, b]$ is*

$$\sum_{j=1}^{n-1} \frac{(n-j)}{j! n} |a_j(t)| (b-a)^j + \frac{(n-1)^{n-1}}{n! n^n} |a_n(t)| (b-a)^n \leqq 1$$

for each $t \in [a, b]$.

(b) *A sufficient condition that L be right(left)-*$(1, \cdots, 1, n-m+1)$*-invertible on* $[a, b]$ *is*

$$\sum_{j=1}^{n-m+1} \frac{1}{j!}|a_j(t)|(b-a)^j + \sum_{j=n-m+2}^{n} \frac{1}{j[(j-1)/2]![j/2]!}|a_j(t)|(b-a)^j \leqq 1$$

for each $t \in [a, b]$.

Part (a) of this corollary is due to Bessmertnyh and Levin [1]; the present proof is also given in [12]. In the case $m = n$, Part (b) states that L is right or left disfocal on $[a, b]$ if

$$\sum_{j=1}^{n} \frac{1}{j[(j-1)/2]![j/2]!}|a_j(t)|(b-a)^j \leqq 1$$

for each $t \in [a, b]$. This case is also known and is essentially part of the proof of a result of Levin [8] which states that L is disconjugate on $[a, b]$ if this inequality holds with $(b-a)$ replaced by $(b-a)/2$ (cf. Coppel [2, p. 86]. Observe that this disconjugacy criterion follows immediately from (b) since, by Rolle's theorem, L is disconjugate on $[a, b]$ if it is right disfocal on $[a, (a+b)/2]$ and left disfocal on $[(a+b)/2, b]$.

Proof of Corollary 6. Part (a) follows from Theorem 4 and the fact that the functions

$$\psi_k(t) = \psi(n, k, 1; t), \qquad k = 1, \cdots, n-1,$$

satisfy (4.1) $[a, b]$ with $l = 1$. The condition given implies, from (4.6), that $(-1)^k L\psi_k \geqq 0$ with strict inequality holding almost everywhere. Furthermore, since the condition of (a) is monotone in $(b-a)$, a similar set of functions may be constructed on every subinterval $[a, c]$ of $[a, b]$, as required by Theorem 4.

To prove Part (b), let $l = m$, $n_j = n-j+1$, $j = 1, \cdots, m$, and consider

$$\psi_k(t) = \psi(n, k, m; t), \qquad k = 1, \cdots, n-m+1,$$

$$\psi_k(t) = \psi(n, k, n-k+1; t), \qquad k = n-m+2, \cdots, n-1.$$

The functions satisfy the conditions (4.1) $[a, b]$ for the sequence $\{n_j\}$ and $(-1)^k L\psi_k \geqq 0$ from (4.9), and the result follows as in Part (a).

Several improvements are possible in Corollary 6, especially for lower order operators, since the inequalities (4.9) are quite rough. Also, if one is prepared to consider conditions which imply $(-1)^k L\psi_k \geqq 0$ for each individual k, then one can take advantage of the fact that $\psi^{(j)}(n, k, m; t) > 0$, $j = 0, \cdots, m-1$, so that $|a_{n-j}(t)|$ may be replaced by $a_{n-j}(t)_+$ or $a_{n-j}(t)_-$ as appropriate in the corresponding inequalities. Finally a more exhaustive study may be conducted by considering sequences $\{n_j\}$ other than those treated here.

REFERENCES

[1] G. A. Bessmertnyh and A. Ju. Levin, *Some inequalities satisfied by differentiable functions of one variable*, Dokl. Akad. Nauk SSSR, 144 (1962), pp. 471–474 = Soviet Math. Dokl., 3 (1962), pp. 737–740.

[2] W. A. Coppel, *Disconjugacy*, Springer-Verlag, New York, 1971.

[3] F. R. Gantmacher, *The Theory of Matrices*, vol. I, Chelsea, New York, 1960.

[4] Philip Hartman, *Principal solutions of disconjugate n-th order linear differential equations*, Amer. J. Math., 91 (1969), pp. 306–362; *Corrigendum and addendum*, Ibid., 93 (1971), pp. 439–451.

[5] Philip Hartman, *Ordinary Differential Equations*, Hartman, Baltimore, 1973.

[6] S. Karlin and W. J. Studden, *Tchebycheff Systems with Applications in Analysis and Statistics*, Interscience, New York, 1966.

[7] A. JU. LEVIN, *Some problems bearing on the oscillation of solutions of linear differential equations*, Dokl. Akad. Nauk SSSR, 148 (1963), pp. 512–515 = Soviet Math. Dokl., 4 (1963), pp. 121–124.
[8] ———, *A bound for a function with monotonely distributed zeros of successive derivatives*, Mat. Sb., 64, 106 (1964), pp. 396–409.
[9] ———, *Non-oscillation of solutions of the equations* $x^{(n)}+p_1(t)x^{(n-1)}+\cdots+p_n(t)x=0$, Uspehi Mat. Nauk., 24 (1969), pp. 43–100 = Russian Math Surveys, 24 (1969), pp. 43–100.
[10] J. S. MULDOWNEY, *On an inequality of Čaplygin and Pólya*, Proc. Royal Irish Acad. Sect. A, 76 (1976), pp. 85–99.
[11] ———, *A disconjugacy criterion for linear scalar differential operators*, Proc. Amer. Math. Soc., 65 (1977), pp. 93–96.
[12] ———, *Comparison theorems for linear boundary value problems*, this Journal, 9 (1978), pp. 943–955.
[13] ———, *Linear differential inequalities*, this Journal, 9 (1978), pp. 106–120.
[14] ———, *A necessary and sufficient condition for disfocality*, Proc. Amer. Math. Soc., 74 (1979), pp. 49–55.
[15] ZEEV NEHARI, *Disconjugate linear differential operators*, Trans. Amer. Math. Soc., 129 (1967), pp. 500–576.
[16] G. PÓLYA, *On the mean-value theorem corresponding to a given linear homogeneous differential equation*, Trans. Amer. Math. Soc., 24 (1922), pp. 312–324.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0008 $01.00/0

VARIATIONAL INEQUALITIES IN SEQUENTIAL ANALYSIS*

AVNER FRIEDMAN†

Abstract. Several models in sequential analysis are studied by reducing them to variational inequalities.

Introduction. Given a stochastic process with some unknown parameter of its probability distribution, several hypotheses are introduced regarding the true value of the parameter. One then begins to observe the process (i.e., take samples of the random variables) and after some time τ make a decision as to which hypothesis to accept. The point of view of sequential analysis [17] is that one should choose a random τ, depending on the past observations, which is "as small as possible"; account must be taken that the *risk* W due to accepting an incorrect hypothesis satisfies some a priori constraints.

There are different ways of incorporating the risk W. One way is to consider the total cost

$$J = E\tau + EW \qquad (E = \text{expectation}); \tag{0.1}$$

J depends on τ and on a final decision δ of accepting an hypothesis ($W = W(\delta)$). The problem is then to minimize J. Another model is:

$$\text{minimize } E\tau, \text{ given the restriction } EW \leqq \lambda, \tag{0.2}$$

where λ is prescribed.

If in (0.1) we first minimize on δ, then we are left with a problem of

$$\text{minimize } \tilde{J}(\tau), \qquad \tilde{J}(\tau) = \min_{\delta} J. \tag{0.3}$$

A large class of stopping time problems with respect to diffusion Markov processes are known to lead to variational inequalities [5], [12] and to quasi-variational inequalities [1], [3], [4]. The stopping time problems of the type (0.3) are *not* Markovian. However, using the theory of filters they can sometimes be reduced to Markovian problems and then studied by methods of variational inequalities. In this paper we give such models.

In §1 we consider a model involving two composite hypotheses. This is generalized in §2 to three composite hypotheses. In § 3 we consider a model with m simple hypotheses, $m \geqq 2$. Finally in § 4 we give an example of type (0.2) which can be reduced to a Stefan problem.

1. Two composite hypotheses. A one-dimensional stochastic process $z(t)$ is to be observed. It is known to have the probability distribution of a Brownian motion with drift μ and variance σ^2; σ^2 is given but μ is unknown. We do know, however, that μ has the normal law $N(\mu_0, \sigma_0^2)$. The two composite hypotheses are

H_1: accept that $\mu > 0$,

H_2: accept that $\mu < 0$.

* Received by the editors May 5, 1980. This work was partially supported by the National Science Foundation under grant MCS 791 5171.

† Mathematics Department, Northwestern University, Evanston, Illinois 60201.

Let δ be a variable taking the values $\delta = 1$ if H_1 is accepted and $\delta = 2$ if H_2 is accepted. We define the risk function

$$W(\mu, \delta) = k|\mu| \quad \text{if } \delta = 1, \mu < 0 \quad \text{or } \delta = 2, \mu > 0,$$

$$= 0 \qquad \text{in all other cases,}$$

where k is a positive constant. If $c > 0$ is the cost of observation per unit time, then the total cost of observation and accepting some hypothesis is

$$J_{\mu_0}(\tau, \delta) = E[c\tau + W(\mu, \delta(\omega))], \tag{1.1}$$

where $\tau = \tau(\omega)$ is the random time of observation and $\delta(\omega)$ is the final decision as to which hypothesis to accept.

Here τ is to be a stopping time with respect to the σ-fields $\mathcal{F}_t = \sigma(z(s), 0 \leqq s \leqq t)$ and $\delta(\omega)$ is to be $\mathcal{F}_\tau$-measurable.

The process $z(t)$ is not a Markov process, but we can write

$$dz(t) = y\,dt + \sigma\,dw(t), \qquad z(0) = 0,$$

where $y = \mu$ or, more specifically,

$$y = \mu_0 + \xi,$$

and ξ is $N(0, \sigma_0^2)$ variable; here $w(t)$ is a Brownian motion independent of the random variable ξ.

Using the Kalman–Bucy linear filtering theory [7], [16] we introduce the *filter*

$$\hat{y}(t) = E[y|\mathcal{F}_t]$$

and the *error*

$$\varepsilon(t) = y - \hat{y}(t);$$

they satisfy the stochastic differential equations

$$d\hat{y}(t) = \frac{p(t)}{\sigma}\,d\hat{w}(t), \qquad \hat{y}(0) = \mu_0, \tag{1.2}$$

where $\hat{w}(t)$ is a Brownian motion with respect to $\mathcal{F}_t$ and

$$d\varepsilon(t) = -\frac{p(t)}{\sigma^2}\,\varepsilon(t)\,dt - \frac{p(t)}{\sigma}\,dw(t), \qquad \varepsilon(0) = \xi; \tag{1.3}$$

$p(t)$ is a solution of the Ricatti equation

$$p'(t) = -\frac{p^2(t)}{\sigma^2}, \qquad p(0) = \sigma_0^2.$$

Hence

$$p(t) = \frac{1}{t/\sigma^2 + 1/\sigma_0^2}.$$

We can now solve (1.3):

$$\varepsilon(t) = \frac{\xi/\sigma_0^2}{t/\sigma^2 + 1/\sigma_0^2} - \frac{w(t)/\sigma}{t/\sigma^2 + 1/\sigma_0^2}.$$

Set

$$\phi(u)=\frac{e^{-u^2/2}}{\sqrt{2\pi}}, \qquad \Phi(u)=\int_{-\infty}^{u}\frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt \tag{1.4}$$

and

$$s=\frac{1}{t/\sigma^2+1/\sigma_0^2}. \tag{1.5}$$

By direct computation we find that

$$\begin{aligned} EW(x+\varepsilon(t),2) &= Ek|x+\varepsilon(t)|I_{x+\varepsilon(t)>0} \\ &= k\left[\sqrt{s}\phi\left(\frac{x}{\sqrt{s}}\right)+x\Phi\left(\frac{x}{\sqrt{s}}\right)\right]. \end{aligned} \tag{1.6}$$

Similarly,

$$EW(x+\varepsilon(t),1)=k\left[\sqrt{s}\phi\left(\frac{x}{\sqrt{s}}\right)-x\left(1-\Phi\left(\frac{x}{\sqrt{s}}\right)\right)\right]. \tag{1.7}$$

It follows that

$$\min_{\delta} W(x+\varepsilon(t),\delta)=\begin{cases} W(x+\varepsilon(t),1) & \text{if } x\geqq 0, \\ W(x+\varepsilon(t),2) & \text{if } x\leqq 0. \end{cases} \tag{1.8}$$

Set

$$\psi(u)=\begin{cases} \phi(u)+u\Phi(u) & \text{if } u\leqq 0, \\ \phi(u)-u(1-\Phi(u)) & \text{if } u\geqq 0, \end{cases} \tag{1.9}$$

$$\Psi(x,s)=k\sqrt{s}\,\psi\left(\frac{x}{\sqrt{s}}\right). \tag{1.10}$$

Then

$$\min_{\delta} W(x+\varepsilon(t),\delta)=\psi(x,s).$$

The "non-Markovian" cost $J_{\mu_0}(\tau,\delta)$ can now be reduced to the "Markovian" cost

$$J_{\mu_0}(\tau)=E[c\tau+\Psi(\hat{y}(\tau),s(\tau))], \qquad \hat{y}(0)=\mu_0, \tag{1.11}$$

where $s=s(t)$ is defined by (1.5); this is done in [5] for general cost functions.

The problem of minimizing the cost function (1.11) is a standard stopping time problem. We first consider the truncated problem

$$U^T(x,t)=\inf_{t\leqq\tau\leqq T} J_x(\tau) \qquad (0<T<\infty).$$

Then U^T satisfies the variational inequality

$$U_t^T + \frac{1}{2}\frac{s^2(t)}{\sigma^2} U_{xx}^T + c \geqq 0,$$

$$U^T(x,t) \leqq \Psi(x, s(t)), \tag{1.12}$$

$$\left(U_t^T + \frac{1}{2}\frac{s(t)}{\sigma^2} U_{xx}^T + c\right)(U^T - \Psi(x, s(t))) = 0$$

and the terminal condition

$$U^T(x,T) = \Psi(x, s(T)). \tag{1.13}$$

Setting

$$u^T(x,s) = U^T(x,t)$$

and taking $T \uparrow \infty$, we find that $u^T(x,s) \to u(x,s)$, where u is the solution of

$$u_s - \tfrac{1}{2} u_{xx} \leqq \frac{c\sigma^2}{s^2},$$

$$u(x,s) \leqq \Psi(x,s), \tag{1.14}$$

$$\left(u_s - \tfrac{1}{2} u_{xx} - \frac{c\sigma^2}{s^2}\right)(u - \Psi(x,s)) = 0,$$

a.e. in $x \in R^1$, $0 < s < \sigma_0^2$, and $u \geqq 0$. From the definition of U^T it also follows that

$$u(\mu_0, \sigma_0^2) = \inf_\tau J_{\mu_0}(\tau). \tag{1.15}$$

The sets C: $u < \Psi$ and S: $u = \psi$ are called, respectively, the *continuation* set (or noncoincidence set) and the *stopping* set (or coincidence set). Consider these sets in the variables (x,t), rather than the variables (x,s); then the optimal stopping time τ is the first time $(\hat{y}(t), t)$ hits the set S. Further, the optimal δ is given by (in view of (1.8))

$$\delta = 1 \quad \text{if } \hat{y}(\tau) \geqq 0, \qquad \delta = 2 \quad \text{if } \hat{y}(\tau) \leqq 0.$$

We shall now briefly study the variational problem (1.14). Since

$$\Psi(x,s) = k \min\left\{\sqrt{s}\phi\left(\frac{x}{\sqrt{s}}\right) + \frac{x}{\sqrt{s}}\Phi\left(\frac{x}{\sqrt{s}}\right), \sqrt{s}\phi\left(\frac{x}{\sqrt{s}}\right) - \frac{x}{\sqrt{s}}\left(1 - \Phi\left(\frac{x}{\sqrt{s}}\right)\right)\right\} \tag{1.16}$$

for all $x \in R$, $s > 0$, we have

$$\Psi_x(0+, s) < 0, \qquad \Psi_x(0-, s) > 0 \tag{1.17}$$

and

$$\frac{\partial^2}{\partial x^2}\Psi \leqq C \tag{1.18}$$

in the distribution sense. By a regularity result for variational inequalities [6] we conclude that

$$u,\ u_x,\ u_{xx},\ u_t \text{ are locally in } L^\infty. \tag{1.19}$$

Observing that

$$\Psi_s - \tfrac{1}{2}\Psi_{ss} = 0 \quad \text{if } x \neq 0,$$

we can write for $w = u - \Psi$:

$$w_s - \tfrac{1}{2} w_{ss} = \frac{c\sigma^2}{s^2} \quad \text{in } C.$$

Also, by (1.17),

$$w_x|_{x=0+} > 0 \qquad (\text{since } u_x|_{x=0} = 0 \text{ by symmetry}).$$

Since finally $w(x, 0) = 0$ and grad $w = 0$ at the points of ∂C where $x > 0$, $s > 0$, we can apply the maximum principle to w_x and conclude that

$$w_x > 0 \quad \text{in } C \cap \{x > 0\}. \tag{1.20}$$

Similarly,

$$w_s < 0 \quad \text{in } C. \tag{1.21}$$

For the proof it is convenient to work with the "penalized problem"

$$w_s^\varepsilon - \tfrac{1}{2} w_{xx}^\varepsilon + \beta_\varepsilon(w^\varepsilon) = \frac{c\sigma^2}{(s+\varepsilon)^2}, \tag{1.22}$$

where $\beta_\varepsilon'(t) \geqq 0$, $\beta_\varepsilon(t) \to 0$ if $t < 0$, $\varepsilon \to 0$ and $\beta_\varepsilon(t) \to \infty$ if $t > 0$, $\varepsilon \to 0$. Notice that (1.22) holds for $x \neq 0$; since, however,

$$\frac{\partial}{\partial x} \frac{\partial}{\partial s} \Psi(x, s) \text{ is continuous across } x = 0,$$

the equation obtained for w_s^ε after differentiating (1.22) with respect to s holds for all x. Since $w_s^\varepsilon \leqq 0$ for $s = 0$ and w_s^ε is continuous, the maximum principle gives $w_s^\varepsilon \leqq 0$; hence, $w_s \leqq 0$ a.e. and (1.21) follows.

From (1.20), (1.21) it follows that the free boundary (i.e., the boundary of C in $s >$) is given by the two curves

$$x = \pm\zeta(s), \qquad \zeta(s) \text{ monotone increasing};$$

$\zeta(s)$ is C^∞ as in the case of the Stefan problem.

Recently Knerr [14] proved that

$$s^{-2}\zeta(s) \to \gamma \quad \text{if } s \to 0 \qquad (\gamma > 0); \tag{1.23}$$

he also studied more general variational inequalities of the type (1.14).

Chernoff [10], [11] has formally derived the variational inequality (1.14) as a limit of the corresponding discretized problem. He also proved an asymptotic formula

$$s^{-1/2}\zeta(s) \sim \gamma s^{3/2}\{1 + a_1 s^3 + a_2 s^6 + \cdots\}$$

as $s \to 0$; this includes (1.23).

2. Three composite hypotheses. The assumptions on $z(t)$, μ, σ are the same as in § 1, but we now make three composite hypotheses:

H_1: $\mu > a$,

H_2: $-a < \mu < a$,

H_3: $\mu < -a$,

where a is a given positive constant. The risk function is defined by

$$W(\mu, 1) = k(a-\mu) \quad \text{if } \mu < a,$$
$$W(\mu, 3) = k(\mu+a) \quad \text{if } \mu > -a,$$
$$W(\mu, 2) = \mu - a \quad \text{if } \mu > a,$$
$$W(\mu, 2) = -a - \mu \quad \text{if } \mu < -a,$$
$$W(\mu, i) = 0 \quad \text{in all other cases.}$$

We compute

$$\psi_1(x, s) = EW(x+\varepsilon(t), 1) = k\left[\sqrt{s}\phi\left(\frac{a-x}{\sqrt{s}}\right) + (a-x)\Phi\left(\frac{a-x}{\sqrt{s}}\right)\right],$$

$$\psi_3(x, s) = EW(x+\varepsilon(t), 3) = k\left[\sqrt{s}\phi\left(\frac{a+x}{\sqrt{s}}\right) + (a+x)\Phi\left(\frac{a+x}{\sqrt{s}}\right)\right],$$

$$\psi_2(x, s) = EW(x+\varepsilon(t), 2) = k\left[\sqrt{s}\phi\left(\frac{x-a}{\sqrt{s}}\right) + (x-a)\Phi\left(\frac{x-a}{\sqrt{s}}\right) + \sqrt{s}\phi\left(\frac{a+x}{\sqrt{s}}\right) - (a+x)\Phi\left(-\frac{a+x}{\sqrt{s}}\right)\right].$$

In order to decide, for $x > 0$, which is the best hypothesis (it obviously should be either H_1 or H_2) we consider the function

$$\phi(x, s) = \psi_2(x, s) - \psi_1(x, s)$$

$$= \frac{k}{\sqrt{2\pi}}\left\{(x-a)\int_{-\infty}^{(x-a)/\sqrt{s}} e^{-\lambda^2/2}\,d\lambda + \sqrt{s}\,e^{-(a+x)^2/2s} - (a+x)\int_{-\infty}^{(a+x)/\sqrt{s}} e^{-\lambda^2/2}\,d\lambda + (x-a)\int_{-\infty}^{(a-x)/\sqrt{s}} e^{-\lambda^2/2}\,d\lambda\right\}. \tag{2.1}$$

We compute

$$\phi_s = \frac{k}{\sqrt{2\pi}}\frac{1}{2\sqrt{s}}e^{-(a+x)^2/2s}, \tag{2.2}$$

so that

$$\phi_s > 0. \tag{2.3}$$

One can also verify that

$$\phi_x = \frac{k}{\sqrt{2\pi}}\left(\int_{-\infty}^{(x-a)/\sqrt{s}} + \int_{-\infty}^{(a-x)/\sqrt{s}} - \int_{-\infty}^{-(a+x)/\sqrt{s}}\right) e^{-\lambda^2/2}\,d\lambda \tag{2.4}$$

$$= k - \frac{k}{\sqrt{2\pi}}\int_{-\infty}^{-(a+x)/\sqrt{s}} e^{-\lambda^2/2}\,d\lambda,$$

so that

$$\phi_x > 0 \quad \text{if } x \geqq 0. \tag{2.5}$$

From (2.3), (2.5) we conclude that the curve

$$\Gamma = \{(x, s);\ x > 0,\ \phi(x, s) = 0\}$$

is given by

$$x = \gamma(s), \qquad \gamma(s) \text{ monotone decreasing.} \tag{2.6}$$

This curve intersects the x-axis at $x = a$ and the s-axis at a point s^* determined by

$$2a\int_{-\infty}^{-a/\sqrt{s^*}} e^{-\lambda^2/2}\,d\lambda + a\int_{-\infty}^{a/\sqrt{s^*}} e^{-\lambda^2/2}\,d\lambda = \sqrt{s^*}\,e^{-a^2/2(s^*)^2}.$$

From (2.2), (2.4) one also deduces that

$$\gamma^{(j)}(0) = 0 \quad \text{for all } j;$$

$$\gamma'(s) = -\frac{1}{2\sqrt{\pi s}} e^{-2a^2/s}(1 + O(s)).$$

Finally, $-\infty < \gamma'(s^*) < 0$.

Denote by Ω_2 the domain given by

$$0 < x < \gamma(s), \qquad 0 < s < s^*,$$

and by Ω_1 the complement of Ω_2 in the quadrant $x > 0$, $s > 0$.

Set

$$\Psi(x, s) = \min_{1 \leqq i \leqq 3} \psi_i(x, s). \tag{2.7}$$

As in § 1 one can show that the optimal cost is $u(\mu_0, \sigma_0^2)$, where u is the solution of the variational inequality (1.14). Furthermore, (1.19) holds; the optimal τ is the hitting time by $(\hat{y}(t), t)$ of the stopping set $\{u = \Psi\}$ and, if $\hat{y}(\tau) > 0$, the optimal δ is:

accept H_1 if $(\hat{y}(\tau), \tau) \in \Omega_1$,

accept H_2 if $(\hat{y}(\tau), \tau) \in \Omega_2$.

The set

$$\Gamma' = \Gamma \cup I \qquad (I = \{(s, 0),\ s > s^*\})$$

is called the *ridge* of Ψ.

LEMMA 2.1. *The ridge belongs to the continuation set C.*

For such a result, in another variational inequality, see [8]; the proof here is similar.

From the lemma it follows that the free boundary in $x \geqq 0$ must lie in $\Omega_1 \cup \Omega_2$; i.e., the free boundary does not intersect the ridge.

LEMMA 2.2.

$$(u - \psi_1)_x \geqq 0 \quad \textit{if } x > 0, \tag{2.8}$$

$$(u - \psi_2)_x \leqq 0 \quad \textit{if } x > 0, \tag{2.9}$$

$$(u - \psi_2)_s \leqq 0 \quad \textit{for all } x. \tag{2.10}$$

The proof uses the inequalities (2.3), (2.5).

From Lemmas 2.1, 2.2 follows:

THEOREM 2.3 *The free boundary in $x \geqq 0$ is given by two curves*:

$$x = \zeta_1(s) \quad \textit{lying in } \Omega_1,$$

$$x = \zeta_2(s) \quad \textit{lying in } \Omega_2,$$

$$\zeta_2(s) < \zeta_1(s),$$

$$\zeta_2(s) \textit{ is monotone decreasing.}$$

One can further show that

$$c_1 \leqq \frac{(-1)^{i-1}(\zeta_i(s) - a)}{s^2} \leqq c_2 \qquad (c_1, c_2 \text{ are positive}).$$

The questions whether

$$\lim \frac{\zeta_i(s)-a}{s^2} \text{ exists}$$

and whether $\zeta_1(s)$ is monotone have not yet been considered.

3. Several simple hypotheses. In this section we consider an m-dimensional stochastic process $z(t)$ which is an m-dimensional Brownian motion with a drift; the drift can be any one of given $n+1$ m-vectors $\lambda_0, \lambda_1, \cdots, \lambda_n$. We also assume that the drift may change in time as follows. There is a Markov process $\theta(t)$ with $n+1$ states $0, 1, \cdots, n$ and an infinitesimal matrix $Q=(q_{ij})$ such that

$$\text{if } \theta(t)=j, \text{ then the drift is } \lambda_j.$$

Thus

$$dz(t)=dw(t)+\sum_{j=0}^{n} I_{\theta(t)=j}\lambda_j\, dt. \tag{3.1}$$

The Brownian motion $w(t)$ and the Markov process $\theta(t)$ are assumed to be independent. The hypotheses are:

(3.2) accepting H_j at time t means accepting that $\theta(t)=j$.

Following the procedure of §§ 1–2 we introduce the sampling cost $c\tau$ and the risk W, where

$$W(\theta,\delta)=a_i \quad \text{if } \delta=i \quad \text{but } \theta\neq i \qquad (a_i>0).$$

The total cost is

$$J_\pi(\tau,\delta)=E_\pi[c\tau+W((\theta(\tau),\delta(\omega))], \tag{3.3}$$

where the index $\pi=(\pi_0,\pi_1,\cdots,\pi_n)$ indicates that the initial distribution of θ is $\theta(0)=j$ with probability π_j.

Setting

$$J_\pi(\tau)=\inf_\delta J_\pi(\tau,\delta),$$

one can now proceed to study the problem of minimizing $J_\pi(\tau)$ by introducing the nonlinear filter

$$\pi_j(t)=P_\pi[\theta(t)=j|\mathcal{F}_t] \qquad (0\leqq j\leqq n), \tag{3.4}$$

where $\mathcal{F}_t=\sigma(z(s),\ 0\leqq s\leqq t)$. Set

$$\Pi=\{\pi=(\pi_0,\cdots,\pi_n);\ \pi_i>0, \textstyle\sum \pi_i=1\}.$$

It is known [1], [16] that $\pi(t)=(\pi_0(t),\cdots,\pi_n(t))$ is a Markov diffusion process in Π with generator

$$Mu(\pi)=\tfrac{1}{2}\sum_{i,j=0}^{n}\pi_i\pi_j\left(\lambda_i-\sum_{k=0}^{n}\lambda_k\pi_k\right)\cdot\left(\lambda_j-\sum_{l=0}^{n}\lambda_l\pi_l\right)\frac{\partial^2 u(\pi)}{\partial\pi_i\partial\pi_j}$$
$$+\sum_{i,j=0}^{n} q_{ij}\pi_i\frac{\partial u(\pi)}{\partial\pi_j}.$$

This elliptic operator is nondegenerate in Π if and only if the vectors

$$\lambda_1-\lambda_0,\quad \lambda_2-\lambda_0,\quad \cdots,\quad \lambda_n-\lambda_0 \tag{3.5}$$

are linearly independent (which implies that $m \geqq n$). However, M is always degenerate on all of $\partial\Pi$.

The study of the function

$$u(\pi)=\inf_{\tau} J_{\pi}(\tau)$$

can be reduced to that of the variational inequality:

$$Mu+c \geqq 0,\quad u(\pi) \leqq g(\pi),\quad (Mu+c)(u-g)=0,$$

a.e. in Π, where

$$g(\pi)=\min_{0 \leqq i \leqq n} a_i(1-\pi_i).$$

THEOREM 3.1 [9]. *If* $q_{i,j} \equiv 0$, *then*

$$u(\pi) \sim \left(\sum_{i=0}^{n} \gamma_i \pi_i\right) c \log \frac{1}{c} \quad \text{if } c \to 0,$$

where

$$\gamma_i=\frac{2}{[\min_{k \neq i}|\lambda_k-\lambda_i|]^2}.$$

The *ridge* of g (i.e., the set where ∇g is discontinuous) divides Π into $n+1$ domains R_j; each R_j is a convex polyhedron and ∂R_j contains precisely one vertex of Π, say V_j.

THEOREM 3.2 [9]. *If* $q_{i,j} \equiv 0$ *then the stopping set consists of* $n+1$ *convex domains* S_j; S_j *lies in* R_j *and contains a* Π*-neighborhood of* V_j.

The proofs of both Theorems 3.1 and 3.2 are probabilistic. If we make, however, the assumption that the vectors in (3.5) are linearly independent, then we can use elliptic estimates in order to obtain further results (also in case $q_{i,j} \not\equiv 0$). For example (see [9]) one can prove, in this case, that the free boundary is analytic.

Theorems 3.1 and 3.2 should extend to other models, for instance, when $z(t)$ is a Poisson process with $n+1$ possible parameters.

4. A problem of type (0.2). Consider a process

$$z(t)=w(t)+\theta t, \tag{4.1}$$

where $w(t)$ is a one-dimensional Brownian motion and θ is unknown, $-\infty<\theta<\infty$; we observe the process $z(t)$ and have to choose between the two simple hypotheses:

H_1: accept $\theta=\theta_1$,

H_2: accept $\theta=\theta_2$.

Here $\theta_2=-\theta_1<0$. For any fixed θ, denote by P_θ the probability, on the space $C[0,\infty)$, determined by the process (4.1). We impose the following restriction on the pair of decision variables (τ, δ):

τ is a stopping time with respect to the σ-fields $\mathcal{F}_t=\sigma(z(s), 0 \leqq s \leqq t)$, and δ is such that

$$P_{\theta_i}[H_i \text{ is rejected}] \leqq \lambda, \tag{4.2}$$

where $\lambda < \frac{1}{2}$ is given. The objective now is to choose τ such that

$$\tau \text{ minimizes } \max_{-\infty<\theta<\infty} E_\theta \tau. \tag{4.3}$$

The motivation for this problem is given in [2].

Weiss [18] has shown that if we restrict δ to be "symmetric" then (τ, δ) solves (4.3) if there exists a number $p \in (0, 1)$ such that (τ, δ) minimizes the functional

$$J = \frac{p}{2} P_{\theta_2}[H_1 \text{ is accepted}] + \frac{p}{2} P_{\theta_1}[H_2 \text{ is accepted}] + (1-p)E_0\tau.$$

An easy calculation shows (see Lai [15]) that

$$J = E_x g(w(\tau), \tau) \equiv J_x(\tau), \qquad w(0) = x, \tag{4.4}$$

where

$$g(x, t) = \alpha t + \exp\left(-\frac{t}{2} - |x|\right), \qquad \alpha = \frac{2(1-p)}{p}. \tag{4.5}$$

But then the function

$$u(x, t) = \inf_\tau J_x(\tau)$$

is the solution of the variational inequality

$$u_t + \tfrac{1}{2}u_{xx} \geqq 0, \quad u \leqq g(x, t), \quad (u_t + \tfrac{1}{2}u_{xx})(u - g) = 0 \tag{4.6}$$

a.e. in $x \in R^1$, $t > 0$.

By symmetry, $u_x(x, 0) = 0$. Setting

$$w = \frac{\partial}{\partial t}(u - g),$$

one can prove that $w \geqq 0$, and then easily verify that (4.6) is equivalent to following Stefan problem:

$$\begin{aligned}
&w_t + \tfrac{1}{2}w_{xx} = 0 && \text{if } 0 < x < s(t), \quad t > 0,\\
&w_x(0, t) = -\tfrac{1}{2}e^{-t/2} && \text{if } t > 0,\\
&w(s(t), t) = 0 && \text{if } t > 0,\\
&w_x(s(t), t) = \alpha \dot{s}(t) && \text{if } t > 0.
\end{aligned} \tag{4.7}$$

The strong maximum principle implies that $w > 0$ if $0 < x < s(t)$ and $\dot{s}(t) < 0$.

Lai [15] studied the free boundary for (4.7). He proved:

THEOREM 4.1. *For fixed* α,

$$s(t) \sim \frac{1}{2\alpha} e^{-t/2} \quad \text{as } t \to \infty; \tag{4.8}$$

for $\alpha \to 0$,

$$\begin{aligned}
&s\left(t \log \frac{1}{\alpha}\right) \sim \left(1 - \frac{t}{2}\right) \log \frac{1}{\alpha} && \text{if } 0 \leqq t < 2,\\
&s\left(t \log \frac{1}{\alpha}\right) \sim \tfrac{1}{2}\alpha^{t/2-1} && \text{if } t > 2,\\
&s\left(2 \log \frac{1}{\alpha}\right) && \text{is a positive number independent of } \alpha.
\end{aligned} \tag{4.9}$$

We shall now complement this result with the following theorem.

THEOREM 4.2. *The free boundary for* (4.7) *is convex*; *i.e.*, $\dot{s}(t)$ *is monotone decreasing.*

This will follow from a more general result for the Stefan problem:

$$
\begin{aligned}
&u_t - \tfrac{1}{2}u_{xx} = 0 && \text{if } 0 < x < s(t), \quad 0 < t < \infty,\\
&u(x, 0) = h(x) && \text{if } 0 < x < s(0) = b, \quad b > 0,\\
&u_x(0, t) = f(t) && \text{if } t > 0,\\
&u(s(t), t) = s(t) && \text{if } t > 0,\\
&u_x(s(t), t) = -\alpha \dot{s}(t) && \text{if } t > 0 \qquad (\alpha > 0).
\end{aligned} \tag{4.10}
$$

We shall assume:

$$
h(x) > 0, \quad h'(x) < 0, \quad h''(x) > 0, \quad \frac{h''(x)}{h'(x)} \text{ is decreasing for } 0 < x < b, \tag{4.11}
$$

$$
h(b) = 0,
$$

$$
h'(0) = f(0), \qquad h''(b) = (h'(b))^2, \tag{4.12}
$$

$$
f < 0, \qquad f' \leqq 0. \tag{4.13}
$$

The conditions in (4.12) ensure that u_t and u_{xx} are continuous functions up to the boundary (at $t = 0$).

THEOREM 4.3. *If* (4.11)–(4.13) *hold then* $\dot{s}(t)$ *is monotone decreasing.*

Proof. The proof is based on an extension of a method of Friedman and Jensen [13] for proving convexity in case the condition at $x = 0$ is $u(0, t) =$ constant. Analogously to [13], the conditions of the theorem ensure that

$$
u_x < 0, \qquad u_t > 0
$$

and

$$
\left.\frac{u_t}{u_x}\right|_{t=0} \text{ is decreasing.}
$$

The function $z = u_t/u_x$ satisfies a parabolic equation, and we consider the regular curves Γ_β: $z = \beta$ which initiate on the free boundary and go into the domain where $u > 0$. As shown in [13], z cannot take a local extremum at a point of the free boundary; hence, Γ_β cannot end at a point on the free boundary and, further, its t-coordinate is monotone decreasing. Suppose

Γ_{β_1} starts at $(s(t_1), t_1)$,

Γ_{β_2} starts at $(s(t_2), t_2)$, and $t_2 > t_1$.

If Γ_{β_1}, Γ_{β_2} both end on $t = 0$, say at x_1 and x_2 respectively, then $x_2 < x_1$ and, consequently,

$$
\beta_1 = \frac{h''(x_1)}{h'(x_1)} < \frac{h''(x_2)}{h'(x_1)} = \beta_2.
$$

Since $z = -\dot{s}(t)$ on the free boundary, we get

$$
-\dot{s}(t_1) < -\dot{s}(t_2). \tag{4.14}
$$

We now make the crucial observation that

(4.15) z cannot take a minimum on $x=0$.

Indeed, on $x=0$,

$$z_x=\frac{u_{tx}}{u_x}-\frac{u_tu_{xx}}{u_x^2}=\frac{(u_x)_t}{u_x}-\frac{u_t^2}{u_x^2}=\frac{f'}{f}-z^2\leqq 0 \quad \text{by (4.13).}$$

Suppose now that Γ_{β_1}, Γ_{β_2} both end on the t-axis, say at $\bar{t}_1$ and $\bar{t}_2$ respectively. Since $t_2>t_1$, also $\bar{t}_2>\bar{t}_1$, and there is a region Q bounded by

$$\bar{t}_1\leqq t\leqq \bar{\bar{t}}_2, \quad x=0; \qquad t=\bar{\bar{t}}_2; \qquad \Gamma_{\beta_1}\cap\{t<\bar{\bar{t}}_2\},$$

where $\bar{\bar{t}}_2\leqq \bar{t}_1$, such that

$$\bar{Q}\cap\Gamma_{\beta_2}=\partial Q\cap\Gamma_{\beta_2}=\{t=\bar{\bar{t}}_2\}\cap\Gamma_{\beta_2},$$

and this intersection is nonempty. In view of (4.15), the minimum of z in $\bar{Q}$ is attained on Γ_{β_1} and consequently $\beta_2>\beta_1$, again giving (4.14).

Consider finally the case where

$$\Gamma_{\beta_1} \text{ ends on } t=0, \qquad \Gamma_{\beta_2} \text{ ends on } x=0.$$

Then we can introduce curves Γ_β that end arbitrarily close to $(0, 0)$ on the x-axis and similar curves Γ_β, that end near $(0, 0)$ on the t-axis. Applying the previous results to the pairs Γ_{β_1}, Γ_β and to $\Gamma_{\beta'}$, Γ_{β_2}, completes the proof.

Proof of Theorem 4.2. We can consider w as a limit of solutions w_T of truncated problems in $t<T$ with $T\to\infty$; the terminal condition is

$$w(x, T)=\varepsilon^2(b-x)+\frac{\varepsilon}{2}(b-x)^2 \quad \text{for } 0<x<b,$$

and we choose ε, b so that

$$\tfrac{1}{2}e^{-T/2}=\varepsilon^2+\varepsilon b, \qquad \text{(and say } b=\varepsilon\text{)}.$$

Changing variables $\tau=T-t$ we arrive at the setting of Theorem 4.3 with

$$f(\tau)=-ce^{-\tau/2}, \qquad c=\tfrac{1}{2}e^{-T/2},$$

and the conclusion follows.

REFERENCES

[1] R. F. ANDERSON AND A. FRIEDMAN, *Multi-dimensional quality control problems and quasi-variational inequalities*, Trans. Amer. Math. Soc., 246 (1978), pp. 31–76.

[2] R. E. BECHHOFER, *Optimal stochastic control*, Sankhyā, Ser. A, 30 (1968), pp. 221–252.

[3] A. BENSOUSSAN AND A. FRIEDMAN, *Nonzero sum stochastic differential games with stopping times and new free boundary problems*, Trans. Amer. Math. Soc., 231 (1977), pp. 275–327.

[4] A. BENSOUSSAN AND J. L. LIONS, *Nouvelles methods–contrôle impulsionnel*, Appl. Math. Optimization , 1 (1975), pp. 289–312.

[5] ———, *Variationnelles en Contrôle Stochastique*, Dunod, Paris, 1978.

[6] H. BREZIS AND D. KINDERLEHRER, *The smoothness of solutions to nonlinear variational inequalities*, Indiana Univ. Math. J., 23 (1974), pp. 831–844.

[7] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Applications to Guidance*, Interscience, New York, 1968.

[8] L. A. CAFFARELLI AND A. FRIEDMAN, *The free boundary for elastic-plastic torsion problems*, Trans. Amer. Math. Soc., 252 (1979), pp. 65–97.

[9] ———, *Sequential testing of several simple hypotheses for a diffusion process and the corresponding free boundary problem*, Pacific J. Math., to appear.
[10] H. CHERNOFF, *Optimal stochastic control*, Sankhyā, Ser. A, 30 (1968), pp. 221–252.
[11] ———, *Sequential Analysis and Optimal Design*, Regional Conference Series in Applied Mathematics, 8, Society for Industrial and Applied Mathematics, Philadelphia, 1972.
[12] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, vol. 2, Academic Press, New York, 1976.
[13] A. FRIEDMAN AND R. JENSEN, *Convexity of the free boundary in the Stefan problem and in the dam problem*, Arch. Rat. Mech. Anal., 67 (1978), pp. 1–24.
[14] B. F. KNERR, *A singular free boundary problem*, Illinois J. Math., 23 (1979), pp. 438–458.
[15] T. L. LAI, *Optimal stopping and sequential tests which minimize the maximum expected sample size*, Ann. Statist., 1 (1973), pp. 659–673.
[16] R. S. LIPSTER AND A. N. SHIRYAEV, *Statistics of Random Processes*, vols. 1, 2, Springer-Verlag, Berlin, 1977, 1978.
[17] A. WALD, *Sequential Analysis*, John Wiley, New York, 1947.
[18] L. WEISS, *On sequential tests which minimize the maximum expected sample size*, J. Amer. Statist. Assoc., 57 (1962), pp. 551–556.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0009 $01.00/0

PRINCIPAL SOLUTIONS FOR LIÉNARD'S EQUATION*

DONALD C. BENSON†

Abstract. It is found that for a certain class of nonlinear ordinary differential equations there exists a one-parameter family of solutions which display lower order growth at infinity than the remaining solutions. This result is found with the help of an auxiliary equation which plays a role similar to that of the Riccati equation in the linear case. Results on the oscillation of Liénard's equation are included.

1. Introduction. In [1] this author gives a generalization of Picone's identity (see [8, p. 5]) which is used to obtain comparison and oscillation results for Liénard's equation. In the present article, another device of the theory of ordinary second order linear differential equations is adapted for application to Liénard's equation, namely, the Riccati equation.

It is well known in the theory of linear ordinary differential equations (see [6, p. 332]) that, if $x(t)$ is a solution of the linear equation

$$\frac{d^2x}{dt^2}+p(t)\frac{dx}{dt}+q(t)x=0, \tag{1.1}$$

then on any interval where $x(t)$ is nonzero the function $u=(dx/dt)x^{-1}$ satisfies a Riccati equation. It is easy to see that, if in addition $p(t)$ is differentiable and nonzero, then

$$u=\frac{-\dfrac{dx}{dt}}{p(t)x(t)} \tag{1.2}$$

also satisfies a Riccati equation, namely

$$\frac{du}{dt}=u^2p-u\frac{p^2+\dfrac{dp}{dt}}{p}+\frac{q}{p}.$$

Now consider the more general equation

$$\frac{d^2x}{dt^2}+p(x,t)\frac{dx}{dt}+q(x,t)x=0. \tag{1.3}$$

Let $P(x,t)$ satisfy

$$\frac{\partial}{\partial x}P(x,t)=p(x,t)$$

and $P(0,t)\equiv 0$ identically. One is led to generalize (perhaps to overgeneralize) the Riccati substitution (1.2) by putting

$$u=-\frac{\dfrac{dx}{dt}}{P(x,t)}. \tag{1.4}$$

In this paper it is seen that this substitution leads to useful results in an important special

* Received by the editors November 1, 1979, and in revised form September 18, 1980.
† Department of Mathematics, University of California, Davis, California 95616.

case, the case in which $p(x, t)$ and $q(x, t)$ are functions of x only, i.e., in the autonomous case, the Liénard equation.

To this end consider the Liénard equation

$$\frac{d^2x}{dt^2}+k(x)\frac{dx}{dt}+h(x)=0, \tag{1.5}$$

where it is assumed that k and h are continuous on $(-\infty, \infty)$ and that $k(x)$ is positive if x is nonzero. Put $K(x)=\int_0^x k(\xi)\,d\xi$, and suppose that $x(t)$ is a nonvanishing solution with nonvanishing derivative of (1.5) on an interval $[a, b)$. Note that $\log|K(x(t))|$ is strictly monotone. Now put

$$u=\frac{-\dfrac{dx}{dt}}{K(x)}, \tag{1.6}$$

and introduce as a new independent variable

$$y=-\log|K(x(t))|. \tag{1.7}$$

A calculation from (1.5), (1.6) and (1.7) yields

$$u\frac{du}{dy}=u^2-u+F(y). \tag{1.8}$$

The function $F(y)$ is equal to $h_1(y)/k_1(y)K_1(y)$, where h_1, k_1, and K_1 are determined by the following equations:

$$\begin{aligned} &h_1(-\log|K(x)|)=h(x),\\ &k_1(-\log|K(x)|)=k(x),\\ &K_1(-\log|K(x)|)=K(x), \quad \text{i.e., } |K_1(y)|=e^{-y}. \end{aligned} \tag{1.9}$$

Equation (1.8) will be called the extended Riccati equation of (1.5) because it arises from (1.5) by means of the substitution (1.4), the same substitution which gives the usual Riccati equation in the linear case. In this paper the study of (1.8) yields information concerning solutions of (1.5).

In § 2, we consider questions of existence, uniqueness and continuation of solutions for (1.5) and (1.8). In § 3, some results on the asymptotic behavior of solutions of (1.8) are obtained. In § 4, the results of the preceding sections are applied to obtain results concerning oscillation and asymptotic behavior of solutions of (1.5). (To be more precise we should speak of asymptotic behavior of the *inverse* functions of *monotone* solutions of (1.5).) The concept of principal solution, well known in the linear case (see [6, p. 350]), is generalized in § 4, and it is shown that, in the nonoscillatory case, (1.5) has a principal solution. We conclude with an example to illustrate that in the case of linear equations with constant coefficients, the new concept of principal solution coincides with the usual concept.

2. Existence, uniqueness and continuation of solutions for the extended Riccati equation. In this section we see that, under certain conditions, solutions of initial value problems for both (1.5) and (1.8) exist and are unique and can be continued to infinity to the right. First we consider (1.5).

PROPOSITION 2.1. *Let $h(x)$ and $k(x)$ be continuous on $(-\infty, \infty)$, and let $k(x)$ and $xh(x)$ be positive for x nonzero. Then, for any real numbers a, x_0, v_0, there exists a unique*

solution $x(t)$ of (1.5) *satisfying $x(a)=x_0$ and $x'(a)=v_0$ defined on a maximal interval of existence I such that $I \supset [a, \infty)$.*

Proof. First note that, according to the standard theorems on existence and extension of solutions of differential equations [6, pp. 10–13], it follows that solutions of (1.5) can be continued to the boundary of the region D in (x, x', t)-space (i.e., R^3) in which the second and third terms of (1.5) are continuous.

Put

$$W(t)=\frac{1}{2}x'(t)^2+\int_0^{x(t)} h(\xi)\,d\xi. \tag{2.1}$$

For $x(t)$ an arbitrary solution of (1.5), we have

$$\frac{dW}{dt}=x'(x''+h(x))=-x'^2k(x). \tag{2.2}$$

Since $W(t)$ is decreasing, and since the integral is nonnegative, we see that $x'(t)$ is bounded; in fact,

$$x'(t)^2 \leqq v_0^2+2\int_0^{x_0} h(\xi)\,d\xi \tag{2.3}$$

for all $t \geqq a$ in the interval of existence.

Suppose that, contrary to the assertion of the theorem, the maximal interval of existence is bounded to the right. Since $x'(t)$ is bounded, $x(t)$ must be unbounded; otherwise (x, x', t) could not approach the boundary of the region D. But it follows easily from the law of the mean that in a finite interval $x(t)$ cannot be unbounded unless $x'(t)$ is also unbounded, contrary to (2.3). This contradiction shows that the maximal interval of existence is unbounded to the right.

The uniqueness of $x(t)$ is established in [1, Thm. 4]. This concludes the proof.

PROPOSITION 2.2. *In addition to the hypotheses of Proposition* 2.1, *suppose $x_0>0$ and $v_0<0$. Then there exists b $(a<b\leqq\infty)$ such that $x(t)$ satisfying* (1.5) *and $x(a)=x_0$, $x'(a)=v_0$ is monotone decreasing on $[a, b)$ and $\lim_{t\to b} x(t)=0$.*

Proof. First we observe that $x(t)$ cannot have a local minimum at a point t_1 where $x(t_1)>0$. In fact, if such were the case, we would have $x'(t_1)=0$ and hence from (1.5), $x''(t_1)=-h(x(t_1))<0$, which is not possible at a local minimum.

If $x(t)$ has a zero greater than a, then we are finished because $x(t)$ must be strictly monotone decreasing up to the first zero.

Now suppose that there is no zero of $x(t)$ greater than a. We need to show $\lim_{t\to\infty} x(t)=0$. Since $x(t)$ is strictly monotone decreasing, $\lim_{t\to\infty} x(t)$ exists; suppose it is positive, i.e., $\lim_{t\to\infty} x(t)=x_\infty>0$. As above, put $W(t)=\frac{1}{2}x'^2+\int_0^x h(\xi)\,d\xi$. Since $W(t)$ is decreasing and positive, $W(t)$ tends to a limit W_∞ as $t\to\infty$. Hence, since $x'(t)$ is negative,

$$x'(t)\to -\left(2W_\infty-2\int_0^{x_\infty} h(\xi)\,d\xi\right)^{1/2}=v_\infty \quad \text{as } t\to\infty.$$

Clearly v_∞ must equal zero. From (1.5)

$$\ddot{x}(t)\to v_\infty k(x_\infty)+h(x_\infty)=h(x_\infty)$$

as $t\to\infty$. But the only possible limit for $\ddot{x}(t)$ is zero. Hence $h(x_\infty)=0$, which implies $x_\infty=0$, since $xh(x)$ is positive for $x\neq 0$.

PROPOSITION 2.3. *In addition to the hypotheses of Proposition* 2.1, *assume that* $x_0 > 0$ *and* $v_0 > 0$. *Then there exists* $c < a$ *such that* $(c, a]$ *is in the interval of existence of* $x(t)$, *the unique solution of* (1.5) *subject to* $x(a) = x_0$ *and* $x'(a) = v_0$, *and such that* $\lim_{x \to c^+} x(t) = 0$.

Proof. Let $(a, d]$ be an interval which is maximal in the sense that the solution of (1.5) subject to the initial conditions satisfies $x(t) > 0$ and $x'(t) > 0$ on $(a, d]$, but this solution cannot be continued to the left of a without violating these inequalities. Then $x''(t) < 0$ on $(a, d]$, which implies $d > -\infty$. Moreover, $w = x'(t)$ is a decreasing function which satisfies $w'(t) \geqq -Aw(t) - B$, where $A = \max_{0 \leqq x \leqq x_0} k(x)$ and $B = \max_{0 \leqq x \leqq x_0} h(x)$. It follows that $\lim_{t \to d^-} x'(t) = \lim_{t \to d^-} w(t) < \infty$. Thus both $x(t)$ and $x'(t)$ have finite limits as $t \to d^-$, so that $x(t)$ is continuable to the left at $t = d$. Since $\lim_{t \to d^-} x'(t) > 0$, the maximality of $(d, a]$ implies $x(d) = 0$.

PROPOSITION 2.4. *Let* $h(x)$ *and* $k(x)$ *be continuous on* $(-\infty, \infty)$; *let* $k(x)$ *and* $xh(x)$ *be positive for* x *nonzero. Let* y_0 *be in the range of* $-\log |K(x)|$ *and let* u_0 *be arbitrary. Then there exists a unique solution of* (1.8) *on* $[y_0, \infty)$ *satisfying* $u(y_0) = u_0$.

Proof. There is a unique positive number x_0 such that $|K(x_0)| = e^{-y_0}$. Consider the initial value problem for (1.5),

$$x(0) = x_0, \qquad x'(0) = -u_0 K(x_0). \tag{2.4}$$

If $u_0 K(x_0) \geqq 0$, then by Proposition 2.2 there exists b $(0 < b \leqq \infty)$ such that $x(t)$ is monotone decreasing on $[0, b)$ and $\lim_{t \to b^-} x(t) = 0$.

On the other hand, if $u_0 K(x_0) \leqq 0$, then by Proposition 2.3 there exists c $(-\infty \leqq c < 0)$ such that $x(t)$ is monotone increasing on $(c, 0)$ and $\lim_{t \to c^+} x(t) = 0$.

In either case ($u_0 K(x_0) \geqq 0$ or $u_0 K(x_0) \leqq 0$) the substitutions (1.6) and (1.7) yield a solution of (1.8) on $[y_0, \infty)$ satisfying $u(y_0) = u_0$.

To show that the solution is unique we observe that a solution to (1.8) satisfying $u(y_0) = u_0$ determines a solution of (1.5) by means of (1.6), i.e.,

$$\frac{dx}{dt} = -K(x) u(-\log |K(x)|), \tag{2.5}$$

together with the condition $x(0) = x_0$ where x_0 satisfies $|K(x_0)| = e^{-y_0}$. Moreover (2.5) implies $x'(0) = -u_0 K(x_0)$. By Proposition 2.1, this initial value problem for (1.5) has a unique solution, and therefore (2.5) determines the function $u(y)$ uniquely. This concludes the proof.

PROPOSITION 2.5. *Let* $F(y)$ *be continuous and positive on* $[a, \infty)$, *and let* u_0 *be positive. Then* (1.8) *subject to the initial condition* $u(a) = u_0$ *has a unique positive solution on* $[a, \infty)$. *If two positive solutions* $u_1(y)$ *and* $u_2(y)$ *on* $[a, \infty)$ *satisfy* $u_1(y_0) = u_2(y_0)$ *for some* y_0 *in* $[a, \infty)$, *then* $u_1(y) = u_2(y)$ *for all* y *in* $[a, \infty)$.

Proof. Let y_0 be greater than a. Let $m = \inf \{F(y) | a \leqq y \leqq y_0\}$ and $M = \sup \{F(y) | a \leqq y \leqq y_0\}$. Now, for y in $[a, y_0]$ and u positive, estimate the right-hand side of (1.8) above and below as follows:

$$-\frac{1}{2}\frac{u^2}{m} + \frac{1}{2}m = -\frac{u}{2}\left(\frac{m}{u} + \frac{u}{m}\right) + m \leqq -u + m \leqq u^2 - u + F(y) \leqq u^2 + M.$$

Now consider the auxiliary initial value problems

$$u\frac{du}{dy} = \frac{(-u^2 m^{-1} + m)}{2}, \qquad u(a) = u_0,$$

$$u\frac{du}{dy} = u^2 + M, \qquad u(a) = u_0.$$

These have solutions, respectively,

$$u_1(y) = ((u_0^2 - m^2)\, e^{-(y-a)/m} + m^2)^{1/2} \tag{2.6}$$

and

$$u_2(y) = ((u_0^2 + M)\, e^{2(y-a)} - M)^{1/2}. \tag{2.7}$$

From standard results on differential inequalities [6, p. 26], any solution $u(y)$ of (1.8) satisfying $u(a) = u_0$ must satisfy

$$u_1(y) \leqq u(y) \leqq u_2(y) \tag{2.8}$$

for all y in $[a, y_0]$. From standard results on continuation of solutions [6, p. 12], either there is a solution of the initial value problem which is unbounded somewhere in $[a, y_0]$ or a solution which is not bounded away from 0, or there is a positive solution on the entire interval $[a, y_0]$. From (2.8), the first two possibilities cannot occur because $\inf_{a \leqq y \leqq y_0} u_1(y) > 0$ and $\sup_{a \leqq y \leqq y_0} u_2(y) < \infty$. From the usual Picard theorem [6, p. 8] it follows that this solution is unique. Since $y_0 > a$ is arbitrary, $u(t)$ can be extended to all of $[a, \infty)$. The last assertion of the theorem also follows from the usual Picard theorem.

We have proved something stronger that the theorem requires; namely, (2.6) and (2.7) give specific upper and lower estimates for the solution.

3. Asymptotic behavior of solutions of the extended Riccati equation. In this section, we study the asymptotic behavior of positive solutions of (1.8) with the purpose of later determining the asymptotic behavior as $t \to \infty$ of solutions of (1.5). More specifically, we wish to determine the asymptotic behavior as $y \to \infty$ of solutions of (1.8) if the asymptotic behavior of $F(y)$ as $y \to \infty$ is known.

In case $C_0 < \frac{1}{4}$, the equation

$$x^2 - x + C_0 = 0 \tag{3.1}$$

has two real roots. We denote the smaller root

$$B_0 = \frac{1}{2} - \frac{1}{2}(1 - 4C_0)^{1/2}. \tag{3.2}$$

PROPOSITION 3.1. *Let $0 \leqq C_0 < \frac{1}{4}$, and let $\varepsilon > 0$ be less than $\frac{1}{4} - C_0$. Let $u(y)$ be a solution of* (1.8) *on some interval $[a, b)$ such that*

$$0 < u(a) < \frac{1}{2} + \frac{1}{2}(1 - 4(C_0 + \varepsilon))^{1/2}. \tag{3.3}$$

Moreover, let the function F be positive and continuous and let

$$|F(y) - C_0| < \varepsilon \tag{3.4}$$

for all $y \geqq a$. Then $u(y)$ can be continued as a positive solution of (1.8) *to the interval $[a, \infty)$, and*

$$\begin{aligned} \frac{1}{2} - \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2} &\leqq \liminf_{y \to \infty} u(y) \\ &\leqq \limsup_{y \to \infty} u(y) \leqq \frac{1}{2} - \frac{1}{2}(1 - 4(C_0 + \varepsilon))^{1/2}. \end{aligned} \tag{3.5}$$

Proof. At the risk of some confusion we use $u(y)$ to denote the original solution and also its continuation.

That $u(y)$ can be continued as a positive solution of (1.8) follows directly from Proposition 2.5.

The quadratic $u^2-u+C_0+\varepsilon$ is negative on the interval $I_1: |u-\frac{1}{2}|< \frac{1}{2}(1-4(C_0+\varepsilon))^{1/2}$; moreover, $u^2-u+C_0-\varepsilon$ is positive on the two infinite intervals

$$I_2=\left\{u\middle| u<\frac{1}{2}-\frac{1}{2}(1-4(C_0-\varepsilon))^{1/2}\right\}$$

and

$$I_3=\left\{u\middle| u>\frac{1}{2}+\frac{1}{2}(1-4(C_0-\varepsilon))^{1/2}\right\}.$$

Moreover, let I_4 be the interval $\frac{1}{2}-\frac{1}{2}(1-4(C_0-\varepsilon))^{1/2}\leqq u\leqq\frac{1}{2}-\frac{1}{2}(1-4(C_0+\varepsilon))^{1/2}$. We shall show that there exists $y_1\geqq a$ such that $u(y)$ is in I_4 for all $y\geqq y_1$. In fact, it is sufficient to show that there exists $y_1\geqq a$ such that the solution can be continued to the interval $[a, y_1]$ and $u(y_1)$ is in I_4. Indeed, if this is so and the continuation of $u(y)$ does not remain in I_4 for all $y>y_1$, let $y_2=\inf\{y>y_1: u(y)\notin I_4\}$. Then $u(y_2)$ must be either the left or the right endpoint of I_4, i.e., respectively either

$$u(y_2)=\frac{1}{2}-\frac{1}{2}(1-4(C_0-\varepsilon))^{1/2} \tag{3.6}$$

or

$$u(y_2)=\frac{1}{2}-\frac{1}{2}(1-4(C_0+\varepsilon))^{1/2}. \tag{3.7}$$

If (3.6), then from (1.8), (3.5), and the positivity of u,

$$\begin{aligned}u'(y_2)&=\frac{1}{u(y_2)}(u(y_2)^2-u(y_2)+F(y_2))\\&=\frac{1}{u(y_2)}(-C_0+\varepsilon+F(y_2))>0,\end{aligned}$$

which is impossible if $u(y_2)$ is the left endpoint of I_4, since by definition y_2 is the point at which $u(y)$ leaves the interval I_4. A similar contradiction follows from assuming (3.7).

It remains to show that there exists $y_1\geqq a$ such that $u(y_1)\in I_4$.

For any $y\geqq a$ such that $u(y)\in I_1$,

$$u'(y)=\frac{1}{u}(u^2-u+F(y))\leqq\frac{1}{u}(u^2-u+C_0+\varepsilon)<0. \tag{3.8}$$

Thus if $u(y)\in I_1(y\geqq a)$, it is impossible for $u(y)$ to remain in I_1 for all sufficiently large y; if such were the case, then $u(y)$ would decrease monotonically to a limit $u_\infty\geqq\frac{1}{2}-\frac{1}{2}(1-4(C_0+\varepsilon))^{1/2}$; and, since $0\leqq C_0<\frac{1}{4}$ implies that this bound is positive,

$$\lim_{y\to\infty}u'(y)=\frac{1}{u_\infty}(u_\infty^2-u_\infty+C_0)\leqq-\frac{\varepsilon}{u_\infty}<0,$$

which is impossible. Hence $u(y)$ must leave I_1 at some point y_1; by (3.8) $u(y_1)$ cannot be the right endpoint of I_1, and the only other alternative is that $u(y_1)$ is the left endpoint of I_1, specifically

$$u(y_1)=\frac{1}{2}-\frac{1}{2}(1-4(C_0+\varepsilon))^{1/2},$$

which is the right endpoint of I_4.

A similar argument shows that for $y \geqq a$, if $u(y)$ enters I_2 (which cannot occur if $C_0 < \varepsilon$, i.e., if the left endpoint of I_4 is nonpositive), then there exists $y_1 \geqq a$, such that

$$u(y_1) = \frac{1}{2} - \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2},$$

which is the left endpoint of I_4. (This argument uses the positivity of the left endpoint of I_4 in the same manner that the positivity of the right endpoint was used in the previous argument.)

Since $u(a) \in I_1 \cup I_2 \cup I_4$, it follows from the above that there exists $y_1 \geqq a$ such that $u(y_1) \in I_4$.

As remarked above, it follows that $u(y) \in I_4$ for all $y \geqq y_1$. This is equivalent to one of the assertions of the proposition.

This concludes the proof.

PROPOSITION 3.2. *Let* $0 \leqq C_0 < \frac{1}{4}$, *and let* $u(y)$ *be a positive solution of* (1.8) *on* $[a, \infty)$ *such that*

$$\liminf_{u \to \infty} u(y) < 1 - B_0.$$

Let $F: [a, \infty) \to R$ *be continuous and positive, and let*

$$\lim_{y \to \infty} F(y) = C_0.$$

Then $\lim_{y \to \infty} u(y)$ *exists and is equal to* B_0.

Proof. Let $\varepsilon > 0$ satisfy $\varepsilon < \frac{1}{4} - C_0$ and $\frac{1}{2} + \frac{1}{2}(1 - 4(C_0 + \varepsilon))^{1/2} > \liminf_{y \to \infty} u(y)$. There exists y_0 such that

$$|F(y) - C_0| < \varepsilon \quad \text{for all } y \geqq y_0$$

and

$$0 < u(y_0) < \frac{1}{2} + \frac{1}{2}(1 - 4(C_0 + \varepsilon))^{1/2}.$$

The assertion follows by applying Proposition 3.1 and observing that ε may be taken arbitarily small.

PROPOSITION 3.3. *Let* $0 \leqq C_0 < \frac{1}{4}$, *and let* $\varepsilon > 0$ *be less than* $\frac{1}{4} - C_0$. *Let* $u(y)$ *be a positive solution of* (1.8) *on* $[a, \infty)$ *such that*

$$u(a) > \frac{1}{2} + \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2}. \tag{3.9}$$

Moreover, let $F: [a, \infty) \to R$ *be continuous and positive, and let*

$$F(y) > C_0 - \varepsilon \quad \text{for all } y \geqq a. \tag{3.10}$$

Then $u(y) \to \infty$ *monotonically as* $y \to \infty$; *moreover, for some positive constant* P *and some* $y_0 \geqq a$, $u(y) > Pe^y$ *for all* $y > y_0$.

Proof. The quadratic $u^2 - u + C_0 - \varepsilon$ is positive in the interval

$$I_3 = \left\{ u : u > \frac{1}{2} + \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2} \right\}.$$

For all $y \geqq a$

$$u'(y) = \frac{1}{u}(u^2 - u + F(y)) > \frac{1}{u}(u^2 - u + C_0 - \varepsilon).$$

Since $u(a)$ belongs to I_3 it follows that $u(y)$ is increasing and belongs to I_3 for all $y \geqq a$. Moreover, $u(y)$ tends monotonically to ∞ as $y \to \infty$, because if $u(y)$ had a finite limit then $u'(y)$ would have a positive lower bound, which would be impossible. Consequently there exists $y_1 \geqq a$ such that $u(y_1) > 1$. Hence for all $y > y_1$

$$u'(y) = u(y) - 1 + \frac{F(y)}{u} > u(y) - 1. \tag{3.11}$$

Integrating (3.11) one obtains for $y > y_1$

$$u(y) > (u(y_1) - 1)\, e^{y - y_1} + 1; \tag{3.12}$$

since $u(y_1) - 1$ is positive, the order-of-magnitude assertion of the propositions follows, and the proof is complete.

Next follows a proposition related to Proposition 3.3 more or less as Proposition 3.2 is related to Proposition 3.1.

PROPOSITION 3.4. *Let $F: [a, \infty) \to R$ be continuous and positive. Let $0 \leqq C_0 < \frac{1}{4}$ and let $\liminf_{y \to \infty} F(y) \geqq C_0$. Let $u(y)$ be a positive solution of (1.8) such that $\limsup_{y \to \infty} u(y) > 1 - B_0$. Then $u(y) \to \infty$ as $y \to \infty$; more specifically, $u(y) > P e^y$ for some positive constant P for all y sufficiently large.*

Proof. For any $\varepsilon > 0$ satisfying both $\varepsilon < \frac{1}{4} - C_0$ and

$$\frac{1}{2} + \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2} < \limsup_{y \to \infty} u(y),$$

there exists $y_0 \geqq a$ such that

$$u(y_0) > \frac{1}{2} + \frac{1}{2}(1 - 4(C_0 - \varepsilon))^{1/2}$$

and

$$F(y) > C_0 - \varepsilon \quad \text{for all } y \geqq y_0.$$

The result now follows by an application of Proposition 3.3.

Propositions 3.2 and 3.4 leave open the analysis of the case $\lim_{y \to \infty} F(y) = C_0$ and $\lim_{y \to \infty} u(y) = 1 - B_0$. The following discussion fills in this gap.

THEOREM 3.1. *Let $F: [a, \infty) \to R$ be continuous and positive. Let $0 \leqq C_0 < \frac{1}{4}$ and let $\lim_{y \to \infty} F(y) = C_0$. Let a be a real number. Let there exist at least one bounded solution u_b of (1.8) on $[a, \infty)$. Then there exists a unique positive solution $U(y)$ of (1.8) on $[a, \infty)$ such that $\lim_{y \to \infty} U(y) = 1 - B_0$. Moreover, if $u(y)$ is any positive solution of (1.8) on $[a, \infty)$, then exactly one of the following conditions holds:*

(a) $\lim_{y \to \infty} u(y) = B_0$;

(b) $\lim_{y \to \infty} u(y) = 1 - B_0$;

(c) $\lim_{y \to \infty} u(y) = \infty$, *and there exists $P > 0$ such that $u(y) > P e^y$ for all large y.*

The three conditions (a), (b), (c) *hold accordingly as $u(a) < U(a)$, $u(a) = U(a)$, or $u(a) > U(a)$, respectively.*

Proof. That any positive solution $u(y)$ on $[a, \infty)$ satisfies exactly one of conditions (a), (b), (c), follows immediately from Propositions 3.2 and 3.4.

Let y_0 be in $[a, \infty)$ and let u_0 be positive. Let $u(y; y_0, u_0)$ denote the unique solution of (1.8) on $[y_0, \infty)$ satisfying $u(y_0) = u_0$. The existence of such a unique solution is assured by Proposition 2.5.

If the given bounded solution u_b of (1.8) satisfies (b), then we take this solution for U. If not, then this solution must satisfy $\liminf_{y \to \infty} u_b(y) < 1 - B_0$ or $\limsup_{y \to \infty} u_b(y) >$

$1-B_0$. The latter case is impossible, since otherwise Proposition 3.4 would imply that the solution is unbounded. Hence Proposition 3.2 would imply that the given bounded solution satisfies $\lim_{y\to\infty} u_b(y) = B_0$.

There exist unbounded solutions of (1.8). In fact, (1.8) and the positivity of F and u imply $du/dy > u-1$; integrating, we obtain

$$u(y) > (u(a)-1)\, e^{y-a} + 1$$

for all $y > a$. If $u(a) > 1$, then $u(y)$ is unbounded. Moreover, the conclusion of Proposition 3.4 applies.

Now let

$$R = \left\{ u_0 > 0 \,\middle|\, \limsup_{y\to\infty} u(y; a, u_0) > 1 - B_0 \right\},$$

$$L = \left\{ u_0 > 0 \,\middle|\, \liminf_{y\to\infty} u(y; a, u_0) < 1 - B_0 \right\}.$$

From the foregoing, R and L are disjoint and nonempty. They are intervals, because otherwise the graphs of distinct solutions would have to cross, contrary to the last assertion of Proposition 2.5. An arbitrary element of R is greater than any element of L. Let $u_* = \sup L$, and let $u^* = \inf R$.

We shall show that $u_* = u^*$ and that $U(y) = u(y; a, u^*)$ is the unique solution satisfying condition (b).

First we show that no more than one solution can satisfy (b). Suppose there are two such solutions u_1 and u_2. Then there exists y_1 such that $u_1(y) > \frac{1}{2}$ and $u_2(y) > \frac{1}{2}$ for all $y \geqq y_1$. Let $V(y) = |u_1(y)^2 - u_2(y)^2|$ for $y \geqq a$. At points where $V(y) \neq 0$, V is differentiable and, for $y \geqq y_1$, (1.8) implies

$$\begin{aligned} V'(y) &= 2(u_1(y)^2 - u_1(y) - u_2(y)^2 + u_2(y)) \operatorname{sgn}(u_1(y) - u_2(y)) \\ &= 2\left[\left(u_1(y) - \frac{1}{2}\right)^2 - \left(u_2(y) - \frac{1}{2}\right)^2\right] \operatorname{sgn}(u_1(y) - u_2(y)) > 0. \end{aligned}$$

On the other hand, condition (b) implies $V(y) \to 0$ as $y \to \infty$. Since V is also nonnegative and nondecreasing, $V(y)$ must vanish identically on $[y_1, \infty)$. By Proposition 2.2, it follows that $u_1(y) = u_2(y)$ for all y on $[a, \infty)$.

It now follows that $u_* = u^*$. Indeed, suppose $u_* < u^*$; then there exist numbers ω_1 and ω_2 such that $u_* < \omega_1 < \omega_2 < u^*$. By Proposition 2.5, $u(y; a, \omega_1)$ and $u(y; a, \omega_2)$ must be distinct. But now, by the definitions of u_*, u^*, L and R,

$$\lim_{y\to\infty} u(y; a, \omega_1) = \lim_{y\to\infty} u(y; a, \omega_2) = 1 - B_0.$$

We have just shown that this implies $u(y; a\, \omega_1) = u(y; a, \omega_2)$ for all $y \geqq a$, contrary to the distinctness of these two functions.

Now put $U(y) = u(y; a, u^*) = u(y; a, u_*)$. We have seen that $U(y)$ must satisfy exactly one of the conditions (a), (b), (c).

In particular, suppose (a) is satisfied. Let $\varepsilon = \frac{1}{8} - (C_0/2)$. Choose $y_2 \geqq a$ such that

(i) $$U(y_2) < \tfrac{1}{2}$$

and

(ii) $$|F(y) - C_0| < \varepsilon$$

for all $y \geqq y_2$. The usual theorem on continuous dependence on initial values [6, p. 94]

implies $u(y_2; a, u_0)$ is continuous as a function of u_0. Hence there exists a number $u_0 > u^*$ such that $u(y_2; a, u_0) < \frac{1}{2}$. On the one hand Proposition 3.1 now implies $\limsup_{y\to\infty} u(y; a, u_0) < \frac{1}{2}$; on the other hand $u_0 \in R$, and hence Proposition 3.4 implies $\lim_{y\to\infty} u(y; a, u_0) = \infty$. This contradiction shows that, in fact, $U(y)$ cannot satisfy condition (a).

Now suppose (c) is satisfied. As above, let $\varepsilon = \frac{1}{8} - (C_0/2)$. Choose $y_3 \geqq a$ such that

(i) $$U(y_3) > 1,$$

(ii) $$F(y) > C_0 - \varepsilon,$$

for all $y \geqq y_3$. Arguing as above, we conclude that there exists a number $u_1 < u_*$ such that $u(y_3; a, u_1) > 1$. On the one hand, Proposition 3.3 implies $\lim_{y\to\infty} u(y; a, u_1) = \infty$; on the other hand, $u_1 \in L$, and hence Proposition 3.2 implies $\lim_{y\to\infty} u(y; a, u_1) < \frac{1}{2}$. This contradiction shows that $U(y)$ cannot satisfy condition (c).

Since $U(y)$ does not satisfy (a) or (c), it must satisfy (b). Since $U(a) = u^* = u_*$, it follows that if $u(y)$ is another solution of (1.8) on $[a, \infty)$, then (a), (b), (c) hold accordingly as $u(a) < U(a)$, $u(a) = U(a)$, or $u(a) > U(a)$, respectively. This concludes the proof.

This section concludes with an examination of the case $C_0 > \frac{1}{4}$.

PROPOSITION 3.5. *Let $C_0 > \frac{1}{4}$, and let $\varepsilon > 0$ be less than $C_0 - \frac{1}{4}$. Let $u(y)$ be a positive solution of* (1.8) *on* $[a, \infty)$. *Moreover, let $F(y)$ be positive and continuous, and let*

(3.13) $$F(y) > C_0 - \varepsilon \quad \text{for all } y \geqq a.$$

Then $u(y) \to \infty$ monotonically as $y \to \infty$; moreover, for some positive constant P and some $y_0 \geqq a$, $u(y) > Pe^y$ for all $y > y_0$.

Proof. The right-hand side of (1.8) must be positive under hypothesis (3.14). Now the order of magnitude assertion ($u(y) > Pe^y$) is shown exactly as in the proof of Proposition 3.3.

PROPOSITION 3.6. *Let $C_0 > \frac{1}{4}$ and let $\liminf_{y\to\infty} F(y) \geqq C_0$. Let $u(y)$ be a positive solution of* (1.8) *on an interval $[a', \infty)$, and let F be positive and continuous on this interval. Then $u(y) \to \infty$ as $y \to \infty$; more specifically $u(y) > Pe^y$ for some positive constant and for all y sufficiently large.*

Proof. Let $\varepsilon > 0$ satisfy $\varepsilon < C_0 - \frac{1}{4}$. There exists $a \geqq a'$ such that the hypotheses of Proposition 3.5 are satisfied. This proposition yields the desired assertion.

The critical case, $C_0 = \frac{1}{4}$, will not be discussed in this paper.

4. Oscillation, nonoscillation, and asymptotic behavior of solutions of the Liénard equation. Principal solutions. In this section, we see some applications of the results of the previous section. We shall consider (1.5) in the case that all solutions can be continued on the right to a half infinite interval $[a, \infty)$ and that $x \equiv 0$ is a globally asymptotically stable (g.a.s.) solution of (1.5). Conditions that imply this are well known (see [9, p. 67]); for example, the following conditions suffice:

(4.1) $$xh(x) > 0 \quad \text{and} \quad k(x) > 0 \quad \text{for } x \neq 0$$

and

(4.2) $$\left| \int_0^x h(\xi)\, d\xi \right| \to \infty \quad \text{as } |x| \to \infty.$$

Other conditions for global asymptotic stability of (1.5) and related equations are studied in [2], [4], [5], [7], and [10].

We shall say that a solution of (1.5), not identically zero, is *oscillatory* if it has a sequence of zeros tending to $+\infty$. If this condition does not hold, then the solution is said to be *nonoscillatory*.

By Proposition 2.1, if h and k are continuous and (4.1) holds, then a maximal solution of (1.5) is uniquely determined by its initial values. Moreover, if x and x' vanish simultaneously at a single point, then x vanishes identically. Because of these facts, nonoscillatory solutions in the present case, just as in the more familiar linear case, are characterized by the fact that they have finitely many zeros.

Oscillation problems for equations related to (1.5) have been studied by other authors, e.g., [11]. Burton and Townsend [3] give necessary and sufficient conditions for oscillation of the *forced* Liénard equation, but their restrictive assumptions on the forcing term exclude the unforced case which is presently under consideration. The author [1] has previously treated oscillation problems in the unforced case.

THEOREM 4.1. *Let $x \equiv 0$ be a* g.a.s. *solution of* (1.5) *on $0 \leqq t < \infty$. Let condition* (4.1) *hold, and let h and k be continuous. Let $K(x) = \int_0^x k(\xi)\, d\xi$, and let*

$$\lim_{\alpha \to 0} \frac{h(x)}{k(x)K(x)} = C_0. \tag{4.3}$$

If $C_0 > \frac{1}{4}$, then all solutions of (1.5) *are oscillatory. If $C_0 < \frac{1}{4}$, then all solutions of* (1.5) *are nonoscillatory.*

Proof. Suppose $C_0 > \frac{1}{4}$. Let $x(t)$ be any solution of (1.5). Suppose $x(t)$ is nonoscillatory. Then $x(t)$ must tend monotonically to zero on some interval $[a, \infty)$. In fact, making use of (4.1), we see that the interval can be chosen such that $x'(t)$ and $x(t)$, are nonzero. We have seen in § 1 that under these conditions transformations (1.6) and (1.7) yield a solution of (1.8) on the interval $-\log |K(x(a))| \leqq y < \infty$. Moreover, condition (4.3) becomes $F(y) \to C_0$ as $y \to \infty$. If $x(t)$ is monotone increasing, then we replace (1.5) by the equation

$$\frac{d^2x}{dt^2} + k(-x)\frac{dx}{dt} - h(-x) = 0, \tag{4.4}$$

and apply (1.6) and (1.7) to $-x(t)$ accordingly to obtain a solution of (1.8); again $F(y) \to C_0$ as $y \to \infty$. In either case, Proposition 3.6 asserts that there exists y_0 such that $u(y) > Pe^y$ for some $P > 0$ for all $y > y_0$. In the case that $x(t)$ is monotone decreasing, this implies that there exists t_0 such that

$$\frac{-\dfrac{dx}{dt}}{K(x)} > P\frac{1}{K(x)} \tag{4.5}$$

for all $t > t_0$. This is not possible since $x(t)$ is assumed to tend to zero on an *infinite* interval. The case in which $x(t)$ is monotone increasing is handled similarly. Thus the assumption that a nonoscillatory solution exists yields a contradiction.

Now let $C_0 < \frac{1}{4}$. Suppose that $x(t)$ is an oscillatory solution of (1.5). Let $\varepsilon > 0$ be less than $\frac{1}{4} - C_0$. Since $x(t) \to 0$ as $t \to \infty$, there exists a such that $x'(a) = 0$, $x(a) > 0$, and

$$\left| \frac{h(x)}{k(x)K(x)} - C_0 \right| < \varepsilon$$

for all $x < x(a)$. Let $b > a$ satisfy $x(b) = 0$ and $x(t) > 0$ for $a \leqq t < b$. As above, (1.6) and (1.7) yield a solution $u(y)$ of (1.8) on the interval $-\log |K(x(a))| \leqq y < \infty$. Since $x'(b)$ is

nonzero, $u(y)\to\infty$ as $y\to\infty$. On the other hand, Proposition 3.1 asserts

$$\limsup_{y\to\infty} u(y)\leqq\frac{1}{2}-\frac{1}{2}(1-4(C_0+\varepsilon))^{1/2}.$$

Thus the assumption that an oscillatory solution exists yields a contradiction. This concludes the proof.

This theorem may be compared to the oscillation and nonoscillation results in [1]. In particular, the oscillation condition given here is implied by [1, Thm. 5], which gives information even in the critical case $C_0=\frac{1}{4}$. However, the foregoing theorem still seems of interest because the methods used here are quite different from those of [1], which are based on a generalization of the Picone identity.

The remainder of this section is concerned with the nonoscillatory case $C_0<\frac{1}{4}$. We continue to assume that $x\equiv 0$ is a g.a.s. solution of (1.5). We shall give an analogue for (1.5) of the theory of *principal solutions* (see [6, pp. 350–361]) of second order *linear* equations.

The following lemma is needed.

LEMMA 4.1. *Let $k(x)$ be continuous on $[0, x_0]$ and positive on $(0, x_0]$, and let $K(x)=\int_0^x k(\xi)\,d\xi$. Then*

$$\int_0^{x_0}\frac{dx}{K(x)}=\infty. \tag{4.6}$$

Proof. By l'Hôpital's rule,

$$\lim_{x\to 0^+}\frac{1/x}{1/K(x)}=\lim_{x\to 0^+}\frac{K(x)}{x}=\lim_{x\to 0^+}k(x)=k(0).$$

Hence (4.6) holds by the limit form of the comparison test for improper integrals.

Consider the initial value problem for (1.5),

$$x(0)=x_0>0 \quad\text{and}\quad x'(0)=v_0<0.$$

In the present case $x(t)=x(t; x_0, v_0)$ is strictly monotone decreasing on $[0,\infty)$ or else $x(t)$ vanishes at some point, in which case $x(t)$ is strictly monotone decreasing on an interval $[0, a)$ and $x(a)=0$. In either case let $T(x)=T(x; x_0, v_0)$ denote the inverse of $x(t)$ which is defined on the interval $(0, x(0))$. Let $u(y)=\omega(y; x_0, v_0)$ be the solution of (1.8) which corresponds by means of (1.6) and (1.7) to the solution $x(t; x_0, v_0)$. Let $U(y)$ be the unique solution of (1.8) such that

$$\lim_{y\to\infty} U(y)=1-B_0=\frac{1}{2}+\frac{1}{2}(1-4C_0)^{1/2}; \tag{4.7}$$

the existence and uniqueness of $U(y)$ is asserted by Theorem 3.1. Let

$$U^*(x)=U(-\log|K(x)|).$$

Then (4.7) implies

$$\lim_{x\to 0} U^*(x)=1-B_0.$$

THEOREM 4.2. *Let condition* (4.1) *hold, and let h and k be continuous. Let $K(x)=\int_0^x k(\xi)\,d\xi$. Let* (4.3) *hold with $0<C_0<\frac{1}{4}$. Let $x_0>0$ and $v_0<0$. Let $x\equiv 0$ be a* g.a.s. *solution of* (1.5) *on* $[0,\infty)$. *Let $x(t; x_0, v_0)$, $T(x; x_0, v_0)$, and U^* be as defined in the preceding paragraph.*

(a) *If* $-v_0 > K(x_0)U^*(x_0)$, *then there exists* a, *positive and finite, such that* $x(a; x_0, v_0) = 0$.

(b) *If* $-v_0 < K(x_0)U^*(x_0)$, *then* $x(t; x_0, v_0) \neq 0$ *on* $[0, \infty)$ *and*

$$T(x; x_0, v_0) \sim \frac{1}{B_0} \int_x^{x_0} \frac{d\xi}{K(\xi)}$$

as $x \to 0$. $(B_0 = \frac{1}{2} - \frac{1}{2}(1 - 4C_0)^{1/2})$.

(c) *If* $-v_0 = K(x_0)U^*(x_0)$, *then* $x(t; x_0, v_0) \neq 0$ *on* $[0, \infty)$, *and*

$$T(x; x_0, v_0) \sim \frac{1}{1 - B_0} \int_x^{x_0} \frac{d\xi}{K(\xi)}.$$

Proof. Suppose $-v_0 > K(x_0)U^*(x_0)$. Then, by Theorem 3.1, there exists $P > 0$ such that $\omega(y; x_0, v_0) > Pe^y$ for all sufficiently large y. Taking account of (1.6) and (1.7), we have $-x' > P$ for x in some right-handed neighborhood of zero, i.e., for x in an open interval of the form $0 < x < \varepsilon$ where $\varepsilon > 0$ is suitably chosen.

If there does not exist a, positive and finite, such that $x(a; x_0, v_0) = 0$, then $T(x) \to \infty$ as $x \to 0^+$. Moreover, from the g.a.s. hypothesis, $x'(t) \to 0$ as $t \to \infty$; or, in other words, we would have $x' \to 0$ as $x \to 0^+$. But this is impossible because it has been shown above that $-x' > P > 0$ for x in a right-handed neighborhood of zero. This contradiction shows that assertion (a) of the theorem is correct.

Now suppose $-v_0 < K(x_0)U(x_0)$. By Theorem 3.1, $\omega(y; x_0, v_0) \to B_0$ as $y \to \infty$. Taking account of (1.6) and (1.7), using l'Hôpital's rule and Lemma 4.1, we have

$$\lim_{x \to 0^+} \frac{B_0 T(x; x_0, v_0)}{\int_x^{x_0} K(\xi)^{-1}\, d\xi} = \lim_{x \to 0^+} \frac{B_0\left(1 \Big/ \dfrac{dx}{dt}\right)}{-K(x)^{-1}} = \lim_{y \to \infty} \frac{B_0}{u(y)} = 1. \tag{4.8}$$

The assertion $x(t; x_0, v_0) \neq 0$ on $[0, \infty)$ is now a consequence of

$$T(x; x_0, v_0) \to \infty \quad \text{as } x \to 0,$$

which follows from (4.8) and (4.6). This proves assertion (b) of the theorem. Assertion (c) is proved in a similar manner by using (b) of Theorem 3.1.

One obtains an interpretation of Theorem 4.2 by introducing an extension of the concept of *principal solution* (see [6, p. 350]). In the theory of the linear second order nonoscillatory equation, a principal solution is a solution, unique apart from a scalar multiple, which is, roughly speaking, smaller at infinity than any nonprincipal solution. The extension offered here coincides with the usual concept in the case of second order linear equations with constant coefficients.

The new principal solutions are small at infinity in a different sense. As is reasonable for solutions of an autonomous equation, translation of the independent variable takes a principal solution into another principal solution.

THEOREM 4.3. *Let condition* (4.1) *hold, and let* h *and* k *be continuous. Let* $K(x) = \int_0^k k(\xi)\, d\xi$. *Let* $x \equiv 0$ *be a* g.a.s. *solution of* (1.5). *Let* (4.3) *hold with* $0 < C_0 < \frac{1}{4}$. *There exists a positive strictly monotone solution* $x^*(t)$ *of* (1.5) *defined on a right-half line, which we shall call a principal solution, such that*

$$\lim_{t \to \infty} \frac{-x^{*\prime}(t)}{K(x^*(t))} = 1 - B_0 = \frac{1}{2} + \frac{1}{2}(1 - 4C_0)^{1/2}. \tag{4.9}$$

The solution x^* *is uniquely defined by* (4.9) *apart from a translation of the independent variable; i.e., any other positive solution* $x(t)$ *of* (1.5) *satisfying* (4.9) *must satisfy, for all*

sufficiently large t,

$$x(t) = x^*(t + t_1)$$

for some real number t_1. *Let* $T^*(x)$ *denote the inverse of* $x^*(t)$. *Let* $x(t)$ *be any other positive monotone solution of* (1.5) *on a right-half line, and let* $T(x)$ *be the inverse of* $x(t)$. (*From the* g.a.s. *assumption it follows that* T^* *and* T *are both defined in a right-handed neighborhood of* 0.) *Exactly one of the following two alternatives must hold*:

(i) $$\lim_{x \to 0^+} \frac{T(x)}{T^*(x)} = \frac{1 - B_0}{B_0}$$

or

(ii) $$\lim_{x \to 0^+} \frac{T(x)}{T^*(x)} = 1.$$

In case (i), $x(t)$ *is a nonprincipal solution*; *in case* (ii) , $x(t)$ *is a principal solution.*

Proof. By Theorem 3.1, there exists a unique positive solution $U(y)$ of (1.8) on some interval $[y_1, \infty)$ such that

$$\lim_{y \to \infty} U(y) = 1 - B_0.$$

Let $x^*(t)$ be a solution of

(4.10) $$\frac{dx}{dt} = -K(x)U(-\log |K(x)|)$$

satisfying $x(0) = x_1$, where $K(x_1) = e^{-y_1}$. We must show that $x^*(t)$ is well defined and has the asserted properties. Since $k(x)$ is positive, x_1 is uniquely determined. It is clear that x^* must satisfy (4.9). Differentiating both sides of (4.10) and making use of (1.6), (1.7) and (1.8), we see that $x^*(t)$ must be a solution of (1.5). Let $x^{**}(t)$ be another solution of (1.5) satisfying (4.9). By putting $x(t) \equiv x^{**}(t)$ in (1.6) and (1.7), a solution $u(y)$ of (1.8) is obtained which satisfies

$$\lim_{y \to \infty} u(y) = 1 - B_0.$$

From the uniqueness assertion of Theorem 3.1, $u(y) \equiv U(y)$. Therefore $x^{**}(t)$ satisfies (4.10). Since (4.10) is autonomous and the right-hand side is continuously differentiable, solutions of (4.10) are unique apart from a translation of the independent variable. Thus we have shown the uniqueness assertion of the theorem.

If $x(t)$ is any positive monotone solution of (1.5) defined on a right-half line, then either case (b) or case (c) of Theorem 4.2 applies. If (b), then assertion (i) of the theorem holds, and $x(t)$ is nonprincipal; if (c), then (ii) holds and $x(t)$ is principal. This concludes the proof.

The following example shows that the new concept of principal solution coincides with the old in the only case in which it is possible to compare them, namely, the case of constant coefficients.

Example 4.1. Consider the equation

(4.11) $$\frac{d^2x}{dt^2} + 3\frac{dx}{dt} + 2x = 0.$$

We have (see § 1):

$$h(x)=2x, \qquad h_1(y)=\tfrac{2}{3}e^{-y},$$
$$k(x)=3, \qquad k_1(y)=3,$$
$$K(x)=3x, \qquad K_1(y)=e^{-y},$$
$$F(y)=\frac{h_1(y)}{k_1(y)K_1(y)}=\frac{2}{9},$$
$$C_0=\lim_{y\to\infty}F(y)=\frac{2}{9}.$$

Since $0<C_0<\frac{1}{4}$, the preceding theorem applies. The principal solutions are characterized by (4.9), which reads in our special case

$$\lim_{t\to\infty}\frac{-x^{*\prime}(t)}{3x^*(t)}=\frac{1}{2}+\frac{1}{2}(1-4C_0)^{1/2}=\frac{2}{3}. \tag{4.12}$$

The solutions of (4.11) are of the form

$$x=C_1e^{-t}+C_2e^{-2t}.$$

The only functions of this form which satisfy (4.12) are those for which $C_1=0$. According to Theorem 4.3, the principal solutions are of the form

$$C_2e^{-2(t+t_1)};$$

by replacing C_2 by another suitable constant C_2', we see that the principal solutions in the new sense have the form

$$C_2'e^{-2t};$$

these are precisely the principal solutions in the old sense.

Acknowledgment. The author is grateful to the referee for the present simplified version of the proof of Theorem 2.3.

REFERENCES

[1] D. C. BENSON, *Comparison and oscillation theory for Liénard's equation with positive damping*, SIAM J. Appl. Math., 24 (1973), pp. 251–271.

[2] T. A. BURTON, *On the equation* $x''+f(x)h(x')x'+g(x)=e(t)$, Ann. Mat. Pura Appl., 85 (1970), pp. 277–285.

[3] T. A. BURTON AND C. G. TOWNSEND, *On the generalized Liénard equation with forcing function*, J. Differential Equations, 4 (1968), pp. 620–633.

[4] J. R. GRAEF AND P. W. SPIKES, *Asymptotic behavior of solutions of a second order nonlinear differential equation*, J. Differential Equations, 17 (1975), pp. 461–476.

[5] R. GRIMMER, *On nonoscillatory solutions of a nonlinear differential equation*, Proc. Amer. Math. Soc., 34 (1972), pp. 118–120.

[6] PHILIP HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[7] J. W. HEIDEL, *A Liapunov function for a generalized Liénard equation*, J. Math. Anal. Appl., 39 (1972), pp. 192–197.

[8] KURT KREITH, *Oscillation Theory*, Springer-Verlag, Berlin, 1973.

[9] J. P. LASALLE AND S. LEFSCHETZ, *Stability by Liapunov's Direct Method*, Academic Press, New York, 1961.

[10] S.-O. LONDEN, *Some nonoscillation theorems for a second order nonlinear differential equation*, this Journal, 4 (1973), pp. 460–465.

[11] W. R. UTZ, *Periodic solutions of a nonlinear second order differential equation*, SIAM J. Appl. Math., 19 (1970), pp. 56–59.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0010 $01.00/0

HIERARCHIES OF ITERATED AVERAGES FOR SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS WITH A SMALL PARAMETER*

STEPHEN C. PERSEK†

Abstract. A hierarchy of iterated or nonlinear averaging methods is developed for periodic systems by the introduction of the general nth order iterated average. Approximations are then derived with $O(\varepsilon)$ accuracy, uniformly valid on intervals $0 \leqq t < \infty$ or (the lesser case) $0 \leqq t \leqq O(1/\varepsilon^n)$.

1. Introduction. With $\varepsilon > 0$ and w in R^k, consider the initial value problem on $0 \leqq t < \infty$ for the system of ordinary differential equations

$$\frac{dw}{dt} = \varepsilon E(w, t, \varepsilon), \qquad w|_{t=0} = \mathring{w}_0 + \varepsilon \mathring{w}_1(\varepsilon), \tag{1), (2}$$

with E quasiperiodic in t, and with $\varepsilon > 0$ small. Bogoliubov–Mitropolskii [1] obtained an approximating system to (1), (2) by replacing the vector $E(w, t, \varepsilon)$ with its average on $0 \leqq t < \infty$ at $\varepsilon = 0$, denoted $\bar{E}^{(1)}(w)$. This approximation results in solutions accurate to $O(\varepsilon)$ uniformly on $0 \leqq t \leqq O(1/\varepsilon)$, and this accuracy extends uniformly to $0 \leqq t < \infty$ if the approximating system is exponentially asymptotically stable.

Now if $\bar{E}^{(1)}(w) \equiv 0$, then for periodic systems in the form (1), Laričeva [2] replaces E by an iterated average $\varepsilon \bar{E}^{(2)}(w)$, and obtains an approximation to the solution of (1) and (2), accurate to $O(\varepsilon)$ on $0 \leqq t \leqq O(1/\varepsilon^2)$. Persek and Hoppensteadt [3] extend the definition of $\bar{E}^{(2)}(w)$ to aperiodic systems including those in the form (1), show that the Laričeva approximation holds equally well for such systems, and show that, provided the obtained approximating system is asymptotically stable, the $O(\varepsilon)$ accuracy of the approximation can be extended uniformly to the semi-infinite interval $0 \leqq t < \infty$. Finally, in the latter case, Persek [4] demonstrates that the original aperiodic system is also asymptotically stable.

Now if $\bar{E}^{(1)}(w) \equiv 0 \equiv \bar{E}^{(2)}(w)$ for the periodic system (1), Laričeva [2] then replaces E by its nth order iterated average $\varepsilon^{n-1}\bar{E}^{(n)}(w)$ for some choice of $n \geqq 3$, resulting in a system which approximates solutions of (1) and (2) to $O(\varepsilon)$ accuracy on $0 \leqq t \leqq O(1/\varepsilon^n)$. In this paper, we will show that if the approximating system is asymptotically stable, the $O(\varepsilon)$ accuracy of the approximation can be extended to the full interval $0 \leqq t < \infty$. This is our main result and forms Case B of the theorem given in § 3. And since Laričeva's work restricts itself only to systems where $E = E(w, t)$ is independent of ε, we mention the results for the general situation where $E = E(w, t, \varepsilon)$ in Case A of our theorem.

2. Preliminaries. With w in R^k, consider the system

$$\frac{dw}{dt} = \varepsilon E(w, t, \varepsilon), \tag{3}$$

$$w|_{t=t_0(\varepsilon)} = \mathring{w}_0 + \varepsilon \mathring{w}_1(\varepsilon), \tag{4}$$

defined for $t_0(\varepsilon) \leqq t < \infty$ and $\varepsilon > 0$ small, where $t_0(\varepsilon) \geqq 0$ has been arbitrarily chosen. Now, for some $\varepsilon_D > 0$, define the set D consisting of points (w, t, ε) by $D =$

* Received by the editors October 5, 1979, and in revised form June 18, 1980.
† CBA-MGT, St. John's University, 160 Banbury Road, Mineola, New York 11501.

$D_w \times [0, \infty) \times [0, \varepsilon_D]$, with D_w a bounded convex open set in R^k. Let $\mathscr{S}_w$ be an open set with its closure contained in D_w, and let $\mathring{\mathscr{S}}_w$ be a subset of $\mathscr{S}_w$.

HYPOTHESIS H1 (periodicity, smoothness). *$E(w, t, \varepsilon)$ is periodic in t on the set D of fixed period $P > 0$. E and several of its derivatives with respect to (w, ε) are bounded uniformly on the set D, and for each fixed t $(0 \leqq t \leqq P)$ are smooth functions of (w, ε) on $D_w \times [0, \varepsilon_D]$. Finally, any $\mathring{w}_1(\varepsilon)$ used in (4) is assumed to be bounded in norm by a given constant on $0 \leqq \varepsilon \leqq \varepsilon_D$.*

Now consider a solution of (3), (4), at any two points t and $\hat{t}$, where $|t - \hat{t}| = O(1)$. With $w = w(t)$, $\hat{w} = w(\hat{t})$, $E(w, t, \varepsilon)$ can be expanded as a power series to some order in ε with coefficients in $(\hat{w}, t)$, provided

$$\sup_{\hat{t} \leqq s \leqq t} |w(s) - \hat{w}| = O(\varepsilon).$$

In fact,

$$\begin{aligned} &E(w(t), t, \varepsilon) \\ &\quad = E(\hat{w}, t, 0) + \left(\frac{\partial E}{\partial w}\right)(\hat{w}, t, 0)(w(t) - \hat{w}) + \left(\frac{\partial E}{\partial \varepsilon}\right)(\hat{w}, t, 0)\varepsilon \\ &\qquad + \frac{1}{2!}\left(\frac{\partial^2 E}{\partial w\, \partial w}\right)(\hat{w}, t, 0)(w(t) - \hat{w})(w(t) - \hat{w}) + \left(\frac{\partial^2 E}{\partial w\, \partial \varepsilon}\right)(\hat{w}, t, 0)(w(t) - \hat{w})\varepsilon \\ &\qquad + \frac{1}{2!}\left(\frac{\partial^2 E}{\partial \varepsilon^2}\right)(\hat{w}, t, 0)\varepsilon^2 + \cdots + \frac{1}{(n-1)!}\left(\frac{\partial^{n-1} E}{\partial \varepsilon^{n-1}}\right)(\hat{w}, t, 0)\varepsilon^{n-1} \\ &\qquad + O([\varepsilon + |w(t) - \hat{w}|]^n), \end{aligned}$$

with $(\partial E/\partial w)$ a Jacobian matrix, $(\partial^2 E/\partial w\, \partial w)(\hat{w}, t, 0)(w(t) - \hat{w})(w(t) - \hat{w})$ a tensor product, etc. Hence,

$$\begin{aligned} &E(w(t), t, \varepsilon) \\ &\quad = E(\hat{w}, t, 0) \\ &\qquad + \varepsilon\left(\frac{\partial E}{\partial w}\right)(\hat{w}, t, 0)\left[\int_{\hat{t}}^{t} E(\hat{w}, s, 0)\, ds + \varepsilon \int_{\hat{t}}^{t} \left(\frac{\partial E}{\partial w}\right)(\hat{w}, s, 0) \left\{\int_{t}^{s} E(\hat{w}, \phi, 0)\, d\phi\right\} ds \right. \\ &\qquad\qquad \left. + \varepsilon \int_{\hat{t}}^{t} \left(\frac{\partial E}{\partial \varepsilon}\right)(\hat{w}, s, 0)\, ds + \cdots\right] \\ &\qquad + \varepsilon\left(\frac{\partial E}{\partial \varepsilon}\right)(\hat{w}, t, 0) + \frac{\varepsilon^2}{2!}\left(\frac{\partial^2 E}{\partial w\, \partial w}\right)(\hat{w}, t, 0)\left[\int_{\hat{t}}^{t} E(\hat{w}, s, 0)\, ds + \cdots\right] \\ &\qquad\qquad \cdot \left[\int_{\hat{t}}^{t} E(\hat{w}, s, 0)\, ds + \cdots\right] \\ &\qquad + \varepsilon^2\left(\frac{\partial^2 E}{\partial w\, \partial \varepsilon}\right)(\hat{w}, t, 0)\left[\int_{\hat{t}}^{t} E(\hat{w}, s, 0)\, ds + \cdots\right] \\ &\qquad + \frac{\varepsilon^2}{2!}\left(\frac{\partial^2 E}{\partial \varepsilon^2}\right)(\hat{w}, t, 0) + \cdots + \frac{\varepsilon^{n-1}}{(n-1)!}\left(\frac{\partial^{n-1} E}{\partial \varepsilon^{n-1}}\right)(\hat{w}, t, 0) + O(\varepsilon^n). \end{aligned}$$

So for $|t - \hat{t}| = O(1)$ and $\sup_{\hat{t} \leqq s \leqq t} |w(s) - \hat{w}| = O(\varepsilon)$, a sufficiently smooth $E(w(t), t, \varepsilon)$

can be expanded as

$$E(w(t), t, \varepsilon) = \sum_{i=1}^{n} \varepsilon^{i-1} E^{(i)}(\hat{w}, t, \hat{t}) + R_n(w(t) - \hat{w}, t, \hat{t}, \varepsilon),$$

with

$$|R_n(w(t) - \hat{w}, t, \hat{t}, \varepsilon)| = O(\varepsilon^n).$$

HYPOTHESIS H2 (the nth iterated-average system). *For w in D_w, $0 \leqq \tau < P$ and $1 \leqq i \leqq n$, define the iterated averages*

$$\bar{E}^{(i)}(w, \tau) = \frac{1}{P} \int_{\tau}^{\tau+P} E^{(i)}(w, t, \tau)\, dt,$$

and assume $\bar{E}^{(i)}(w, \tau) \equiv 0$ for $1 \leqq i \leqq n-1$. Assume that $\bar{E}^{(n)}(w, \tau)$ is independent of τ, i.e., $\bar{E}^{(n)} = \bar{E}^{(n)}(w)$, and define the nth iterated-average system by

$$\frac{d\rho}{dt} = \varepsilon^n \bar{E}^{(n)}(\rho), \tag{5}$$

$$\rho|_{t=t_0(\varepsilon)} = \mathring{w}_0, \tag{6}$$

with $\bar{E}^{(n)}(w)$ smooth on D_w. Assume a constant $M > 0$ exists (M may be infinite) such that the solution $\rho = \rho(t, \varepsilon)$ to (5), (6), exists and remains in $\mathcal{S}_w$ for $t_0(\varepsilon) \leqq t \leqq t_0(\varepsilon) + M/\varepsilon^n$, $0 < \varepsilon < \varepsilon_D$, and for all $\mathring{w}_0$ in $\mathring{\mathcal{S}}_w$.

HYPOTHESIS H3 (stability of the variational system). *Assume $M = \infty$ and let $U(t, s)$ be the fundamental solution to*

$$\frac{dU}{dt} = \varepsilon^n \left(\frac{\partial \bar{E}^{(n)}}{\partial w} \right) (\rho(t, \varepsilon)) U, \qquad U|_{t=s} = I_{k \times k},$$

with $I_{k \times k}$ the identity matrix in $R^{k \times k}$. We assume constants $\bar{K}_n$ and $\lambda_n > 0$ exist (independent of ε, s, $t_0(\varepsilon)$ and the $\rho(t, \varepsilon)$ chosen from Hypothesis H2) such that, for $t_0(\varepsilon) \leqq s \leqq t < \infty$ and $0 < \varepsilon < \varepsilon_D$,

$$\|U(t, s)\| \leqq \bar{K}_n \exp(-\varepsilon^n \lambda_n (t - s)),$$

independent of the chosen ρ and $t_0(\varepsilon)$.

3. Result and applications.

THEOREM: *Let $w(t, \varepsilon)$ be the solution to the initial value problem (3), (4), with $t_0(\varepsilon) \geqq 0$ arbitrarily chosen, and let $\rho(t, \varepsilon)$ be a solution to the iterated average system (5), (6), for $0 < \varepsilon \leqq \varepsilon_D$:*

Case A. *Let Hypotheses* H1 *and H2 hold with M finite. Then constants $K_n(M)$ and $\varepsilon_n(M) > 0$ exist, (with values depending on D, $\mathcal{S}_w$, $\mathring{\mathcal{S}}_w$ and the bounds in* H1, H2, *but independent of $t_0(\varepsilon)$), such that for $0 < \varepsilon \leqq \varepsilon_n(M)$, the solutions $w(t, \varepsilon)$ and $\rho(t, \varepsilon)$ exist on $t_0(\varepsilon) \leqq t \leqq t_0(\varepsilon) + M/\varepsilon^n$, and*

$$\sup_{t_0(\varepsilon) \leqq t \leqq t_0(\varepsilon) + M/\varepsilon^n} |w(t, \varepsilon) - \rho(t, \varepsilon)| \leqq K_n(M)\varepsilon$$

uniformly for all $\mathring{w}_0$ in $\mathring{\mathcal{S}}_w$.

Case B. *Let Hypotheses* H1–H3 *hold ($M = \infty$). Then constants K_n^*, $\varepsilon_n^* > 0$ exist (with values independent of $t_0(\varepsilon)$, but depending on D, $\mathcal{S}_w$, $\mathring{\mathcal{S}}_w$, and the bounds in* H1–H3) *such that, for $0 < \varepsilon < \varepsilon_n^*$, the solutions $w(t, \varepsilon)$ and $\rho(t, \varepsilon)$ exist on $t_0(\varepsilon) \leqq t < \infty$,*

and

$$\sup_{t_0(\varepsilon)\leqq t<\infty} |w(t,\varepsilon)-\rho(t,\varepsilon)|\leqq K_n^*\varepsilon,$$

uniformly for all $\mathring{w}_0$ *in* $\mathring{\mathcal{S}}_w$.

Before beginning applications, we note the expansions preceding Hypothesis H2, and define the following iterated-averages through them:

$$\bar{E}^{(1)}(w,\tau)=\frac{1}{P}\int_\tau^{\tau+P} E(w,t,0)\,dt,$$

$$\bar{E}^{(2)}(w,\tau)=\frac{1}{P}\int_\tau^{\tau+P}\left(\frac{\partial E}{\partial w}\right)(w,t,0)\left\{\int_\tau^t E(w,s,0)\,ds\right\}dt+\frac{1}{P}\int_\tau^{\tau+P}\left(\frac{\partial E}{\partial \varepsilon}\right)(w,t,0)\,dt,$$

$$\begin{aligned}\bar{E}^{(3)}(w,\tau)=&\frac{1}{P}\int_\tau^{\tau+P}\left(\frac{\partial E}{\partial w}\right)(w,t,0)\\&\cdot\left[\int_\tau^t\left(\frac{\partial E}{\partial w}\right)(w,s,0)\left\{\int_\tau^s E(w,\phi,0)\,d\phi\right\}ds+\int_\tau^t\left(\frac{\partial E}{\partial \varepsilon}\right)(w,s,0)\,ds\right]dt\\&+\frac{1}{2P}\int_\tau^{\tau+P}\left(\frac{\partial^2 E}{\partial w\,\partial w}\right)(w,t,0)\left[\int_\tau^t E(w,s,0)\,ds\right]\left[\int_\tau^t E(w,\phi,0)\,d\phi\right]dt\\&+\frac{1}{P}\int_\tau^{\tau+P}\left(\frac{\partial^2 E}{\partial w\,\partial \varepsilon}\right)(w,t,0)\left[\int_\tau^t E(w,s,0)\,ds\right]dt\\&+\frac{1}{2P}\int_\tau^{\tau+P}\left(\frac{\partial^2 E}{\partial \varepsilon^2}\right)(w,t,0)\,dt.\end{aligned}$$

Example 1. With u scalar and α, β, γ constants, consider the equation

$$\frac{d^2u}{dt^2}+\varepsilon^2(\beta u-\gamma\varepsilon)\frac{du}{dt}+(1-\varepsilon\alpha u)u=0,$$

which in system form becomes

$$\frac{du}{dt}=-v,$$

$$\frac{dv}{dt}=u-\varepsilon(\alpha u^2+\varepsilon\beta uv-\gamma\varepsilon^2 v).$$

Then, letting $u=w\cos\theta$ and $v=w\sin\theta$, and using the (w,θ)-phase plane, we obtain

$$\begin{aligned}\frac{dw}{d\theta}&=-\varepsilon\frac{\alpha w^2\cos^2\theta\sin\theta+\varepsilon\beta w^2\cos\theta\sin^2\theta-\gamma\varepsilon^2 w\sin^2\theta}{1-\varepsilon(\alpha w\cos^3\theta+\varepsilon\beta w\cos^2\theta\sin\theta-\gamma\varepsilon^2\sin\theta\cos\theta)}\\&=-\varepsilon\alpha w^2\cos^2\theta\sin\theta-\varepsilon^2(\beta w^2\cos\theta\sin^2\theta+\alpha^2 w^3\cos^5\theta\sin\theta)\\&\quad+\varepsilon^3(\gamma w\sin^2\theta-2\alpha\beta w^3\cos^4\theta\sin^2\theta-\alpha^3 w^4\cos^8\theta\sin\theta)+\varepsilon^4 E_R(w,\theta,\varepsilon),\end{aligned}$$

with E_R a smooth function of ε. As $\bar{E}^{(1)}(w)=0=\bar{E}^{(2)}(w)$, we obtain $\bar{E}^{(3)}(w)=\gamma w/2-\alpha\beta w^3/8$ and the approximating system

$$\frac{d\rho}{d\theta}=\frac{\varepsilon^3}{8}(4\gamma\rho-\alpha\beta\rho^3),\qquad \rho|_{\theta=\theta_0}=w|_{\theta=\theta_0}=\mathring{w}_0,$$

with $\theta_0 = \theta(t=0)$. Since the ρ-system has a stable rest-point at $\rho = 2\sqrt{\gamma/(\alpha\beta)}$, then, by the theorem, constants K_3^* and $\varepsilon_3^* > 0$ exist independent of the choice of θ_0 and $w(\theta_0)$, (provided the latter is restricted to a bounded set with $w(\theta_0) > 0$) such that, for all θ_0 and for all $w(\theta_0)$ in this set,

$$\sup_{\theta_0 \leqq \theta < \infty} |w(\theta, \varepsilon) - \rho(\theta, \varepsilon)| \leqq K_3^* \varepsilon,$$

for all ε satisfying $0 < \varepsilon \leqq \varepsilon_3^*$. Note that we have restricted ourselves to the (w, θ)-phase plane to avoid having to deal with stable subsystems, as these will be treated in a later paper.

Example 2. With u and v scalars and $w = (u, v)$ a vector, consider the system

$$\frac{du}{dt} = \varepsilon f(u, v) \sin t,$$

$$\frac{dv}{dt} = \varepsilon g(u, v) \cos 2t.$$

With $\rho = (\zeta, \eta)$ corresponding to $w = (u, v)$, we obtain the averaged system

$$\frac{d\zeta}{dt} = \frac{\varepsilon^3}{8}\left\{\frac{\partial f(\zeta, \eta)}{\partial \zeta}\frac{\partial f(\zeta, \eta)}{\partial \eta} g(\zeta, \eta) - \frac{\partial^2 f(\zeta, \eta)}{\partial \zeta\, \partial \eta} f(\zeta, \eta) g(\zeta, \eta) - 2f(\zeta, \eta)\frac{\partial f(\zeta, \eta)}{\partial \eta}\frac{\partial g(\zeta, \eta)}{\partial \zeta}\right\},$$

$$\frac{d\eta}{dt} = \frac{\varepsilon^3}{8}\left\{f(\zeta, \eta)\frac{\partial f(\zeta, \eta)}{\partial \zeta}\frac{\partial g(\zeta, \eta)}{\partial \zeta} + f^2(\zeta, \eta)\frac{\partial^2 g(\zeta, \eta)}{\partial \zeta\, \partial \zeta}\right\}.$$

In particular, choosing $f(u, v) = uv$ and $g(u, v) = \alpha u + \beta u^2 + \gamma u^3 v$ with α, β, γ constants, we obtain the iterated-average system

$$\frac{d\zeta}{dt} = -\frac{\varepsilon^3}{4}\zeta^2\eta(\alpha + 2\beta\zeta + 3\gamma\zeta^2\eta), \qquad \zeta|_{t=0} = u(t=0),$$

$$\frac{d\eta}{dt} = \frac{\varepsilon^3}{8}\zeta\eta^2(\alpha + 4\beta\zeta + 9\gamma\zeta^2\eta), \qquad \eta|_{t=0} = v(t=0),$$

which has a stable rest-point located at $(\zeta, \eta) = (-\alpha/\beta, \beta^2/(3\alpha\gamma))$, provided $\alpha < 0$, $\beta > 0$, $\gamma < 0$ (actually, provided $\alpha\gamma/\beta > 0$). Then by the theorem K_3^* and $\varepsilon_3^* > 0$ exist such that, for all $t_0(\varepsilon) \geqq 0$ and all initial values in a bounded subdomain of stability of the averaged system,

$$\sup_{t_0(\varepsilon) \leqq t < \infty} \{|u(t, \varepsilon) - \zeta(t, \varepsilon)| + |v(t, \varepsilon) - \eta(t, \varepsilon)|\} \leqq K_3^* \varepsilon,$$

for $0 < \varepsilon \leqq \varepsilon_3^*$, independent of the choice of initial values.

4. Proof of the theorem. With $w = w(t, \varepsilon)$ the solution to (3), (4), and $\rho = \rho(t, \varepsilon)$ the solution to (5), (6), let $w = \rho + \varepsilon W$. Let d be the distance between the boundaries of $\mathcal{S}_w$ and D_w, and let N_1 majorize $E(w, t, \varepsilon)$ and its appropriate derivatives on the set D,

and bound $\mathring{w}_1(\varepsilon)$, the initial data for $0 \leqq \varepsilon \leqq \varepsilon_D$. We now write

$$\frac{dW}{dt} = E(w, t, \varepsilon) - \varepsilon^{n-1}\bar{E}^{(n)}(\rho), \tag{7}$$

$$W|_{t=t_0(\varepsilon)} = \mathring{w}_1(\varepsilon). \tag{8}$$

Now, given any constant $a > \sup_{0<\varepsilon\leqq\varepsilon_D} |\mathring{w}_1(\varepsilon)|$, there exists some $t_1(\varepsilon) > t_0(\varepsilon)$, $(t_1(\varepsilon) \leqq t_0(\varepsilon) + M/\varepsilon^n)$ such that $w(t, \varepsilon)$ exists on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$ for $0 < \varepsilon \leqq \varepsilon_D$, and $\sup_{t_0(\varepsilon)\leqq t\leqq t_1(\varepsilon)} |W(t, \varepsilon)| \leqq a$ for $0 < \varepsilon \leqq \varepsilon_D$. Choose $\varepsilon_1 = \min(\varepsilon_D, d/(2a), 1)$; then, by Hypothesis H2, $w(t, \varepsilon)$ lies in D_w for $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$, $0 < \varepsilon \leqq \varepsilon_1$. Now with $[\cdot]$ denoting the greatest integer function define the function $\hat{t}$ by $\hat{t} = [(t - t_0(\varepsilon))/P]P + t_0(\varepsilon)$, and $\hat{w}$ by $\hat{w} = \hat{w}(t, \varepsilon) \equiv w(\hat{t}, \varepsilon)$. Since $|w - \hat{w}| = O(\varepsilon)$ on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$ and $0 < \varepsilon \leqq \varepsilon_1$, then $E(w, t, \varepsilon)$ in (7) can be expanded in powers of ε according to the procedure outlined following Hypothesis H1. Hence (7) becomes

$$\frac{dW}{dt} = \sum_{i=1}^{n} \varepsilon^{i-1} E^{(i)}(\hat{w}, t, \hat{t}) + R_n(w - \hat{w}, t, \hat{t}, \varepsilon) - \varepsilon^{n-1}\bar{E}^{(n)}(\rho),$$

and a constant N_2 also exists independent of ε, a, $t_0(\varepsilon)$, $t_1(\varepsilon)$ such that, for $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$ and $0 < \varepsilon \leqq \varepsilon_1$, $|R_n(w(t, \varepsilon) - \hat{w}(t, \varepsilon), t, \varepsilon)| \leqq N_2\varepsilon^n$. Since, by Hypothesis H2, $\bar{E}^{(i)}(w, \tau) \equiv 0$ on D_w for $1 \leqq i \leqq n-1$ and $\bar{E}^{(n)}(w, \tau) = \bar{E}^{(n)}(w)$, then, with $w = \rho + \varepsilon W$,

$$\begin{aligned}\frac{dW}{dt} = \sum_{i=1}^{n} \varepsilon^{i-1}\{E^{(i)}(\hat{w}, t, \hat{t}) - \bar{E}^{(i)}(\hat{w}, \hat{t})\} + \varepsilon^{n-1}\{\bar{E}^{(n)}(\hat{w}) - \bar{E}^{(n)}(w)\} \\ + \varepsilon^{n-1}\{\bar{E}^{(n)}(\rho + \varepsilon W) - \bar{E}^{(n)}(\rho)\} + R_n(w - \hat{w}, t, \hat{t}, \varepsilon).\end{aligned}$$

So, with $\varepsilon_2 = \min(\varepsilon_1, 1/a^2)$, a constant N_3 exists independent of ε, a, $t_0(\varepsilon)$ and $t_1(\varepsilon)$ such that

$$\frac{dW}{dt} = \sum_{i=1}^{n} \varepsilon^{i-1}\{E^{(i)}(\hat{w}, t, \hat{t}) - \bar{E}^{(i)}(\hat{w}, \hat{t})\} + \varepsilon^n\left(\frac{\partial \bar{E}^{(n)}}{\partial w}\right)(\rho)W + C(W, t, \varepsilon), \tag{9}$$

where $|C(W(t, \varepsilon), t, \varepsilon)| \leqq N_3\varepsilon^n(1 + \varepsilon a^2) \leqq 2N_3\varepsilon^n$ on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$, for $0 < \varepsilon < \varepsilon_2$. Defining $U(t, s)$ as in Hypothesis H3 and noting that

$$\frac{dU^{-1}(s, t_0(\varepsilon))}{ds} = -U^{-1}(s, t_0(\varepsilon))\frac{dU(s, t_0(\varepsilon))}{ds}U^{-1}(s, t_0(\varepsilon))$$

and

$$U(t, s) = U(t, t_0(\varepsilon))U^{-1}(s, t_0(\varepsilon)),$$

we can write (8) and (9) as

$$\begin{aligned}W(t) &= U(t, t_0(\varepsilon))\mathring{w}_1(\varepsilon) \\ &+ \sum_{i=1}^{n} \varepsilon^{i-1} \int_{t_0(\varepsilon)}^{t} U(t, s) \times \{E^{(i)}(w(\hat{s}), s, \hat{s}) - \bar{E}^{(i)}(w(\hat{s}), \hat{s})\}\, ds \\ &+ \int_{t_0(\varepsilon)}^{t} U(t, s)C(W(s), s, \varepsilon)\, ds,\end{aligned} \tag{10}$$

where $\hat{s}$ is defined like $\hat{t}$, and we have abbreviated $W(t, \varepsilon)$ by $W(t)$, $w(\hat{t}, \varepsilon)$ by $w(\hat{t})$, etc.

Using the properties of U, we integrate (10) by parts:

$$
\begin{aligned}
W(t) &= U(t, t_0(\varepsilon))\mathring{w}_1(\varepsilon) \\
&\quad + \sum_{i=1}^{n} \varepsilon^{i-1} \int_{t_0(\varepsilon)}^{t} \{E^{(i)}(w(\hat{s}), s, \hat{s}) - \bar{E}^{(i)}(w(\hat{s}), \hat{s})\}\, ds \\
&\quad + \sum_{i=1}^{n} \varepsilon^{i-1} \int_{t_0(\varepsilon)}^{t} U(t, s)\varepsilon^{n}\left(\frac{\partial \bar{E}^{(n)}}{\partial w}\right)(\rho(s, \varepsilon)) \\
&\qquad\qquad \cdot \left[\int_{t_0(\varepsilon)}^{s} \{E^{(i)}(w(\hat{\phi}), \phi, \hat{\phi}) - \bar{E}^{(i)}(w(\hat{\phi}), \hat{\phi})\}\, d\phi\right] ds \\
&\quad + \int_{t_0(\varepsilon)}^{t} U(t, s)C(W(s), s, \varepsilon)\, ds.
\end{aligned} \tag{11}
$$

Then a constant N_4 exists independent of ε, a, $t_0(\varepsilon)$, $t_1(\varepsilon)$, majorizing $(\partial \bar{E}^{(n)}/dw)(w)$ in norm on D_w and, by Hypotheses H1–H2, majorizing

$$\int_{t_0(\varepsilon)}^{t} \{E^{(i)}(w(\hat{s}), s, \hat{s}) - \bar{E}^{(i)}(w(\hat{s}), \hat{s})\}\, ds,$$

for $1 \leqq i \leqq n$, on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$, $0 < \varepsilon \leqq \varepsilon_2$. Then, for $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$, $0 < \varepsilon \leqq \varepsilon_2$, with $\|\cdot\|$ the vector or matrix norm,

$$
\begin{aligned}
\|W(t)\| &\leqq \|U(t, t_0(\varepsilon))\|N_1 + N_4 \sum_{i=1}^{n} \varepsilon^{i-1} \\
&\quad + \int_{t_0(\varepsilon)}^{t} \|U(t, s)\|\varepsilon^{n}\left\{N_4^2 \sum_{i=1}^{n} \varepsilon^{i-1} + 2N_3\right\} ds.
\end{aligned} \tag{12}
$$

The rest of the proof for Case A. Now, from Hypothesis H2 and the above, $\|U(t, s)\| \leqq e^{\varepsilon^n N_4(t-s)}$ on $t_0(\varepsilon) \leqq s \leqq t \leqq t_0(\varepsilon) + M/\varepsilon^n$, $0 < \varepsilon \leqq \varepsilon_D$. Hence, for $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon) \leqq t_0(\varepsilon) + M/\varepsilon^n$ and $0 < \varepsilon \leqq \varepsilon_2 \equiv \min(\varepsilon_D, d/(2a), 1, 1/a^2)$, by a Gronwall inequality,

$$
\begin{aligned}
\|W(t)\| &\leqq \left(N_1 + N_4 \sum_{i=1}^{n} \varepsilon^{i-1} + \frac{2N_3}{N_4}\right) \exp(\varepsilon^n N_4(t - t_0(\varepsilon))) + N_4 \sum_{i=1}^{n} \varepsilon^{i-1} \\
&\leqq \left(N_1 + nN_4 + \frac{2N_3}{N_4}\right) e^{N_4 M} + nN_4 \equiv \frac{1}{2}K_n(M),
\end{aligned}
$$

with $K_n(M)$ independent of ε, a, $t_0(\varepsilon)$ and $t_1(\varepsilon)$. Hence, if a is taken as equal to $K_n(M)$ and $\varepsilon_n(M) = \min\{\varepsilon_D, d/[2K_n(M)], 1, 1/[K_n(M)]^2\}$, then, if $t_1(\varepsilon)$ is picked such that $\|W(t)\| \leqq a$ on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$ and $0 < \varepsilon \leqq \varepsilon_n(M)$, it follows that $\|W(t)\| \leqq a/2$ there. Thus we are free to choose $t_1(\varepsilon) = t_0(\varepsilon) + M/\varepsilon^n$, from which it follows that $\sup_{t_0(\varepsilon) \leqq t \leqq t_0(\varepsilon) + M/\varepsilon^n} \|W(t)\| < K_n(M)$, independent of all $\mathring{w}_0$ in $\mathring{\mathcal{S}}_w$ and with $0 < \varepsilon \leqq \varepsilon_n(M)$.

The rest of the proof for Case B. Now from Hypothesis H3, constants $\bar{K}_n$ and $\lambda_n > 0$ exist such that $\|U(t, s)\| \leqq \bar{K}_n \exp(-\varepsilon^n \lambda_n(t - s))$ for $t_0(\varepsilon) \leqq s \leqq t < \infty$, $0 < \varepsilon < \varepsilon_D$, and all

$\mathring{w}_0$ in $\mathring{\mathscr{S}}_w$. So from (12), for $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$, $0 < \varepsilon \leqq \varepsilon_2 \equiv \min(\varepsilon_D, d/(2a), 1, 1/a^2)$,

$$\begin{aligned}\|W(t)\| &\leqq N_1 \bar{K}_n \exp(-\varepsilon^n \lambda_n (t - t_0(\varepsilon))) + N_4 \sum_{i=1}^{n} \varepsilon^{i-1} \\ &\quad + \left\{ N_4^2 \sum_{i=1}^{n} \varepsilon^{i-1} + 2N_3 \right\} \bar{K}_n (1 - \exp(-\varepsilon^n \lambda_n (t - t_0(\varepsilon)))/\lambda_n) \\ &\leqq N_1 \bar{K}_n + nN_4 + (nN_4^2 + 2N_3) \frac{\bar{K}_n}{\lambda_n} \equiv \frac{1}{2} K_n^*,\end{aligned}$$

with K_n^* independent of ε, a, $t_0(\varepsilon)$ and $t_1(\varepsilon)$. Hence, if we choose $a \equiv K_n^*$ and $\varepsilon_n^* \equiv \min[\varepsilon_D, d/(2K_n^*), 1, 1/(K_n^*)^2]$, then if $t_1(\varepsilon)$ is picked such that $\|W(t)\| \leqq a$ on $t_0(\varepsilon) \leqq t \leqq t_1(\varepsilon)$ and $0 < \varepsilon \leqq \varepsilon_n^*$, it follows that $\|W(t)\| \leqq a/2$ there. Thus we are free to choose $t_1(\varepsilon) = \infty$, from which it follows that $\sup_{t_0(\varepsilon) \leqq t < \infty} \|W(t)\| < K_n^*$, for $0 < \varepsilon \leqq \varepsilon_n^*$ and for all $\mathring{w}_0$ in $\mathring{\mathscr{S}}_w$. The theorem is proved.

REFERENCES

[1] N. Bogoliubov and Yu. Mitropolskii, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Gordon and Breach, New York, 1961.

[2] V. V. Ličeva, *On the extension of the averaging interval*, Soviet Math. Dokl., 16 (1975), pp. 146–150.

[3] S. Persek and F. Hoppensteadt, *Iterated averaging methods for systems of ordinary differential equations with a small parameter*, Comm. Pure Appl. Math., XXXI (1978), pp. 135–156.

[4] S. Persek, *Asymptotic stability and the method of iterated averages*, in preparation.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0011 $01.00/0

STABLE SOLUTION OF THE INVERSE REFLECTION PROBLEM FOR A SMOOTHLY STRATIFIED ELASTIC MEDIUM*

W. W. SYMES†

Abstract. The subject of this paper is a version of the inverse reflection problem for a smoothly stratified elastic medium. The same mathematics describes the inverse problem of the vibrating string. This problem is solved in a constructive way. Also, a priori estimates are derived which exhibit the continuous dependence of the solution (index of refraction, relative sound speed) on the data (scattering or reflection measurements).

1. Introduction. The subject of this paper is a version of the inverse reflection problem for a smoothly stratified elastic medium. The same mathematics describes the inverse problem of the vibrating string. This problem is solved in a constructive way. Also, a priori estimates are derived which exhibit the continuous dependence of the solution (index of refraction, relative sound speed) on the data (scattering or reflection measurements).

We take particular care to use only constructions which apply in principle to higher dimensional problems of a similar sort. Most other treatments of this problem proceed via a reduction to an inverse Sturm–Liouville problem (see, e.g., [1] and [2]). This coordinate transformation is no longer available in higher-dimensional problems, so we avoid it, except where it can be interpreted as a coordinate transformation along the rays of geometric topics; this is the case for the a priori estimates of c at the end of § 4. (See § 4 for a discussion of this point.)

Other authors who avoid dependence on peculiarly one-dimensional tricks have invariably used approximate methods (JWKB, Born series: see [3], [4, Ch. XIII] and practically any article in the literature of inverse scattering in exploration geophysics, physical chemistry, ultrasound radiology, etc.). In contrast, our methods are exact: we construct the index of refraction *exactly* by an iterative approximation procedure.

Our formulation of the inverse reflection problem is time dependent, which also contrasts with many other treatments (e.g., [2], [5]). The solution propounded here can also be adapted to steady state (frequency domain) problem formulations, other boundary conditions, etc.

Our presentation shares some features with the work of Weston and Kreuger (see [17] and references cited therein), and Kay [18]. Sondhi and Gopinath [21] seem to have been the first to notice the attractive computational (stability) aspects of the nonlinear Gel'fand–Levitan Volterra equation, which is related to the main analytical device of the present work. None of these authors, however, prove stability results. Regularization techniques from the theory of ill-posed problems have been applied to reflection data inversion by M. Gerver [20] and, more recently, by A. Bamberger et al. [19]. Our results show that the problem for smooth stratification may be regarded as a well-posed problem, rather than as a regularization of an ill-posed problem. We remark that we have recently shown that the crucial parameter, denoted by ε below, which determines the condition number or local Lipshitz constant of the problem, is closely related to the *rate of local energy decay*; see [12] for details. Finally, we mention the

* Received by the editors November 1, 1979, and in revised form August 25, 1980. This work was sponsored by the United States Army under contract DAAG29-75-C-0024. This material is based on work supported by the National Science Foundation under grant MCS78-09525, and performed in part at the Mathematics Research Center, University of Wisconsin, Madison. A preliminary form of the paper appeared as Technical Summary Report 2007, MRC, University of Wisconsin.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.

work of O. Hald [8] on inverse Sturm–Liouville problems and P. Deift and E. Trubowitz [9] on inverse Schrödinger scattering, who also derive stability results in those contexts. Their techniques are closely related in several respects to those presented here.

We abjure further discussion of the literature and proceed to describe our results.

The physical setting of our problem is as follows. Consider a stratified elastic medium with unit elastic moduli (Lamé constants). (This latter restriction may be removed, in which case we recover the elastic impedance rather than the density). The density ρ varies as a function of only one of the coordinates, say x. We assume that ρ is known and constant in the halfspace $\{x \leqq 0\}$; in fact, for the sake of simplicity, $\rho \equiv 1$ for $x \leqq 0$ and smooth (of class C^2) otherwise. We set up a measuring device at $x = 0$, and record the (infinitesimal) reflected wave $F(t)$, $t > 0$ resulting from an (infinitesimal) impulsive incident wave of the form $\delta(x - t)$, $t < 0$ impinging on the unknown medium $\{x \geqq 0\}$.

Our results are phrased in terms of the index of refraction $c = \rho^{-1/2}$. We obtain sufficient conditions on F (which are very close to necessary), as in

THEOREM B. *Suppose $T > 0$, $F:[0, 2T] \to \mathbb{R}$ is of class C^1, $F(0) = 0$, and the kernel*

$$H(s, t) = \tfrac{1}{2}F(s + t) - \int_0^s d\tau\, F(s - \tau)F(t - \tau), \qquad 0 \leqq s \leqq t \leqq T,$$

$$H(t, s) \equiv H(s, t),$$

defines a self-adjoint Hilbert–Schmidt operator $\mathbb{H}$ on $L^2[0, T]$ with the property

$$\mathbb{I} + 4\mathbb{H} \geqq \varepsilon > 0.$$

Then there exists a unique $X > 0$ and a unique positive function $c : [0, \mathbb{R}] \to \mathbb{R}^+ = \{c > 0\}$ of class C^2 so that $\int_0^X c^{-1} = T$ and $F(t) = u(0, t)$, $t > 0$ for the solution u of the initial value problem (for which define $c \equiv 1$ for $x < 0$):

$$\left(\frac{\partial}{\partial t^2} - c^2(x)\frac{\partial^2}{\partial x^2}\right)u(x, t) \equiv 0, \tag{1.1a}$$

in the domain of dependence of $(0, 2T)$

$$u(x, 0) = \delta(x), \qquad \frac{\partial u}{\partial t}(x, 0) \equiv 0. \tag{1.1b}$$

Denote by $\mathcal{S}$ the set

$$\mathcal{S} = \{(X, c) : X > 0,\ c \in C^2[0, X]\}.$$

$\mathcal{S}$ is given the topology determined by the following distance function:

$$d_{\mathcal{S}}((X, c), (\bar{X}, \bar{c})) = |X - \bar{X}| + \|c - \bar{c}\|_{C^2[0, \min(X, \bar{X})]}.$$

Then the map $F \to (X, c)$ whose existence is implicit in the above statement is continuous (even locally Lipschitz) as a map

$$\mathcal{T}_T = \{F \in C^1[0, 2T] : F(0) = 0,\ \mathbb{I} + 4\mathbb{H} > 0\} \to \mathcal{S}.$$

The proof proceeds along the lines laid down in [6], where a similar problem was solved with the differential equation (1.1a) replaced by

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} + q(x)\right)u(x, t) = 0.$$

In reference [6], therefore, the characteristics of the problem are known from the outset. The characteristics are exactly what is to be found in the present problem, however, which gives it the nature of a free boundary problem. This nature is unavoidable in higher-dimensional inverse problems for elastic waves, and we meet it head-on, which causes some headaches. In particular, the necessary a priori estimates are more difficult to derive than are the corresponding estimates in [6].

The main tool is the *progressing wave decomposition* of the solution u of an initial value problem (1.1):

$$u(x, t) = \tfrac{1}{2}c^{1/2}(x)[\delta(t+T(x))+\delta(t-T(x))]+K(x, t),$$

where

$$T(x) = \int_0^x c^{-1}$$

is the *travel-time function* and K is smooth inside the light cone with apex $(0, 0)$ (§ 2).

This expansion is an example of the progressing wave construction of Luneburg, Lax, Courant and Ludwig. Detailed calculations of this type seem to be rare in the literature. Also, the discussions in the literature typically assume infinitely differentiable coefficients, whereas we require a result for finitely differentiable coefficients. In view of these circumstances, we have decided to include in § 2 a (rather lengthy) detailed discussion of the progressing wave decomposition for (1.1). We show that the type of expansion depicted above exists when ρ has two measurable, bounded derivatives (Theorem A). In that case (and generally not otherwise) the remainder or reflected wave K has *finite energy*.

We remark that a sharp converse to Theorem B can be formulated: if F has one bounded measurable derivative, then ρ has two. The proof is similar to the proof of the sharp results in [6], and we do not give the details here.

In § 3, it is shown that the pair (K, c) solves the GL system of Volterra equations (Theorem C):

$$H(T(x), t) = \tfrac{1}{2}c^{-1/2}(x)K(x, t) + \int_0^x dy\, c^{-2}(y)K(y, T(x))K(y, t) \quad \text{for} \quad t \geqq T(x), \tag{1.2a}$$

$$K(T(x), x) = \tfrac{1}{4}c^{1/2}(x)\int_0^x c^{1/2}(c^{1/2})''. \tag{1.2b}$$

The initials GL stand for Gel'fand and Levitan, for they introduced integral equations of this sort into inverse scattering theory in their seminal paper [13]. Our work is in direct line of descent from theirs; the "nonlinear integral equation" of [13] is derived by the present techniques in [6]. GL also stands for *group law*; *indeed, the equation* (1.2a) *expresses in compact form the propagation of Cauchy data for* $\Box_c$ *by a one-parameter group of operators, which follows from the time-independence of the coefficient.*

A number of crucial a priori estimates are determined in § 4 for the solution $\{K, c\}$ of (1.2). These involve sup norms of F and the number ε^{-1}. This latter number, although in principle present in the data F, is in practice difficult to extract. On the other hand, in practical problems one often has known bounds on the density. It is shown in § 6 that a priori bounds on c and its first two derivatives determine a lower bound for ε, hence, can be employed in place of ε in the a priori and stability estimates. We use the results of § 4 in § 5 to show that an iteration scheme converges to a global solution of the

GL system (1.2). The solution defines a continuous map

$$\mathcal{T}_T \to \mathcal{T} = \{(X, c, K) : X > 0,\ c \in C^2[0, X],\ K \in C^1(\mathcal{C}(T, c))\},$$

where

$$\mathcal{C}(T, c) = \{(x, t) : 0 \leqq x \leqq X,\ T(x) \leqq t \leqq 2T - T(x)\}.$$

Finally, we establish in § 7 that the solution (K, c) of the GL system actually solves the *Chudov boundary value problem*

$$\left(\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2}\right) K = 0 \quad \text{in } \mathcal{C}(T, c), \tag{1.3a}$$

$$K(0, t) = F(t), \qquad \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) K(0, t) \equiv 0, \tag{1.3b}$$

$$K(x, T(x)) = \tfrac{1}{4} c^{1/2}(x) \int_0^x c^{1/2} (c^{1/2})''. \tag{1.3c}$$

Since it was established in § 2 that the (necessarily unique) solution of (1.3) is the smooth part of the solution of the singular initial value problem (1.1), it follows that c is the solution of the inverse problem sought in Theorem B, which completes its proof.

The result of § 7 also establishes that the GL system (1.2) and the Chudov problem (1.3) are completely equivalent. In particular, the a priori estimates of § 4 also hold for the solution of (1.3). In the context of the inverse problem of [6], a similar observation was used in [7] to prove stability of an optimally efficient numerical scheme based on a Chudov problem. In fact, the Chudov problem of [6], [7] was suggested as an approach to the inverse spectral problem for the Schrödinger operator by Chudov in the mid 1950's (see [14, Appendix]), hence the name. The present results will likewise provide the base for a stability result and consequent a priori error estimates for efficient numerical solution of the present problem. This matter will be discussed elsewhere.

To end this introduction, we note that the present results can be combined with well-known techniques from exploration seismology to solve the inverse problem for piecewise C^2 index of refraction with jump discontinuities.

2. The progressing wave expansion. The goal of this section is to express the solution u of the singular initial value problem

$$\Box_c \bar{u} = \left(\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2}\right) \bar{u} \equiv 0, \tag{2.1a}$$

$$\bar{u}(x, 0) = \delta(x - x_0), \qquad \frac{\partial \bar{u}}{\partial t}(x, 0) \equiv 0, \tag{2.1b}$$

in the form

$$\bar{u}(x, t) = \tfrac{1}{2} c^{1/2}(x) c^{-3/2}(x_0)[\delta(t + T(x, x_0)) + \delta(t - T(x, x_0))] + K(x, t; x_0), \tag{2.2}$$

where

$$T(x, x_0) = \int_{x_0}^x c^{-1}$$

and the remainder $K(x, t; x_0)$ is smooth inside the light cone with apex $(x_0, 0)$ and has a jump discontinuity on the boundary of the light cone given by

$$K(x, \pm T(x, x_0); x_0) = \pm \tfrac{1}{4} c(x_0)^{-3/2} c^{1/2}(x) \int_{x_0}^{x} c^{1/2}(c^{1/2})'', \qquad x > x_0. \tag{2.3}$$

Such representations do not hold without some smoothness restrictions on c. We shall at first assume that c is infinitely differentiable, and at the end determine to what extent this requirement can be relaxed. In fact, a version of (2.2), (2.3) holds when c has two bounded, measurable derivatives (Theorem A).

The decomposition (2.2), (2.3) is a typical *progressing wave expansion*, a device for analyzing the singular parts of solutions of hyperbolic initial value problems, which is closely related to the asymptotic expansions of geometric optics. Indeed, the solution of the initial value problem for singular initial data is reduced, modulo the solution of an initial value problem with smooth data, to the solution of a sequence of linear ordinary differential equations along the rays of geometric optics. For a general discussion of this idea and its background, we refer the reader to the treatise of Courant and Hilbert ([11, Chapt. 6, § 4, esp. pp. 618–635], see also [10, Chapt. 7]). Since detailed calculations of this expansion seem to be rare, we will give a thorough treatment for the general operator of one-dimensional linear elasticity,

$$\begin{aligned} W &= W(x, \partial_t; \partial x) \\ &= \rho(x)\partial_t^2 - \partial_x(E(x)\partial_x), \end{aligned}$$

and initial values

$$u(x, 0; x_0) = \delta(x - x_0), \qquad \partial_t u(x, 0; x_0) \equiv 0.$$

Here ρ, E are smooth and positive on $\mathbb{R}$. We will specialize at the conclusion of the section to the special case $\rho = c^{-2}$, $E \equiv 1$, to obtain the formulas listed above.

The expansions involve *compound distributions* (terminology of Lax [10]) of the form $g\delta^{(k)}(\varphi)$, where $g, \varphi \in C^\infty(\mathbb{R}^n)$, φ satisfies the condition

$$d\varphi(x) \neq 0 \quad \text{if } \varphi(x) = 0$$

and $\delta^{(k)}$ denotes the kth derivative of the Dirac delta distribution. These distributions are given an exhaustive treatment in Gel'fand–Shilov [15, Chapt. III, pp. 209–247].

We first recall the definition of $\delta(\varphi)$ (we refer the reader to [15] for the careful definition of $\delta^{(k)}(\varphi)$, $k \geqq 1$). As shown in [15], there is an $(n-1)$-form ω on $\mathbb{R}^n$ for which

$$d\varphi \wedge \omega = dx_1 \wedge \cdots \wedge dx_n.$$

This form is not unique; however, any two choices differ by an $(n-1)$-form containing $d\varphi$ as a factor, hence

$$\langle \delta(\varphi), u \rangle = \int_{\{\varphi = 0\}} u\omega$$

is well defined and defines the functional $\delta(\varphi)$.

In the following statements, abstracted from [15], we use the notation

$$\delta^{(-1)}(\varphi) = H(\varphi) = \begin{cases} 1, & \varphi \geqq 0, \\ 0, & \varphi < 0, \end{cases}$$

(a) $$\frac{\partial}{\partial x_j}\delta^{(k)}(\varphi) = \frac{\partial \varphi}{\partial x_j}\delta^{(k+1)}(\varphi), \qquad k = -1, 0, 1, \cdots,$$

(b) $$\varphi^j \delta^{(k)}(\varphi) = 0 \quad \text{if } j > k \geqq 0,$$

(c) $$\varphi\delta^{(k)}(\varphi) + k\delta^{(k-1)}(\varphi) = 0, \qquad k = 0, 1, 2, \cdots.$$

The statement (c) also implies the first part of

LEMMA 1. (extension lemma). *A distribution of the form $g\delta^{(k)}(\varphi)$ extends to a continuous linear functional on $C_0^{k-1}(\mathbb{R}^n)$ if and only if $g = \varphi f$ for some $f \in C^\infty(\mathbb{R}^n)$.*

Proof. We give a proof for $k = 1$ only. The proof for general k is similar. It is clear from the definition given above that $\delta^{(k)}(\varphi)$ extends continuously to a continuous linear functional on $C_0^k(\mathbb{R}^n)$, $k = 1, 2, \cdots$. Therefore, in view of (c) above, we need only prove the following statement:

Suppose $g\delta'(\varphi)$ extends to a continuous linear functional on $C_0^0(\mathbb{R}^n)$. Then $g = f\varphi$, some $f \in C^\infty(\mathbb{R}^n)$.

Indeed, from [15, p. 288 ff.] one obtains an expression for $g\delta'(\varphi)$ as follows. One may choose, in a suitable neighborhood U of a point in $\{\varphi = 0\}$, coordinates $y_1, \cdots, y_n$ so that $\varphi = y_1$. Then, for any $u \in C_0^\infty(\mathbb{R}^n)$ with support in U,

$$\begin{aligned} \langle g\delta'(\varphi), u\rangle &= \langle \delta'(\varphi), gu\rangle \\ &= \int \cdots \int dy_2 \cdots dy_n \frac{\partial}{\partial y_1}(Jgu)(0, y_2, \cdots, y_n) \\ &= \int \cdots \int dy_2 \cdots dy_n \left\{ \frac{\partial g}{\partial y_1} Ju + g\frac{\partial}{\partial y_1}(Ju)\right\}. \end{aligned}$$

For any $v \in C_0^\infty(U)$, $\lambda \in \mathbb{R}$, one may choose $u \in C_0^\infty(U)$ so that $Ju = e^{i\lambda y}v$, and in so doing $\sup_{x \in U}|u(x)|$ is bounded uniformly in λ. Then the above may be written

$$\langle g\delta'(\varphi), u\rangle = \int \cdots \int dy_2 \cdots dy_n \frac{\partial g}{\partial y_1} v + g\frac{\partial v}{\partial y_1} + i\lambda \int \cdots \int dy_2 \cdots dy_n\, gv.$$

Under the hypotheses, the left-hand side and the first term on the right-hand side are bounded uniformly in λ. It follows that

$$\int \cdots \int dy_2 \cdots dy_n\, gv = 0.$$

But $v \in C_0^\infty(U)$ was arbitrary, so $g \equiv 0$ on $\{\varphi = 0\} \cap U$. Also, U is an arbitrary suitably small neighborhood of a point in $\{\varphi = 0\}$, so in fact $g \equiv 0$ on $\{\varphi = 0\}$. A standard calculus argument then shows that $g = \varphi f$ for suitable $f \in C^\infty(\mathbb{R}^n)$. Q.E.D.

The starting point for the progressing wave expansion for W is the observation that there exist distributions of the form $v = g\delta(\varphi)$, for suitable choice of g, $\varphi \in C^\infty(\mathbb{R})^2$, for which the singular nature of Wv is no worse than that of v: precisely, Wv extends to a continuous linear functional on the space C_0^0 of continuous functions of compact

support, as does v. To see that this is so, compute Wv using the chain rule (a) to obtain

$$\begin{aligned}(\rho\partial_t^2-\partial_x E\partial_x)g\delta(\varphi) &= g(\rho(\partial_t\varphi)^2-E(\partial_x\varphi)^2)\delta''(\varphi)+(2\rho\partial_t g\partial_t\varphi-2E\partial_x g\partial_x\varphi)\delta'(\varphi)\\ &\quad+[g(\rho\partial_t^2\varphi-\partial_x(E\partial_x\varphi))]\delta'(\varphi)+(\rho\partial_t^2 g-\partial_x(E\partial_x g))\delta(\varphi)\\ &= \mathscr{E}\delta''(\varphi)+\mathscr{T}^0\delta'(\varphi)+\mathscr{R}^0\delta(\varphi),\end{aligned}$$

whereby $\mathscr{E}$, $\mathscr{T}^0$ and $\mathscr{R}^0$ are defined. This distribution is then required to extend continuously to $C_0^0(\mathbb{R}^2)$. A fortiori, it must extend continuously to $C_0^1(\mathbb{R}^2)$. Since the last two terms have this property, the first term $\mathscr{E}\delta''(\varphi)$ must itself extend continuously to $C_0^1(\mathbb{Z}^2)$. According to Lemma 4 above, there must exist $\mathscr{E}_1\in C^\infty(\mathbb{R}^2)$ so that $\mathscr{E}=\varphi\mathscr{E}_1$ and

(e) $$\mathscr{E}=\rho(\partial_t\varphi)^2-E(\partial_x\varphi)^2=0 \quad \text{on } \{\varphi=0\}.$$

Using (c) above, we now obtain

$$Wv=(\mathscr{T}^0-2\mathscr{E}_1)\delta'(\varphi)+\mathscr{R}^0\delta(\varphi).$$

Now we apply the same reasoning as above to conclude the existence of a $\mathscr{T}_1^0\in C^\infty(\mathbb{R}^2)$ with $\mathscr{T}^0-2\mathscr{E}_1=\mathscr{T}_1^0\varphi$, in particular,

(t) $$\mathscr{T}^0-2\mathscr{E}_1=0 \quad \text{on } \{\varphi=0\}.$$

Thus, provided (e) and (t) are satisfied,

$$Wv=(\mathscr{R}^0-\mathscr{T}_1^0)\delta(\varphi),$$

and we have achieved continuity of Wv in $C_0^0(\mathbb{R}^2)$.

We now discuss the conditions (e) and (t) in some detail. To begin with, it is customary to consider the stronger condition

($\bar{\text{e}}$) $$\rho(\partial_t\varphi)^2-E(\partial_x\varphi)^2\equiv 0,$$

called the eikonal equation, rather than (e). Clearly, a solution of the first-order partial differential equation ($\bar{\text{e}}$) provides a function φ which satisfies (e). In fact, no essential generality is lost by replacing (e) with ($\bar{\text{e}}$) (although this point is hardly obvious).

The eikonal equation ($\bar{\text{e}}$) is easily solved: along any level curve $\{\varphi=c\}$ of a solution

$$\frac{dt}{dx}=\frac{\partial_x\varphi}{\partial_t\varphi}=\pm\rho^{1/2}E^{-1/2}.$$

So the level curves of a solution of ($\bar{\text{e}}$) passing through (x_0, t_0) are described by the relation

$$\begin{aligned}t &= \pm\int_{x_0}^{x}\rho^{1/2}E^{-1/2}+t_0\\ &= \pm T(x, x_0)+t_0,\end{aligned}$$

where

$$T(x, x_0)=\int_{x_0}^{x}c^{-1}$$

is the *travel-time function* and

$$c=\rho^{-1/2}E^{1/2}$$

is the local *sound speed*. The function φ is determined, therefore, if its values are given along any curve intersecting the level curves transversally, e.g., along the curves $\{t = 0\}$.

If we wish the curve $\{\varphi = 0\}$ to pass through $(x_0, 0)$, then a convenient choice is

$$\varphi^{\pm}(x, t; x_0) = T(x, x_0) \mp t$$

Of course, many other choices are possible. For ($\bar{e}$), we may replace $\varphi^{\pm}(x, 0, x_0) = T(x, x_0)$ by any other function vanishing at x_0, with nonzero x-derivative (to ensure that $d\varphi \neq 0$ near $\varphi = 0$). On the other hand, any multiple of the above choices by a nonvanishing smooth function satisfies (e).

We note that a vector field tangent to the level curve $\{\varphi^{\pm} = c\}$ is

$$D^{\pm} = \partial_x \pm c^{-1}\partial_t.$$

Of course, any multiple of this vector field by a nonvanishing function also gives a vector field tangent to $\{\varphi^{\pm} = c\}$, but the above choice is convenient. A parameterization of the curve $\{\varphi^{\pm} = 0\}$ is

$$x \to (x, \pm T(x, x_0)).$$

The level curves $\{\varphi = c\}$ are called (*characteristic*) *rays* for W.

Consider next the condition (t): since we have solved ($\bar{e}$) rather than (e), the function $\mathscr{E}_1$ vanishes identically, and (t) becomes

($\bar{t}$) $$\mathscr{T} = 0 \quad \text{on } \{\varphi = 0\}.$$

For our choices $\varphi^{\pm}$ of *phase function* (as φ is called, terminology again from geometric optics)

$$\mathscr{T}^{\pm} = \mp 2\rho\partial_t g^{\pm} - 2Ec^{-1}\partial_x g^{\pm} + g^{\pm}(-\partial_x Ec^{-1}),$$

and this quantity is required to vanish along the ray $\{\varphi^{\pm} = 0\}$. Multiply the right-hand side by $-\frac{1}{2}E^{-1}c$ to obtain

$$\begin{aligned} 0 &= \pm c^{-1}\partial_t g^{\pm} + \partial_x g^{\pm} + -\tfrac{1}{2} g^{\pm}\partial_x \log \eta \\ &= D^{\pm} g^{\pm} + g^{\pm}\partial_x \log \eta^{1/2} \equiv \mathscr{T}^{\pm} g^{\pm}, \end{aligned}$$

where

$$\eta = \rho^{1/2}E^{1/2} = Ec^{-1}$$

is the *acoustical impedance* and $\mathscr{T}^{\pm}$ are the *transport operators*. Note that condition ($\bar{t}$) has now taken the form of a linear scalar first-order ordinary differential equation along the rays $\{\varphi^{\pm} = 0\}$, and is easily solved:

$$g^{\pm}(x, \pm T(x, x_0)) = g^{\pm}(x_0, 0)(\eta(x)^{-1}\eta(x))^{1/2}.$$

Note also that the initial values $g^{\pm}(x, 0)$ are completely immaterial for $x \neq x_0$. We shall take advantage of this situation by constructing a family of solutions of ($\bar{e}$), ($\bar{t}$) parameterized by x_0: define

$$\varphi^{\pm}(x, t; x_0) = T(x, x_0) \mp t,$$

$$g^{\pm}(x, \pm T(x, x_0); x_0) = g^{\pm}(x_0, 0; x_0)(\eta(x)^{-1}\eta(x_0))^{1/2}.$$

Now we superimpose the two compound distributions $g^{\pm}\delta(\varphi^{\pm})$ to construct a distribution

$$u_s(x, t; x_0) = g^{+}(x, t; x_0)\delta(\varphi^{+}(x, t; x_0)) + g^{-}(x, t; x_0)\delta(\varphi^{-}(x, t; x_0)),$$

which satisfies the initial conditions

$$u_s(x, 0; x_0) = \delta(x - x_0), \qquad \partial_t u_s(x, 0; x_0) = 0.$$

That this is possible rests on the following observation:

LEMMA 2. (restriction lemma). *Suppose that the hypersurface* $M = \{\psi = 0\} \subset \mathbb{R}^n$ *intersects* $\{\varphi = 0\}$ *transversally. Then there is a unique distribution* $\gamma_k \in \mathscr{D}'(M)$ *so that, for any sequence* $\{v_n\} \subset C^\infty(\mathbb{R}^n)$ *tending to* $\delta(\psi)$ *in the sense of* $\mathscr{D}'(\mathbb{R}^n)$, *and any* $u \in C_0^\infty(\mathbb{R}^n)$,

$$\langle \gamma_k, u|_M \rangle = \lim_{n \to \infty} \langle \delta^{(k)}(\varphi), v_n u \rangle.$$

γ_k is called the *restriction* of $\delta^{(k)}(\varphi)$ to $M = \{\psi = 0\}$. The name is justified by the observation that, if $\{\tilde{u}_n\}$ is any sequence in $C^\infty(\mathbb{R})$ approximating δ, then the ordinary functional restrictions $\tilde{u}_n^{(k)}(\varphi)|_M$ tend to γ_k in $\mathscr{D}'(M)$.

The proof is not difficult, and we omit it. In the course of the proof, one observes that, if coordinates are chosen so that $\psi = y_1$, so that $y_2 \cdots y_n$ are coordinates on M, then

$$\gamma_k = \delta^{(k)}(\varphi(0, y_2 \cdots y_n)), \qquad k = 0, 1, \cdots.$$

In the application, $\psi(x, t) = t$, M is the x-axis and the restrictions of u_s and $\partial_t u_s$ to $\{t = 0\}$ become

$$\begin{aligned} u_s(x, 0; x_0) &= \{g^+(x, 0; x_0\} + g^-(x, 0; x_0)\}\delta(T(x, x_0)), \\ \partial_t u_s(x, 0; x_0) &= \{\partial_t g^+(x, 0; x_0) + \partial_t g(x, 0; x_0)\}\delta(T(x, x_0)) \\ &\quad + \{-g^+(x, 0; x_0\} + g^-(x, 0; x_0\}\delta'(T(x, x_0)). \end{aligned}$$

According to the change of variable formula for the delta function (or see [15, pp. 236–7]), since $T(x, x_0) = (x - x_0)c^{-1}(x_0) + O(T(x, x_0)^2)$, we have

$$\delta(T(x, x_0)) = c(x_0)\delta(x - x_0),$$

and

$$\delta'(T(x, x_0)) = c^2(x_0)\delta'(x - x_0) + \tfrac{1}{2}\left(\frac{d}{dx}c^2\right)\Big|_{x = x_0} \delta(x - x_0).$$

The initial conditions to be imposed therefore amount to the following restrictions on $g^\pm$:

(i) $g^+(x_0, 0; x_0) + g^-(x_0, 0; x_0) = c^{-1}(x_0)$,

(ii) $g^+(x_0, 0; x_0) = g^-(x_0, 0; x_0)$,

(iii) $c(x_0)\{\partial_t g^+(x_0, 0; x_0) + \partial_t g^-(x_0, 0; x_0)\}$
$\quad - c^2(x_0)[\partial_x(-g^+(x, 0; x_0) + g^-(x, 0; x_0))|_{x = x_0}] = 0$

(where, in deriving the second and third conditions, we have used the recurrence relation (c) and the extension lemma).

To analyze (iii), multiply by $c^{-2}(x_0)$ and use the definition of the vector fields $D^\pm$ to obtain the equivalent condition

$$(D^+ g^+)(x_0, 0; x_0) = D^- g^-(x_0, 0; x_0).$$

Because of the transport equation $\mathscr{T}^{\pm}g^{\pm}=0$, this is equivalent to

$$g^{+}(x_0, 0; x_0)(\partial_x \log \eta^{1/2})(x_0) = g^{-}(x_0, 0; x_0)(\partial_x \log \eta^{1/2})(x_0),$$

which is implied by (ii). Therefore, condition (iii) is redundant.

Conditions (i) and (ii) together determine

$$g^{\pm}(x_0, 0; x_0) = \tfrac{1}{2}c(x_0)^{-1},$$

and therefore

$$g^{\pm}(x, \pm T(x, x_0); x_0) = \tfrac{1}{2}c(x_0)^{-1}(\eta(x)^{-1}\eta(x_0))^{1/2}.$$

To recapitulate, u_s solves the wave equation approximately, in the sense that Wu_s is no more singular than u_s and u_s has the desired initial values. The next step is to add a correction term to u_s, to obtain a solution of the wave equation. The reader is cautioned that our treatment diverges somewhat from the usual, e.g., Courant-Hilbert [11], and is designed to treat coefficients with finite order of smoothness.

We seek a distribution K for which

$$W(u_s + K) = 0.$$

If the coefficients ρ, E are C^{∞}, solutions of the (distribution) initial value problem

$$WK = -Wu_s \quad \text{in } \{t>0\}\times\mathbb{R},$$

$$K = \partial_t K = 0 \quad \text{on } \{t=0\}$$

are unique. This result can be established by standard energy methods and duality arguments (see [16, p. 237 ff.] or [10, Chapt. 7]). A local version of this result can also be proved: distribution solutions of initial value problems for W are unique in domains bounded by space-like hypersurfaces, i.e., in domains of dependence. For smooth coefficients, it follows that K must identically be zero outside the light cone $\{|T(x, x_0)| \leqq t\}$.

The first step in constructing K is to examine the remainder term, $-Wu_s$. This distribution is computed as indicated, following the statement of condition (t). To simplify computations, we choose to require that the transport equations be solved globally, not just along the characteristic rays $\{\varphi^{\pm}=0\}$. A convenient choice of global solution is

$$g^{\pm}(x, t; x_0) = \tfrac{1}{2}c(x_0)(\eta(x)^{-1}\eta(x_0))^{1/2}. \tag{2.4}$$

Then

$$Wu_s = (Wg^{+})\delta(\varphi^{+}) + (Wg^{-})\delta(\varphi^{-}),$$

$$Wg^{\pm}(x, t; x_0) = -\tfrac{1}{2}c(x_0)\partial_x(E(x)\partial_x\eta(x)^{-1/2}\eta(x_0)^{1/2}).$$

Consider now the local problem

$$Wv = \mathscr{R}\delta(\varphi)$$

in some open region in $\mathbb{R}^2$, where φ is assumed to satisfy the eikonal equation ($\bar{e}$). We will construct a solution of the form

$$v = g_1 H(\varphi).$$

According to the chain rule, then

$$Wv = \mathscr{T}_1\delta(\varphi) + \mathscr{R}_1 H(\varphi),$$

where

$$\mathscr{T}_1 = 2\rho\partial_t\varphi\partial_t g_1 - 2E\partial_x\varphi\partial_x g_1 + g_1 W\varphi$$

and

$$\mathscr{R}_1 = Wg_1.$$

By now familiar arguments, $\mathscr{T}_1 - \mathscr{R}$ must vanish on $\{\varphi = 0\}$,

$$\mathscr{T}_1 - R = \mathscr{T}_1^1\varphi,$$

and $\mathscr{R}_1$ must vanish identically in the support of $H(\varphi)$. Thus, g_1 must solve the wave equation in $\{\varphi \geqq 0\}$, and its boundary values on $\{\varphi = 0\}$ must satisfy $\mathscr{T}_1 - \mathscr{R} = 0$.

This local construction suggests an *ansatz* for K:

$$K(x, t; x_0) = g_1^+(x, t; x_0)H(T(x, x_0) - t) + g_1^-(x, t; x_0)H(T(x, x_0) + t).$$

Exactly as above, we obtain that

$$(t_1) \qquad -2Ec^{-1}\mathscr{T}^{\pm}g_1^{\pm} = -Wg^{\pm}, \quad \text{on } T(x, x_0) = \pm t.$$

This ordinary differential equation is similar to the first transport equation (t), and is called the *second transport equation* (in the usual theory, a sequence of higher transport equations appears, whose solutions form the coefficients of an asymptotic series solution of the wave equation, see [11, Chapt. 6]). Its solution is

$$g_1^{\pm}(x, \pm T(x, x_0); x_0) = -\tfrac{1}{4}c(x_0)\eta(x_0)^{1/2}\eta(x)^{-1/2}\int_{x_0}^{x}\eta^{-1/2}(E(\eta^{-1/2})')' + g_1^{\pm}(x_0, 0; x_0)\eta^{1/2}(x_0)\eta^{-1/2}(x).$$

The two characteristic rays through $(x_0, 0)$ divide the upper halfplane $\{t \geqq 0\}$ into three open regions, which we shall number I, II and III from the left. In region I, $K \equiv 0$. In region II, $K = g^-$. In region III, $K = g^+ + g^-$. In all three regions, K solves the (homogeneous) wave equation. In region III, K must also vanish identically, which shows in particular that

$$K(x, T(x, x_0); x_0) = g_1^-(x, T(x, x_0); x) = -g_1^+(x, T(x, x_0); x).$$

To determine appropriate initial conditions for the second transport equation, we proceed as in the case of the first transport equation, by way of the restriction lemma. The conditions

$$K = \partial_t K = 0, \qquad t = 0$$

translate into

(i) $(g_1^+(x, 0; x_0) + g_1^-(x, 0; x_0))H(T(x, x_0)) = 0,$

(ii) $(\partial_t g_1^+(x, 0; x_0) + \partial_t g_1^-(x, 0; x_0))H(T(x, x_0)) + (g_1^-(x, 0; x_0) - g_1^+(x, 0; x_0))\delta(T(x, x_0)) = 0,$

which are clearly equivalent to

$$g_1^+ + g_1^- = \partial_t(g_1^+ + g_1^-) \equiv 0, \qquad x \geqq x_0,$$

and

$$g_1^- - g_1^+ \fallingdotseq 0, \qquad x = x_0,$$

which give the initial conditions

$$g_1^{\pm}(x_0, 0; x_0) = 0$$

for the second transport equation.

Now we collect the above computations into the following statement:

PROPOSITION 1. *If* $K \in L^{\infty}(\mathbb{R} \times (t \geqq 0\})$ *is identically zero in regions* I *and* III, *satisfies the wave equation in region* II (*the forward light cone*), *and on the boundary rays has a jump discontinuity given by*

$$K(x, \pm T(x, x_0); x_0)$$
$$= \pm \tfrac{1}{4} c(x_0)\eta(x_0)^{1/2}\eta(x)^{-1/2}\int_{x_0}^{x} \eta^{-1/2}(E(\eta^{-1/2})')', \qquad x \gtrless x_0,$$

then K *solves the problem*

$$WK = -Wu_s \quad \text{in } \{t \geqq 0\},$$
$$K = \partial_t K = 0, \qquad t = 0.$$

Moreover, K *is the unique distribution solution of this problem, by virtue of the existence theorem quoted above.*

This statement holds for smooth coefficients ρ, E and, in that case, K may also be constructed by the method of characteristics (see [11, p. 476 ff]), whence it follows that K is *smooth* inside the light cone.

To obtain a result which applies to coefficients with a finite degree of smoothness, we apply a theorem on existence and uniqueness for the Cauchy problem for W in the presence of measurable coefficients, which may be proved by the method described in Lions' book [22, pp. 272 ff.]:

PROPOSITION 2. *Suppose that* $\rho, E, \rho^{-1}, E^{-1}$ *are uniformly bounded on* $\mathbb{R}$ *and measurable. Suppose* $f \in H^k_{\mathrm{loc}}(\mathbb{R}, L^2(\mathbb{R}))$, $k \in \mathbb{Z}$, *with* $\operatorname{supp} f \subset \{t \geqq 0\}$. *Then there exists a unique distribution* $y \in H^k_{\mathrm{loc}}(\mathbb{R}, H^1(\mathbb{R}))$ *with* $\operatorname{supp} y \subset \{t \geqq 0\}$ *and* $\partial y_t \in H^k_{\mathrm{loc}}(\mathbb{R}, L^2(\mathbb{R}))$ *so that* $Wy = f$.

The calculation following (2.4) above show that Wu_s has the form $h^{+}\delta(\varphi^{+}) + h^{-}\delta(\varphi^{-})$ with $h^{\pm}$ bounded and measurable when ρ'', E'' are bounded and measurable. Reference to the definition of $h\delta(\varphi)$ shows that it makes sense provided that φ has a bounded, measurable differential and h is bounded and measurable. Also in that case, provided that $h^{\pm}$ depends on x alone, $h^{\pm}\delta(\varphi^{\pm}) = \pm(d/dt)h^{\pm}H(\varphi^{\pm})$ and $h^{\pm}H(\varphi^{\pm}) \in L^2_{\mathrm{loc}}(\mathbb{R}, L^2(\mathbb{R}))$, which shows that $h^{\pm}\delta(\varphi^{\pm}) \in H^{-1}_{\mathrm{loc}}(\mathbb{R}, L^2(\mathbb{R}))$. To summarize:

PROPOSITION 3. *Suppose* ρ'', E'' *are bounded and measurable. Then the problem*

$$WK = -Wu_s,$$
$$K \equiv 0 \quad \text{for } t \leqq 0$$

has a unique solution $K \in H^{-1}(\mathbb{R}, H^1(\mathbb{R}))$ *with* $\partial_t K \in H^{-1}(\mathbb{R}, L^2(\mathbb{R}))$.

The hypothesis $\rho'', E'' \in L^{\infty}$ is strong enough to allow the construction of K by the energy method as described in [10] or [11]. In fact, we can prove the following theorem concerning the characteristic Cauchy problem:

PROPOSITION 4. *Suppose* ρ'', E'' *are bounded and measurable. Then the characteristic boundary value problem*

$$Wv = 0, \qquad |T(x, x_0)| < t, \qquad t > 0,$$
$$v(x, |T(x, x_0)|) = f(x)$$

has a unique solution in $H^1_{\text{loc}}(\text{II})$, *provided that* $f \in H^1(\mathbb{R})$. *Moreover,* v *has finite energy*:

$$\|v(\cdot, t)\|_E \equiv \int_{T(x, x_0)=-t}^{T(x, x_0)=t} dx c^{-2}(\partial_t v)^2 + (\partial_x v)^2$$
$$\leqq \|f\|_{H^1(\mathbb{R})},$$

and so also

$$v \in L^\infty(\{t \geqq 0\}; C^0(\mathbb{R})),$$

i.e., v *is continuous on the lines* $\{t = \text{const.}\}$

To apply this theorem to the construction of K, we note that the boundary values of K are in H^1 when the second derivatives of ρ and E are bounded and measurable. Note also that, when the second derivatives of ρ and E are bounded and measurable, the phase functions $\varphi^\pm$ have three bounded measurable derivatives. On the basis of the theory presented in [15], the statements (1)–(4) made earlier concerning the distributions $\delta^{(k)}(\varphi^\pm)$, $k \leqq 2$, are valid in this case, and therefore the derivations of the conditions (e), (t), and (t_1) proceed exactly as in the smooth case. (This result is somewhat nontrivial, and has to do with the fact that $\varphi^\pm$ are actually smooth in a direction transverse to their level surfaces.)

The foregoing constitutes the proof of the following result, which is the culmination of this section:

THEOREM A. *Suppose that* ρ, E *are positive functions and that* ρ'', E'' *are bounded and measurable. Then the singular initial value problem*

$$W\bar{u} = 0, \quad \bar{u}(x, 0) = \delta(x - x_0), \quad \partial_t \bar{u}(x, 0) = 0$$

has a unique solution in $\mathcal{D}'(\mathbb{R}^2)$,

$$\bar{u}(x, t) = \tfrac{1}{2} c(x_0)^{-1} \eta(x_0)^{1/2} \eta(x)^{-1/2} \{\delta(T(x, x_0) - t) + \delta(T(x, x_0) + t)\}$$
$$+ K(x, t, x_0),$$

where $K \equiv 0$ *if* $|T(x, x_0)| > |t|$, *and* K *is the unique* H^1*-solution of* $WK = 0$ *in the interior* $\{|T(x, x_0)| < |t|\}$ *of the light cone with boundary values*

$$K(x, \pm T(x, x_0); x_0)$$
$$= \pm \tfrac{1}{4} c(x_0) \eta(x_0)^{1/2} \eta(x)^{1/2} \int_{x_0}^{x} \eta^{-1/2} (E(\eta^{-1/2})')'.$$

Also, K *is continuous on the intersections of the lines* $\{t = \text{const.}\}$ *with the closed light cone, and the* sup *norms along these segments are locally bounded in* t.

We mention another regularity result needed later, proved via the method of characteristics (a device peculiar to $1+1$ dimensions; the other results are proved by energy methods, hence generalize to higher-dimensional problems):

PROPOSITION 5. *If* ρ, E *are of class* C^{k+1}, $k \geqq 1$, *then* K *is of class* C^k *in the interior of the light cone.*

Finally, by setting $E \equiv 1$, so that $\rho = c^{-2}$, $\eta = c^{-1}$, we obtain the expressions (2.2), (2.3). These are valid when ρ'' is bounded and measurable and $K \in C^k$ in the interior of the light cone when $\rho \in C^{k+1}$.

3. The GL equation. We assume until further notice that $c \in C^2$. Since any function may be written as a superposition of delta functions, we can write the general

solution $u(x, t)$ of the partial differential equation

$$\Box_c u = 0,$$

satisfying the initial condition

$$\frac{\partial u}{\partial t}(x, 0) \equiv 0,$$

as

$$u(x, t) = \int_{-\infty}^{\infty} dx_0 S(x, t; x_0, 0) u(x_0, 0),$$

where

$$S(x, t; x_0, 0) = \bar{u}(x, t),$$

is the solution examined in the previous section; that is,

$$\Box_c\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}\right) S(x, t; x_0, 0) \equiv 0,$$

$$S(x, 0; x_0, 0) = \delta(x - x_0),$$

$$D_2 S(x, 0; x_0, 0) \equiv 0.$$

S is related to the *Reimann function* R of $\Box_c$, which solves the initial value problem

$$\Box_c\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}\right) R(x, t; x_0, 0) \equiv 0,$$

$$R(x, 0; x_0, 0) \equiv 0,$$

$$D_2 R(x, 0; x_0, 0) = \delta(x - x_0)$$

(see [11, Chapt. V, § 5]) by

$$R(x, t; x_0, 0) = \int_0^t ds\, S(x, s; x_0, 0).$$

Define

$$\begin{aligned} R(x, t; x_0, t_0) &= R(x, t - t_0; x_0, 0), \\ S(x, t; x_0, t_0) &= S(x, t - t_0; x_0, 0). \end{aligned} \tag{3.1}$$

Then the general solution of $\Box_c = 0$ with arbitrary initial data is given by

$$u(x, t) = \int_{-\infty}^{\infty} dx_0 \left\{ S(x, t; x_0, t_0) u(x_0, t_0) + R(x, t; x_0, t_0) \frac{\partial}{\partial t_0} u(x_0, t_0) \right\}.$$

This works because the coefficient of $\Box_c$ is independent of t. For the same reason, the initial value vector $(u(\cdot, t_0); \partial u/\partial t(\cdot, t_0))$ is propagated in t by a group of operators (bounded, in fact, in a suitable function space). The distribution kernel of this group of solution operators is the matrix

$$\mathscr{R} = \begin{pmatrix} S & R \\ D_2 S & D_2 R \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} u(\cdot, t) \\ \frac{\partial u}{\partial t}(\cdot, t) \end{pmatrix} = U(t-t_0)\begin{pmatrix} u(\cdot, t_0) \\ \frac{\partial u}{\partial t}(\cdot, t_0) \end{pmatrix}$$

$$= \begin{pmatrix} \int dx_0 \left\{ S(x, t; x_0, t_0)u(x_0, t_0) + R(x, t; x_0, t_0)\frac{\partial u}{\partial t}(x_0, t_0) \right\} \\ \int dx_0 \left\{ D_2 S(x, t; x_0, t_0)u(x_0, t_0) + D_2 R(x, t; x_0, t_0)\frac{\partial u}{\partial t}(x_0, t_0) \right\} \end{pmatrix}.$$

The operators $U(t)$ are bounded in a suitable sense, and form a *one-parameter group*:

$$U(s)U(t) = U(s+t).$$

In terms of the matrix kernel $\mathcal{R}$, this *group law* reads

$$\mathcal{R}(x, t; x_0, t_0) = \int_{-\infty}^{\infty} dy\, \mathcal{R}(x, t; y, t')\mathcal{R}(y, t'; x_0, t_0). \tag{3.2}$$

This relation is fundamental for what follows. Since the components of $\mathcal{R}$ may all be expressed in terms of the scalar kernel S, it is plausible that (3.2) may be expressed in terms of a property of S. This is indeed the case.

To derive this relation, we require some further symmetries of S, which are as follows:

1 S is *even* in $t-t_0$, whereas R is *odd* in $t-t_0$.

2 The kernel

$$S^*(x, t; x_0, t_0) = S(x_0, t_0; s, t)$$

is a solution of the *adjoint equation*

$$\Box_c^* S^* = 0$$

(see [11, Chapt. V, § 5.3]).

Now

$$\Box_c^* = \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}c^2.$$

So (we suppress for the moment the dependence on x_0, t_0)

$$0 = \left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}c^2(x)\right)S^*(x, t) = \frac{\partial^2}{\partial t^2}S^*(x, t) - \frac{\partial^2}{\partial x^2}(c^2(x)S^*(x, t))$$

$$= \left(c^{-2}(x)\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}\right)(c^2(x)S^*(x, t)) = c^{-2}(x)\Box_c(c^2(x)S^*(s, t)).$$

Since $c > 0$, it follows that $c^2(x)S^*(x, t)$ solves

$$\Box_c c^2(x)S^*(x, t) = 0.$$

Now

$$c^2(x)S^*(x, t_0) = c^2(x)\delta(x-x_0) = c^2(x_0)\delta(x-x_0),$$

and

$$\frac{\partial}{\partial t} c^2(x) S^*(x, t)|_{t=t_0} = 0.$$

Therefore,

$$c^2(x) S^*(x, t) = c^2(x_0) S(x, t),$$

i.e.,

$$S(x_0, t_0; x, t) = c^2(x_0) c^{-2}(x) S(x, t; x_0, t_0),$$

which is the desired symmetry relation.

Now examine the (1, 1)-component of the group law equation (3.2), which reads

$$S(x, t; x_0, t_0) = \int_{-\infty}^{\infty} dy \, \{(S(x, t; y, t') S(y, t'; x_0, t_0) + R(x, t; y, t') D_2 S(y, t'; x_0, t_0)\}.$$

We set $x = x_0$, $t_0 = 0$ and replace t by $t+s$, t' by s to obtain

$$S(x_0, t+s; x_0, 0) = \int_{-\infty}^{\infty} dy \, \{S(x_0, t+s; y, s) S(y, s; x_0, 0) + R(x_0, t+s; y, s) D_2 S(y, s; x_0, 0)\}.$$

Using the time-translation symmetry (2.1), rewrite the right-hand side as

$$= \int_{-\infty}^{\infty} dy \, \{S(x_0, t; y, 0) S(y, s; x_0, 0) + R(x_0, t; y, 0) D_2 S(y, s; x_0, 0)\}.$$

Now note that, since $S(y, s; x_0, 0)$ is *even* in s, its s-derivative $D_2 S(y, s; x_0, 0)$ is odd. The second term in the integrand is therefore odd in s, whereas the first is even. Replace s by $-s$, add, and divide by two to obtain

$$\tfrac{1}{2}[S(x_0, t+s; x_0, 0) + S(x_0, t-s; x_0, 0)] = \int_{-\infty}^{\infty} dy \, S(x_0, t; y, 0) S(y, s; x_0, 0).$$

Now use the adjoint symmetry (2) above to interchange the arguments in the first factor in the integrand:

$$\begin{aligned} &= c^2(x_0) \int_{-\infty}^{\infty} dy \, c^{-2}(y) S(y, 0; x_0, t) S(y, s; x_0, 0) \\ &= c^2(x_0) \int_{-\infty}^{\infty} dy \, c^{-2}(y) S(y, -t; x_0, 0) S(y, s; x_0, 0) \qquad (3.3) \\ &= c^2(x_0) \int_{-\infty}^{\infty} dy \, c^{-2}(y) S(y, t; x_0, 0) S(y, s; x_0, 0). \end{aligned}$$

Conversely, one can show that (3.3) entails the group law equation for the full Riemann matrix $\mathcal{R}$.

Now set

$$F(t) = S(x_0, t; x_0, 0), \qquad t \neq 0.$$

According to the transport equation (2.3), $K(x_0, 0; x_0) = 0$. Since K is continuous in the

closed light cone, and $F(t) = K(x_0, t; x_0)$ for $t \neq 0$, one can set

$$F(0) = 0,$$

to obtain a continuous, even function for all t. Define also

$$G(s, t) = \tfrac{1}{2}[F(t+s) + F(t-s)].$$

Now

$$S(x_0, t; x_0, 0) = \delta(t) + F(t),$$

so the left-hand side of (3.3) is

$$\tfrac{1}{2}[S(x_0, t+s; x_0, 0) + S(x_0, t-s; x_0, 0)] = \tfrac{1}{2}[\delta(s+t) + \delta(s-t)] + G(s, t).$$

Recall also from § 2 the expansion

$$S(x, t; x_0, 0) = \tfrac{1}{2}c^{1/2}(x)c^{-3/2}(x_0)[\delta(t + T(x, x_0)) + \delta(t - T(x, x_0))] + K(x, t; x_0).$$

Since x_0 will remain fixed for the rest of this discussion, we shall henceforth set $x_0 = 0$. The number $c(x_0) = c(0)$ is now surely a positive constant which can be set equal to 1 by *scaling* t, which we assume has been done. We write

$$T(x) = T(x, 0) = \int_0^x c^{-1}.$$

We now have in place of (3.3)

$$\tfrac{1}{2}[\delta(s+t) + \delta(s-t)] + G(x, t) = \int_{-\infty}^{\infty} dy\, c^{-2}(y) S(y, t; 0, 0) S(y, s; 0, 0), \tag{3.4}$$

with

$$S(x, t; 0, 0) = \tfrac{1}{2}c^{1/2}(x)[\delta(t + T(x)) + \delta(t - T(x))] + K(x, t). \tag{3.5}$$

One now substitutes (3.5) into (3.4), and after some computation eliminates the singular terms to obtain

$$\begin{aligned} G(s, t) = {} & \tfrac{1}{2}[c^{-1/2}(x^-(s))K(X^-(s), t) + c^{-1/2}(X^+(s))K(X^+(s), t)] \\ & + \int_{X^-(s)}^{X^+(s)} dy\, c^{-2}(y) K(y, t) K(y, s). \end{aligned} \tag{3.6}$$

Here, $s \to X^{\pm}(s) = x$ is the inverse function to $x \to \pm T(x) = s$, and is thus the solution of

$$\frac{d}{ds} X^{\pm} = \pm c(X^{\pm}), \qquad X^{\pm}(0) = 0.$$

The light cone through $(0, 0)$ is thus described by $\{(x, t): X^-(t) \leqq x \leqq X^+(t)\}$, and $t \to (X^{\pm}(t), t)$, $x \to (x, \pm T(x))$ are equivalent descriptions of the characteristic curves emanating from $(0, 0)$. Recall that (§ 2, (2.3))

$$\begin{aligned} K(x, T(x)) &= \tfrac{1}{4}c^{1/2}(x) \int_0^x c^{1/2}(c^{1/2})'', \qquad x \geqq 0, \\ K(x, -T(x)) &= -\tfrac{1}{4}c^{1/2}(x) \int_0^x c^{1/2}(c^{1/2})'', \quad x \leqq 0. \end{aligned} \tag{3.7}$$

In view of the formulation of the inverse problem (§ 1), we now assume that

$$c(x) \equiv 1, \qquad x \leqq 0.$$

LEMMA A.

$$K(y, t) = F(t+y), \qquad y \leqq 0.$$

Proof. K, being the smooth part of a solution of the wave equation, must solve it in the interior of the light cone. For $x \leqq 0$, $T(x) = x$ and (3.7) show that

$$K(x, -x) = 0, \qquad x \leqq 0. \tag{3.8}$$

Finally,

$$K(0, t) = F(t). \tag{3.9}$$

A solution to the problem

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}\right) K = 0,$$

together with the conditions (3.8) and (3.9) in the region $\{-t \leqq x \leqq 0, t \geqq 0\}$, is

$$K(x, t) = F(x+t).$$

However, since one of the boundaries of this region is characteristic, the solution is unique (see [11, Chapt. V]). Q.E.D.

Now $X^-(s) = -s$, and set $X(s) = X^+(s)$. Then (3.6) reads, for $0 \leqq s \leqq t$,

$$\begin{aligned} G(s, t) &= \tfrac{1}{2} c^{-1/2}(X(s)) K(X(s), t) + \tfrac{1}{2} K(-s, t) + \int_{-s}^{X(s)} dy\, c^{-2}(y) K(y, s) K(y, t), \\ &= \tfrac{1}{2} F(t-s) + \int_{-s}^{0} dy\, F(y+t) F(y+s) + \tfrac{1}{2} c^{-1/2}(X(s)) K(X(s), t) \\ &\quad + \int_{0}^{X(s)} dy\, c^{-2}(y) K(y, s) K(y, t). \end{aligned}$$

Now set

$$H(s, t) = \tfrac{1}{2} F(t+s) - \int_0^s d\tau\, F(t-\tau) F(s-\tau). \tag{3.10c}$$

We have proved:

THEOREM C. *The smooth part K of the solution u of the singular initial value problem* (2.1) *satisfies, for* $0 \leqq s \leqq t$,

$$H(s, t) = \tfrac{1}{2} c^{-1/2}(X(s)) K(X(s), t) + \int_0^{X(s)} dy\, c^{-2}(y) K(y, s) K(y, t), \tag{3.10a}$$

and is related to the coefficient c by the transport equation

$$K(x, T(x)) = \tfrac{1}{4} c^{1/2}(x) \int_0^x c^{1/2} (c^{1/2})''. \tag{3.10b}$$

We shall refer to the system of integral equation (3.10) as the GL system, partly because it expresses the group law for the Cauchy problem for $\Box_c$, and partly because it is related to an equation discovered by Gel'fand and Levitan and derived in a completely different way in their fundamental paper [13].

The converse of Theorem C will be proved in § 7.

4. A priori estimates. In this section, we derive some necessary conditions for scattering data. We shall make use of the substitution $x \mapsto s = T(x)$ throughout. We point out that this is not the same as reduction to the case of the Schrödinger equation. Indeed, in the first part of this section, leading up to the estimate (4.11), the kernel $\tilde{K}$ can be replaced by the kernel J, defined in the next section. The operator $\mathbb{K}$ can be replaced by an operator $\mathbb{J}: L^2(dt) \to L^2(c^{-1}(x)dx)$, and all mention of the coordinate transformation $x \mapsto s$ can be eliminated. In higher-dimensional problems, the volume element of the Riemannian metric associated with the relevant hyperbolic p.d.e. will play the role of $c^{-1}(x)\,dx$, so our methods conform to the rubric laid down in § 1. Also, the last part of the section, leading up to the bounds (4.16), depends only on $x \mapsto s$ as arc-length parameterization of the geodesics of the above-mentioned metric, so this again an admissible trick.

Note that the kernel H is symmetric. It follows that (3.10a) also holds with s and t interchanged for $0 \leqq t \leqq s$.

Now define

$$\tilde{K}(s, t) = 2c^{-1/2}(X(s))K(X(s), t) \quad \text{for } 0 \leqq s \leqq t,$$
$$\tilde{K}(s, t) = 0 \quad \text{for } s > t \geqq 0.$$

Since $c^{-1}(y)dy = ds$ with $s = T(y)$, i.e., $y = X(s)$, we can rewrite (3.10) as

$$4H(s, t) = \tilde{K}(s, t) + \int_0^s d\tau\, \tilde{K}(\tau, s)\tilde{K}(\tau, t). \tag{4.2}$$

For $T > 0$, denote by $\mathbb{K}$ the Volterra operator on $L^2[0, T]$ defined by

$$\mathbb{K}\varphi(\tau) = \int_\tau^T dt\, \tilde{K}(\tau, t)\varphi(t)$$

for $\varphi \in L^2[0, T]$. Denote by $\mathbb{H}$ the symmetric operator with kernel H. The hypotheses on c are sufficient to ensure that $\mathbb{H}$ is a Hilbert-Schmidt operator on $L^2[0, T]$. Denote finally by $\mathbb{I}$ the identity operator on $L^2[0, T]$ (with kernel $\delta(\tau - t)$). Then (4.2) can be written

$$\mathbb{I} + 4\mathbb{H} = (\mathbb{I} + \mathbb{K})^\dagger(\mathbb{I} + \mathbb{K})$$

(see [6, eq. 4.10], also [14, eq. 8.1, 2]). This shows that $\mathbb{I} + 4\mathbb{H}$ must be positive definite (since $\mathbb{I} + \mathbb{K}$ is invertible). According to the Fredholm character of $\mathbb{I} + 4\mathbb{H}$, we must in fact have

$$\mathbb{I} + 4\mathbb{H} \geqq \varepsilon(T) > 0. \tag{4.3}$$

Obviously, $\varepsilon(T)$ is a monotone nonincreasing function of T. We note for later use the identity

$$\varepsilon(T)^{-1} = \sup\{\|(\mathbb{I} + \mathbb{K})^{-1}(\varphi)\|^2_{L^2[0,T]} : \|\varphi\|^2_{L^2[0,T]} = 1\}, \tag{4.4}$$

which follows immediately from (4.3).

Next, we extract from (4.3) some a priori estimates which will be crucial for the next section. Denote by H_t, $\tilde{K}_t$ the functions

$$H_t(s) = H(s, t), \qquad \tilde{K}_t(s) = \tilde{K}(s, t),$$

and note that (4.2) may be written

$$4H_t(s) = ((\mathbb{I} + \mathbb{K})^+ \tilde{K}_t)(s), \qquad 0 \leqq s \leqq t.$$

It follows that

$$16\|H_t\|^2_{L^2[0,T]} \geqq 16\|H_t\|^2_{L^2[0,t]} = \langle \tilde{K}_t, (\mathbb{I}+\mathbb{K})(\mathbb{I}+\mathbb{K})^\dagger \tilde{K}_t \rangle_{L^2[0,t]}. \tag{4.5}$$

Now the invertible self-adjoint operator $(\mathbb{I}+\mathbb{K})(\mathbb{I}+\mathbb{K})^\dagger$ has the same spectrum as the operator $(\mathbb{I}+\mathbb{K})^\dagger(\mathbb{I}+\mathbb{K}) = \mathbb{I}+4\mathbb{H}$, hence, in particular, the same lower bound. Therefore, one may combine (4.3) and (4.5) to obtain

$$16\|H_t\|^2_{L^2[0,T]} \geqq \varepsilon(T)\|\tilde{K}_T\|^2_{L^2[0,T]}. \tag{4.6}$$

Now $\|H_t\|$ may be estimated in the following truly crude fashion:

$$\|H_t\|^2_{L^2[0,T]} \leqq 2(\|F\|^2_{L^2[0,2T]} + T^2\|F\|^4_{L^2[0,T]}). \tag{4.7}$$

Also,

$$\left| \int_0^s d\tau\, \tilde{K}(\tau, s)\tilde{K}(\tau, t) \right| = |\langle \tilde{K}_s, \tilde{K}_t \rangle|_{L^2[0,t]} \leqq \|\tilde{K}_s\|_{L^2[0,s]}\|\tilde{K}_t\|_{L^2[0,t]}.$$

Hence, recalling (4.6) and using (4.7) gives

$$\left| \int_0^{X(s)} dy\, c^{-2}(y)K(y, s)K(y, t) \right| \leqq 8\varepsilon(T)^{-1}(\|F\|^2_{L^2[0,2T]} + T^2\|F\|^4_{L^2[0,T]}). \tag{4.8}$$

Combined with (3.10a), this yields the estimate

$$\begin{aligned} |\tfrac{1}{2}c^{-1/2}(X(s))K(X(s), t)| \leqq{}& \|F\|_{L^\infty[0,2T]} + \|F\|^2_{L^2[0,T]} \\ &+ 8\varepsilon(T)^{-1}[\|F\|^2_{L^2[0,2T]} + T^2\|F\|^4_{L^2[0,T]}], \end{aligned} \tag{4.9}$$

valid in the range $0 \leqq s \leqq t \leqq T$.

Finally, we give a priori bounds for c, in terms of T and $K^* = \sup_{0\leqq t\leqq T}|2c^{-1/2}(x(t))K(x(t), t)|$. Note that the latter quantity has just been estimated in terms of F.

According to (3.10b), for $0 \leqq x \leqq X(T)$,

$$\begin{aligned} 4c^{-1/2}(x)K(x, T(x)) &= \int_0^x c^{1/2}(c^{1/2})'' \\ &= \tfrac{1}{2}c'(x) - \tfrac{1}{4}\int_0^x c^{-1}(c')^2. \end{aligned} \tag{4.10}$$

Set

$$g(x) = 8c^{-1/2}(x)K(x, T(x)).$$

Then

$$c(x)^{-1}g(x) = (\log c)'(x) - \tfrac{1}{2}c^{-1}(x)\int_0^x c^{-1}(c')^2.$$

Since $c(0) = 1$, obtain

$$\log c(x) \geqq \int_0^x c^{-1}g \geqq -4K^* \int_0^x c^{-1} \geqq -4K^*T,$$

i.e.,

$$c \geqq \exp(-4K^*T), \tag{4.11}$$

which is the required lower bound for c.

To obtain upper bounds for c, write

$$\tfrac{1}{4}g(X(s)) = \bar{K}(s) = \tilde{K}(s, s), \qquad \varphi = c^{-1/2}, \qquad = \frac{d}{ds}.$$

Then (4.10) may be rewritten

$$2\bar{K}(t) = (\log \varphi)\quad (t) - \int_0^t ((\log \varphi)^{\cdot})^2.$$

It follows that φ obeys the equation

$$\ddot{\varphi} = Q\varphi,$$

where $Q = 2\dot{K}$. (This relation is also part of the Liouville reduction which figures in other treatments of this inverse problem; see, e.g., [2].) Note that

$$\varphi(0) = 1, \qquad \dot{\varphi}(0) = 0, \qquad K(0) = 0.$$

Hence, φ is the solution of

$$\begin{aligned} \varphi(t) &= 1 + \int_0^t ds\ (t-s)\varphi(s)Q(s) \\ &= 1 + 2\int_0^t ds\ \varphi(s)\bar{K}(s) - 2\int_0^t ds\ (t-s)\dot{\varphi}(s)\bar{K}(s). \end{aligned} \tag{4.12a}$$

Also, $\dot{\varphi}$ is the solution of

$$\dot{\varphi}(t) = 2\varphi(t)\bar{K}(t) - 2\int_0^t ds\ \dot{\varphi}(s)\bar{K}(s). \tag{4.12b}$$

Define $\psi = \dot{\varphi} - 2\bar{K}\varphi$. Then the Volterra system (4.12) may be rewritten as

$$\varphi(t) = 1 + \int_0^t ds\ 2\bar{K}(s)(1 + 2(s-t)K(s))\varphi(s) + \int_0^t ds\ 2\bar{K}(s)(s-t)\psi(s), \tag{4.13a}$$

$$\psi(t) = -4\int_0^t ds\ (\bar{K}(s))^2\varphi(s) - 2\int_0^t ds\ \bar{K}(s)\psi(s). \tag{4.13b}$$

Set

$$K^{**} = \max\{4K^* + 8T(K^*)^2, 8(K^*)^2, 4TK^*, 4K^*\}.$$

Then the following estimate is easily derived for the solution of (4.13):

$$\|\varphi\|_{L^\infty[0,T]} \leqq \exp TK^{**}. \tag{4.14a}$$

Also,

$$\|\psi\|_{L^\infty[0,T]} \leqq \exp TK^{**}. \tag{4.14b}$$

In view of the definition of ψ, this entails

$$\|\dot{\varphi}\|_{L^\infty[0,T]} \leqq (1 + 2K^*) \exp TK^{**}. \tag{4.15}$$

Now $\dot{\varphi} = (d/ds)(c^{-1/2}) = c(d/dx)(c^{1/2}) = \frac{1}{2}c^{1/2}c'$. Thus, combining (4.15), (4.14a) and (4.11), we get the estimates, valid for $0 \leqq x \leqq X(T)$,

$$\exp(-4K^*T) \leqq |c(x)| \leqq \exp(2TK^{**}), \tag{4.16a}$$

and

$$|c'(x)| \leqq 2(1+2K^*)\exp T(K^{**}+4K^*). \tag{4.16b}$$

5. Solution of the GL system. With this section, we begin the solution of the inverse problem as stated in § 1. The first step is to show that the GL system as presented in Theorem C,

$$H(s,t) = \tfrac{1}{2}c^{-1/2}(X(s))K(X(s),t) + \int_0^{X(s)} dy\, c^{-2}(y)K(y,s)K(y,t), \tag{5.1a}$$

$$0 \leqq s \leqq t \leqq T,$$

$$K(x, T(x)) = \tfrac{1}{4}c^{1/2}(x)\int_0^x c^{1/2}(c^{1/2})'', \tag{5.1b}$$

has a unique solution $\{K, c\}$, where c is defined on $0 \leqq x \leqq X(T)$ and K is defined in $C_T = \{(x,t)\colon 0 \leqq x \leqq X(T),\ T(x) \leqq t \leqq 2T - T(x)\}$. Since the domains on which the solutions are defined are themselves defined by part of the solution (namely c), the problem has something of the nature of a free boundary problem.

We note that continuous solutions are trivially unique, in view of the Volterra character of (5.1).

The system (5.1) will only have a solution as described when H has the positivity property

$$\mathbb{I} + 4\mathbb{H} \geqq \varepsilon(T) > 0, \tag{5.2}$$

in the notation of the last section, for the reasons explained there. Our goal is to show that this necessary condition is also sufficient.

First introduce the function

$$J(x,t) = 2c^{-1/2}(x)K(x,t),$$

and rewrite (5.1) as

$$4H(s,t) = J(X(s),t) + \int_0^{X(s)} dy\, c^{-1}(y)J(y,s)J(y,t), \tag{5.3a}$$

$$J(x, T(x)) = \tfrac{1}{2}\int_0^x c^{1/2}(c^{1/2})''. \tag{5.3b}$$

We have from (5.3b), (4.10)

$$J(x, T(x)) = \tfrac{1}{4}c'(x) - \tfrac{1}{8}\int_0^x c^{-1}(c')^2,$$

so

$$c(x) = 1 + 4\int_0^4 dy\, J(y, T(y)) + \tfrac{1}{2}\int_0^x dy \int_0^y c^{-1}(c')^2. \tag{5.4}$$

We shall suppose that (5.3) has been solved for $0 \leqq x \leqq x_0$. Set

$$\tilde{H}(x,t) = 4H(s,t) - \int_0^{x_0} dy\, c^{-1}(y)J(y,s)J(y,t).$$

Then, for $x \geqq x_0$, (5.3) may be rewritten

$$\tilde{H}(s, t) = J(X(s), t) + \int_{x_0}^{X(s)} dy\, c^{-1}(y) J(y, t), \tag{5.5a}$$

$$c(x) = k(x, x_0) + 4 \int_{x_0}^{x} dy\, J(y, T(y)) + \tfrac{1}{2} \int_{x_0}^{x} dy \int_{x_0}^{y} c^{-1}(c')^2, \tag{5.5b}$$

where

$$k(x, x_0) = c(x_0) + \tfrac{1}{2} \int_0^{x_0} dy \int_0^{y} c^{-1}(c')^2 + \tfrac{1}{2}(x - x_0) \int_0^{x_0} c^{-1}(c')^2 \tag{5.5c}$$

and (5.5a) is to be construed for $T(x_0) \leqq s \leqq t$.

Now define a sequence of approximate solutions by iterating the right-hand sides of (5.5): select $\Delta x > 0$ and define

$$c_0(x) \equiv c(x_0) \quad \text{for } x_0 \leqq x \leqq x_0 + \Delta x,$$
$$J_0(x, t) \equiv 0 \quad \text{for } x_0 \leqq x \leqq x_0 + \Delta x, \qquad t \geqq 0.$$

For $n \geqq 1$, define

$$c_n(x) = k(x, x_0) + 4 \int_{x_0}^{x} dy\, J_{n-1}(y, T_{n-1}(y)) + \tfrac{1}{2} \int_{x_0}^{x} dy \int_{x_0}^{y} c_{n-1}^{-1} (c'_{n-1})^2$$
$$\text{for } x_0 \leqq x \leqq x_0 + \Delta x.$$

Here

$$T_n(x) = \int_0^{x_0} c^{-1} + \int_{x_0}^{x} c_n^{-1} \quad \text{for } x_0 \leqq x \leqq x_0 + \Delta x.$$

Similarly,

$$J_n(x, t) = \tilde{H}(T_{n-1}(x), t) - \int_{x_0}^{x} dy\, c_{n-1}^{-1}(y) J_{n-1}(y, T_{n-1}(x)) J_{n-1}(y, t)$$
$$\text{for } x_0 \leqq x \leqq x_0 + \Delta x, \quad t \geqq 0.$$

Now suppose $\delta > 0$, and select c_* with $0 < c_* \leqq c(x_0) - \delta$. We claim that, for Δx small enough, we have $c_n(x) \geqq c_*$, $x_0 \leqq x \leqq x_0 + \Delta x$. In fact, suppose that this is so for c_k, $k = 0, 1, \cdots, n-1$ (it is obviously true for $n = 0$). As the first part of the induction, we estimate J_{n-1},

$$|J_{n-1}(x, t)| \leqq |\tilde{H}(T_{n-2}(x), t)| + \int_{x_0}^{x} dy\, c_*^{-1} |J_{n-2}(y, T_{n-2}(x))|\, |J_{n-2}(y, t)|. \tag{5.6}$$

Now

$$\tilde{H}(s, t) = 4\left[F(s + t) + \int_0^{g} d\tau\, F(s - \tau) F(t - \tau) \right] - \int_0^{x_0} dy\, c^{-1}(y) J(y, s) J(y, t).$$

So long as $T_{n-2}(x_0 + \Delta x) \leqq T$, which we assume for the moment, we have

$$\|H(s, t)\| \leqq \|F\|_{L^\infty[0,2T]} + \|F\|^2_{L^2[0,T]}, \qquad s = T_{n-2}(x).$$

The second summand on the r.h.s. is estimated by (4.10):

$$\left|\int_0^{x_0} dy\, c^{-1}(y)J(y,s)J(y,t)\right| \leqq 32\varepsilon(T)^{-1}[\|F\|^2_{L^2[0,2T]} + T^2\|F\|^4_{L^2[0,T]}].$$

So

$$\begin{aligned}|\tilde{H}(s,t)| &\leqq 4\|F\|_{L^\infty[0,2T]} + (4+32\varepsilon(T)^{-1})\|F\|_{L^2[0,2T]} + 32\varepsilon(T)^{-1}T^2\|F\|^4_{L^2[0,T]} \\ &\equiv N(F,T,\varepsilon).\end{aligned} \tag{57}$$

Combined with (5.6), this yields

$$\|J_{n-1}\|_\infty \leqq N(F,T,\varepsilon) + (x-x_0)c_*^{-1}\|J_{n-2}\|^2_\infty,$$

where $\|J\|_\infty$ means, for the moment, $\sup\{|J(x,t)|: x_0 \leqq x \leqq x_0+\Delta x,\ t \geqq 0\}$. Now suppose that

$$\|J_{n-2}\|_\infty \leqq (1+\delta')N(F,T,\varepsilon).$$

Then

$$\|J_{n-1}\|_\infty \leqq [1+(x-x_0)c_*^{-1}(1+\delta')N(F,T,\varepsilon)]N(F,T,\varepsilon).$$

Suppose that Δx is so small that

$$\Delta x c_*^{-1}(1+\delta')N(F,T,\varepsilon) \leqq \delta'. \tag{5.8}$$

Then we have once again that

$$\|J_{n-1}\|_\infty \leqq (1+\delta')N(F,T,\varepsilon). \tag{5.9}$$

To complete the induction, notice from (5.5c) that

$$k(x,x_0) \geqq c(x_0) \geqq c_* + \delta,$$

so that (from the definition of c_n)

$$c_n(x) \geqq c_* + \delta - 4\left|\int_{x_0}^x dy\, J_{n-1}(y, T_{n-1}(y))\right| \geqq c_* + \delta - 4(1+\delta')N(F,T,\varepsilon)\Delta x,$$

so we have proved:

LEMMA B. *Suppose* $\delta, \delta' > 0$, *and*

$$\Delta x \leqq \min\{c_*[(1+\delta')N(F,T,\varepsilon)]^{-1}\delta', [4(1+\delta')N(F,T,\varepsilon)]^{-1}\delta\}.$$

Then

$$c_n(x) \geqq c_*, \qquad n = 0, 1, 2, \cdots, \qquad x_0 \leqq x \leqq x_0 + \Delta x.$$

Note that $\delta' > 0$ is arbitrary here, whereas δ must be chosen so that $0 < c(x_0) - \delta$.

The next step begins with the assumption of a Lipschitz bound on F:

$$|F(t) - F(s)| \leqq L|s-t|. \tag{5.10}$$

It follows that

$$\begin{aligned}|H(\tau,t) - H(\tau,s)| &\leqq |F(\tau+t) - F(\tau+s)| + \left|\int_0^\tau d\sigma\, F(\tau-\sigma)(F(t-\sigma) - F(s-\sigma))\right| \\ &\leqq L|s-t|(1 + T\|F\|_{L^\infty[0,T]}) \equiv L_1|s-t|.\end{aligned} \tag{5.11}$$

Set $J_{s,t}(y) = J(y, s) - H(y, t)$. It follows from (5.5a) that

$$J_{s,t}(y) = 4[H(T(y), s) - H(T(y), t)] + \int_0^y d\tau\, c^{-1}(\tau) J(\tau, y) J_{s,t}(\tau).$$

This is a linear Volterra equation, whence follows the estimate (for $0 \leqq y \leqq \min(X(s), X(t))$

$$\begin{aligned} |J(y, s) - J(y, t)| &\leqq \|J_{s,t}\|_{L_\infty[0,x_0]} \leqq L_1(1 + T\|F\|_{L^\infty[0,T]}) \times \exp[x_0 c_*^{-1} \|J\|_\infty](|s-t|) \\ &= L_2|s-t|. \end{aligned} \tag{5.12}$$

Here, $\|J\|_\infty = \sup\{|J(x, t)|: 0 \leqq x \leqq x_0,\ t \geqq 0\}$ is estimated by (4.11) in terms of F, so L_2 is estimated in terms of F, c_*, T, x_0.

Next, we observe that

$$T_n(x) - T_{n-1}(x) = \int_{x_0}^x c_n^{-1} - c_{n-1}^{-1} = \int_{x_0}^x c_n^{-1} c_{n-1}^{-1} (c_{n-1} - c_n).$$

So

$$\begin{aligned} |T_n(x) - T_{n-1}(x)| &\leqq (x - x_0) c_*^{-2} \sup_{x_0 \leqq y \leqq x} |c_{n-1}(y) - c_n(y)| \\ &\leqq \Delta x c_*^{-2} \|c_{n-1} - c_n\|_\infty, \end{aligned} \tag{5.13}$$

where for the moment

$$\|c_{n-1} - c_n\|_\infty = \sup_{x_0 \leqq y \leqq x_0 + \Delta_x} |c_{n-1}(y) - c_n(y)|,$$

and so on.

According to the definition of $\tilde{H}$,

$$\tilde{H}(s, t) - \tilde{H}(t, \tau) = 4[H(s, \tau) - H(t, \tau)] - \int_0^{x_0} dy\, c^{-1}(y)[J(y, s) - J(y, t)] J(y, \tau).$$

Hence,

$$\begin{aligned} |\tilde{H}(s, \tau) - \tilde{H}(t, \tau)| &\leqq 4L_1|s-t| + x_0 c_*^{-1} \|J\|_\infty L_2 |s - t| \\ &= L_3|s-t|, \end{aligned} \tag{5.14}$$

where again L_3 is estimated in terms of F, c_*, T, and x_0. Next, estimate for $x_0 \leqq y \leqq x$

$$\begin{aligned} |J_n(y, t) - J_n(y, s)| &\leqq \Big| \tilde{H}(T_{n-1}(y), t) - \tilde{H}(T_{n-1}(y), s) \\ &\quad + \Big| \int_{x_0}^y dz\, c_{n-1}^{-1}(z) J_{n-1}(z, T_{n-1}(y)) \Big| \times \{J_{n-1}(z, t) - J_{n-1}(z, s)\} \Big| \\ &\leqq L_3|t-s| + (y - x_0) c_*^{-1} \|J_{n-1}\|_\infty \sup_{x_0 \leqq z \leqq y} [J_{n-1}(z, t) - J_{n-1}(z, s)]. \end{aligned} \tag{5.15}$$

According to (5.8), (5.9),

$$(y - x_0) c_*^{-1} \|J_{n-1}\|_\infty \leqq \delta'.$$

Select L_4 so that

$$L_4 \leqq (1-\delta')^{-1} L_3,$$

(which introduces the new restriction $\delta' < 1$). Then (5.15) and obvious induction guarantees that

$$\sup_{x_0 \leqq y \leqq x_0 + \Delta x} |J_n(y, s) - J_n(y, t)| \leqq L_4 |s-t|, \qquad n = 0, 1, 2, \cdots. \tag{5.16}$$

Note that c_n' satisfies

$$c_n'(x) = \frac{\partial k}{\partial x}(x, x_0) + 4J_{n-1}(x, T_{n-1}(x)) + \tfrac{1}{2}\int_{x_0}^{x} c_{n-1}^{-1}(c_{n-1}')^2.$$

Hence,

$$|c_n'(x)| \leqq \tfrac{1}{2}\int_0^{x_0} c^{-1}(c')^2 + 4|J_{n-1}(x, T_{n-1}(x))| + \tfrac{1}{2}\int_{x_0}^{x} c_{n-1}^{-1}(c_{n-1}')^2.$$

We suppose that $d^* > 0$ such that

$$\|c'\|_{L^\infty[0, x_0]} \leqq d^*,$$

and $|c_k(x)| \leqq d^{**}$, $k = 0, 1, \cdots, n-1$, $x_0 \leqq x \leqq x_0 + \Delta x$, where

$$d^{**} = \tfrac{1}{2} c_*^{-1} (d^*)^2 x_0 + 4(1+\delta) N(F, T, \varepsilon) + \delta.$$

Then

$$|c_n'(x)| \leqq \tfrac{1}{2} c_*^{-1} (d^*)^2 x_0 + 4(1+\delta) N(F, T, \varepsilon) + \tfrac{1}{2} \Delta x\, c_*^{-1} (d^{**})^2,$$

where we have used (5.9) and Lemma A. So we have proved:

LEMMA C. *Provided Δx satisfies the bounds of Lemma* B *and additionally*

$$\Delta x \leqq 2c_* (d^{**})^{-2} \delta,$$

we have the estimates for all $n \geqq 0$:

$$|c_n'(x)| \leqq d^{**}, \qquad x_0 \leqq x \leqq x_0 + \Delta x. \tag{5.17}$$

We are now ready for the main estimates. First,

$$\begin{aligned}
&|J_n(x, t) - J_{n-1}(x, t)| \\
&\quad \leqq |\tilde{H}(T_{n-1}(x), t) - \tilde{H}(T_{n-2}(x), t)| \\
&\qquad + \left| \int_{x_0}^{x} dy\, \{c_{n-1}^{-1}(y) J_{n-1}(y, T_{n-1}(x)) J_{n-1}(y, t) - c_{n-2}^{-1}(y) J_{n-2}(y, T_{n-2}(x)) J_{n-2}(y, t)\} \right| \\
&\quad \leqq \Delta x L_3 c_*^{-2} \|c_{n-1} - c_{n-2}\|_\infty \\
&\qquad + \int_{x_0}^{x} dy\, |c_{n-1}^{-1}(y) - c_{n-2}^{-1}(y)|\, |J_{n-1}(y, T_{n-1}(x)) J_{n-1}(y, t)| \\
&\qquad + \int_{x_0}^{x} dy\, c_{n-2}^{-1}(y) |J_{n-1}(y, T_{n-1}(x)) - J_{n-1}(y, T_{n-2}(x))|\, |J_{n-1}(y, t)| \qquad (5.18) \\
&\qquad + \int_{x_0}^{x} dy\, c_{n-2}^{-1}(y) |J_{n-1}(y, T_{n-2}(x)) - J_{n-2}(y, T_{n-2}(x))|\, |J_{n-1}(y, t)|
\end{aligned}$$

$$+\int_{x_0}^{x} dy\, c_{n-2}^{-1}(y)|J_{n-2}(y, T_{n-2}(x))|\,|J_{n-1}(y, t)-J_{n-2}(y, t)|$$

$$\begin{aligned}(5.18)\quad &\leqq \Delta x\, c_*^{-2}[L_3+\|J_{n-1}\|_\infty^2]\|c_{n-1}-c_{n-2}\|_\infty+\Delta x\, c_*^{-3}\|J_{n-1}\|_\infty L_4\|c_{n-1}-c_{n-2}\|_\infty\\ &\quad+\Delta x\, c_*^{-1}(\|J_{n-1}\|_\infty+\|J_{n-2}\|_\infty)\|J_{n-1}-J_{n-2}\|_\infty\\ &\leqq \Delta x\, A\|c_{n-1}-C_{n-2}\|_\infty+\Delta x B\|J_{n-1}-J_{n-2}\|_\infty,\end{aligned}$$

where

$$A=c_*^{-2}[L_3+(1+\delta')N(F, T, \varepsilon)[(1+\delta')N(F, T, \varepsilon)+c_*^{-1}L_4]],$$
$$B=2c_*^{-1}(1+\delta')N(F, T, \varepsilon).$$

Next, estimate

$$\begin{aligned}&|c_n(x)-c_{n-1}(x)|\\ &\quad\leqq 4\int_{x_0}^{x} dy\,|J_{n-1}(y, T_{n-1}(y))-J_{n-2}(y, T_{n-2}(y))|\\ &\qquad+\tfrac{1}{2}\int_{x_0}^{x} dy\int_{x_0}^{y}|c_{n-1}^{-1}(c'_{n-1})^2-c_{n-2}^{-1}(c'_{n-2})^2|\\ (5.19)\quad &\quad\leqq 4\int_{x_0}^{x} dy\,|J_{n-1}(y, T_{n-1}(y))-J_{n-1}(y, T_{n-2}(y))|\\ &\qquad+\tfrac{1}{2}\int_{x_0}^{x} dy\int_{x_0}^{y} c_{n-1}^{-1}c_{n-2}^{-1}|c_{n-1}-c_{n-2}|(c'_{n-1})^2\\ &\qquad+\tfrac{1}{2}\int_{x_0}^{x} dy\int_{x_0}^{y} c_{n-2}^{-1}|c'_{n-1}+c'_{n-2}|\,|c'_{n-1}-c'_{n-2}|\\ &\quad\leqq 4(\Delta x)^2c_*^{-2}L_4\|c_{n-1}-c_{n-2}\|_\infty+4\Delta x\|J_{n-1}-J_{n-2}\|_\infty\\ &\qquad+\tfrac{1}{4}(\Delta x)^2c_*^{-2}(d^{**})^2\|c_{n-1}-c_{n-2}\|_\infty+\tfrac{1}{2}(\Delta x)^2c_*^{-1}d^{**}\|c'_{n-1}-c'_{n-2}\|_\infty.\end{aligned}$$

Finally,

$$\begin{aligned}&|c'_n(x)-c'_{n-1}(x)|\\ &\quad\leqq 4|J_{n-1}(x, T_{n-1}(x))-J_{n-2}(x, T_{n-2}(x))|+\tfrac{1}{2}\int_{x_0}^{x}|c_{n-1}^{-1}(c'_{n-1})^2-c_{n-2}^{-1}(c'_{n-2})^2|\\ &\quad\leqq 4|J_{n-1}(x, T_{n-1}(x))-J_{n-1}(x, T_{n-2}(x))|+4|J_{n-1}(x, T_{n-2}(x))-J_{n-2}(x, T_{n-2}(x))|\\ (5.20)\quad &\qquad+\tfrac{1}{2}\int_{x_0}^{x} c_{n-1}^{-1}c_{n-2}^{-1}|c_{n-1}-c_{n-2}|(c'_{n-1})^2+\tfrac{1}{2}\int_{x_0}^{x} c_{n-2}^{-1}|c'_{n-1}+c'_{n-2}|\,|c'_{n-1}-c'_{n-2}|\\ &\quad\leqq 4\Delta x\, c_*^{-2}\|c_{n-1}-c_{n-2}\|_\infty+4\|J_{n-1}-J_{n-2}\|_\infty+\tfrac{1}{2}\Delta x\, c_*^{-2}(d^{**})^2\|c_{n-1}-c_{n-2}\|_\infty\\ &\qquad+\Delta x\, c_*^{-1}d^{**}\|c'_{n-1}-c'_{n-2}\|_\infty.\end{aligned}$$

The estimates (5.18), (5.19) and (5.20), taken together, show that, provided that $\Delta x > 0$ satisfies the bounds in Lemmas B and C, and is possibly smaller yet, the sequences $\{J_n\}$, $\{c_n\}$ and $\{c_n'\}$ converge uniformly on $x_0 \leqq x \leqq x_0 + \Delta x$ to solutions of (5.5). The numbers which determine how small Δx must be are c_*, δ', $N(F, T, \varepsilon)$, δ, L (in (5.10)), $\|F\|_\infty$, T, $\|J\|_\infty$, d^* and x_0. Of these, $N(F, T, \varepsilon)$, L, $\|F\|_\infty$ and T are determined by the data, c_* is estimated from below by (4.16a), d^* is estimated by (4.16b), $\|J\|_\infty$ is governed by the main a priori estimate (4.11), and

$$x_0 \leqq c^* T,$$

where c^* is an upper bound for c, given by (4.16a). It follows that, for given T and F satisfying the positivity condition (5.2), Δx may be chosen independently of x_0 so long as $x_0 \leqq X(T)$. Thus, finitely many repetitions of the iteration scheme outlined above suffice to determine c on the interval $[0, X(T)]$, and J on the corresponding domain. The system (5.5) (and with it (5.1)) has therefore been solved, as promised.

By differentiating the GL system (5.1), one obtains systems of Volterra equations for the derivatives of c and the partial derivatives of K. These are (essentially) linear systems. Without carrying out the details, we state that these systems for the derivatives possess continuous solutions. As in the case of the GL equation itself, c winds up with one more derivative than K, and K has as many derivatives as F. The solution of the GL system, therefore, defines a map $F \mapsto \{K, c\}$. It is easily verified that the positivity condition is stable under perturbation. It follows that the map $F \mapsto \{K, c\}$ is continuous in the obvious sense of C^m-norms, as outlined in the Introduction.

6. The lower bound on ε. For reasons explained in the Introduction, we now estimate $\varepsilon(T)$ (see (4.3)) in terms of the bounds c_*, c^*, d^* and e^* which we suppose given:

$$c_* \leqq c(x) \leqq c^*,$$
$$|c'(x)| \leqq d^*,$$
$$|c''(x)| \leqq e^*, \qquad 0 \leqq x \leqq X(T).$$

Recall that K solves the boundary value problem

$$\left(\frac{\partial^2}{\partial t^2} - c^2(x)\frac{\partial^2}{\partial x^2}\right)K(x, t) = 0, \qquad t \geqq T(x),$$

$$\begin{aligned} K(x, T(x)) &= \tfrac{1}{4}c^{1/2}\int_0^x c^{1/2}(c^{1/2})'' \\ &= \tfrac{1}{8}c^{1/2}(x)c'(x) - \tfrac{1}{16}c^{1/2}(x)\int_0^x c^{-1}(c')^2, \qquad x \geqq 0. \end{aligned} \tag{6.1}$$

Also, according to Lemma A (§ 3) (recall $c \equiv 1$ for $x < 0$),

$$K(x, -x) = 0, \qquad x < 0.$$

It follows, as in [11, Chapt. V], that K is the solution of an integral equation of Volterra type. In fact, if one denotes by $C(x, t)$ the intersection of the backward light

cone with vertex (x, t) with the forward light cone with vertex $(0, 0)$, and by $\Gamma(x, t)$ the x coordinate of the intersection of the characteristic curve of negative shape through (x, t) with the characteristic curve of positive slope through $(0, 0)$, one eventually obtains

$$J(x,t)=\tfrac{1}{2}\int_{C(x,t)}\{\tfrac{1}{2}c''+\tfrac{1}{4}c^{-1}(c')^2\}J$$

$$+\tfrac{1}{8}\Big[c'(\Gamma(x,t),t)-\tfrac{1}{2}\int_0^{\Gamma(x,t)}c^{-1}(c')^2\Big]$$

(recalling that $J(x, t)=2c^{1/2}(x)K(x, t)$). Now

$$\text{vol } C(x,t)\leqq\tfrac{1}{2}c^*t^2.$$

It follows easily that

$$|J(x,t)|\leqq Pf(\sqrt{c^*Q}t), \tag{6.2}$$

where f is the entire function with power series

$$f(z)=\tfrac{1}{2}\Big\{1+z^2+\frac{z^4}{3}+\frac{z^6}{5\cdot 3}+\frac{z^8}{7\cdot 5\cdot 3}+\cdots\Big\},$$

and

$$P=\sup_{0\leqq x\leqq X(T)}\tfrac{1}{8}\Big[c'(x)-\tfrac{1}{2}\int_0^x c^{-1}(c')^2\Big]\leqq\tfrac{1}{8}[d^*+\tfrac{1}{2}c^*c_*^{-1}(d^*)^2T],$$

$$Q=\sup_{0\leqq x\leqq X(T)}[\tfrac{1}{2}c''(x)+\tfrac{1}{4}c^{-1}(x)(c'(x))^2]\leqq\tfrac{1}{2}e^*+\tfrac{1}{4}c_*^{-1}(d^*)^2.$$

Recalling the definition (4.1) of $\tilde{K}$ and of the operator $\mathbb{K}$, one sees that $\tilde{K}$ obeys the same sup norm bound as J, namely (6.2) above. The kernel of the inverse operator $(\mathbb{I}+\mathbb{K})^{-1}$, say $\hat{K}$, is then easily bounded:

$$\|\hat{K}\|_\infty\leqq\|\tilde{K}\|_\infty\exp(\|\tilde{K}\|_\infty T). \tag{6.3}$$

It follows immediately that

$$\|(\mathbb{I}+\mathbb{K})^{-1}\varphi\|^2\leqq(1+\|\hat{K}\|_\infty T)^2\|\varphi\|^2.$$

So (see (4.4)),

$$\varepsilon(T)^{-1}\leqq(1+\|\tilde{K}\|_\infty T\exp(\|\tilde{K}\|_\infty T))^2,$$

which together with the bound (6.2) estimates $\varepsilon(T)$ in terms of T and the a priori information on c, as desired.

7. Equivalence of GL and Chudov systems. We show by direct computation that the solution of the GL system constructed in § 5 solves the Chudov boundary value

problem,

$$\left(\frac{\partial^2}{\partial t^2}-c^2\frac{\partial}{\partial x^2}\right)K\equiv 0 \quad \text{in } \{t\geqq 0,\ 0\leqq x\leqq X(t)\},$$

(7.1a)

$$K(0,t)=F(t),$$

(7.1b)

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial x}\right)K(0,t)\equiv 0,$$

(7.1c)

$$K(x,T(x))=\tfrac{1}{4}c^{1/2}(x)\int_0^x c^{1/2}(c^{1/2})''.$$

Now, it was shown in § 2 that the smooth part of the solution of the initial value problem (2.1) solves (7.1). Since the solution of (7.1) can easily be shown to be unique, it follows that the solution of the GL system is, in fact, the smooth part of the solution of (2.1) with corresponding coefficient c; hence, we have solved the inverse problem.

First note from (3.10c) that

$$\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial s^2}\right)H(s,t)=F(t)F'(s)-F'(t)F(s). \tag{7.2}$$

Now denote by $\{K, c\}$ the solution of the GL system as constructed in § 5. We assume first that F satisfies the positivity condition (5.2) and is of class C^2 on its interval of definition, so that K and c are of classes C^2 and C^3, respectively, on their domains of definition.

Setting $s=0$ in (5.10) and recalling the definition (3.10c) of H one obtains

$$F(t)=K(0,t).$$

Using the definition (3.10c) again, and the requirement $F(0)=0$, one sees that

$$\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial s}\right)H(0,t)\equiv 0.$$

On the other hand, from (5.1) one obtains (recalling $c(0)=1$, $c'(0)=0$ and $K(0,0)=F(0)=0$)

$$0=\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial s}\right)H(0,t)=\tfrac{1}{2}(D_2K(0,t)-D_1K(0,t)).$$

Thus, K obeys the boundary conditions on the Chudov system. It remains only to verify that K solves the wave equation in the interior of the light cone. To do this, first compute

$$\begin{aligned}\left(\frac{\partial^2}{\partial t^2}-\frac{\partial^2}{\partial s^2}\right)H(s,t)=&\tfrac{1}{2}c^{-1/2}(X(s))D_2^2K(X(s),t)\\&+\left[\int_0^{X(s)}dy\,c^{-2}(y)K(y,s)D_2^2K(y,t)\right]\\&-\tfrac{1}{8}c^{-1/2}(X(s))(c'(X(s)))^2K(X(s),t)\\&+\tfrac{1}{4}c^{1/2}(X(s))c''(X(s))K(X(s),t)\\&-\tfrac{1}{2}c^{3/2}(X(s))D_1^2K(X(s),t)\\&-\frac{d}{ds}[c^{-1}(X(s))K(X(s),s)K(X(s),t)]\end{aligned}$$

$$-c^{-1}(X(s))D_2K(X(s), s)K(X(s), t)$$
$$-\int_0^{X(s)} dy\, c^{-2}(y)D_2^2K(y, s)K(y, t).$$

Using (7.1c) one sees that

$$c^{-1/2}(X(s))\frac{d}{ds}[2c^{-1/2}(X(s))K(X(s), s)]$$
$$= 2c^{-1}(X(s))\frac{d}{ds}K(X(s), s) - c^{-1}(X(s))c'(X(s))K(X(s), s)$$
$$= \tfrac{1}{4}c^{1/2}(X(s))c''(X(s)) - \tfrac{1}{8}c^{-1/2}(X(s))(c'(X(s)))^2.$$

Also,

$$\frac{d}{ds}[c^{-1}(X(s))K(X(s), s)K(X(s), t] + c^{-1}(X(s))D_2K(X(s), s)K(X(s), t)$$
$$= -c^{-1}(X(s))c'(X(s))K(X(s), s)K(X(s), t)$$
$$+2c^{-1}(X(s))\frac{d}{ds}K(X(s), s)K(X(s), t) - D_1K(X(s), s)K(X(s), t)$$
$$+K(X(s), s)D_1K(X(s), t)$$
$$= [\tfrac{1}{4}c^{1/2}(X(s))c''(X(s)) - \tfrac{1}{8}c^{-1/2}(X(s))(c'(X(s)))^2]K(X(s), t)$$
$$-D_1K(X(s), s)K(X(s), t) + K(X(s), s)D_1K(X(s), t).$$

Hence,

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial s^2}\right)H(s, t) = \tfrac{1}{2}c^{-1/2}(X(s))[D_2^2K(X(s), t) - c^2(X(s))D_1^2K(X(s), 1)]$$
$$+\int_0^{X(s)} dy\, c^{-2}(y)[K(y, s)D_2^2K(y, t) - D_2^2K(y, s)K(y, t)] \tag{7.3}$$
$$+D_1K(X(s), s)K(X(s), t) - K(X(s), s)D_1K(X(s), t).$$

Now add to (7.3) the identity

$$0 = -D_1K(X(s), s)K(X(s), t) + K(X(s), s)D_1K(X(s), t)$$
$$+D_1K(0, s)K(0, t) - K(0, s)D_1K(0, t)$$
$$-\int_0^{X(s)} dy\, c^{-2}(y)[K(y, s)(c^2(y)D_1^2K(y, t)) - (c^2(y)D_1^2K(y, s))K(y, t)],$$

obtained by integration by parts. After a little manipulation, one gets

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial s^2}\right)H(s, t) - D_1K(0, s)K(0, t) + K(0, s)D_1K(0, t)$$
$$= \tfrac{1}{2}c^{-1/2}(X(s))[D_2^2K(X(s), t) - c^2(X(s))D_1^2K(X(s), t)]$$
$$+\int_0^{X(s)} dy\, c^{-2}(y)[K(y, s)\{D_2^2K(y, t) - c^2(y)D_1^2K(y, t)\} \tag{7.4}$$
$$-\{D_2^2K(y, s) - c^2(y)D_1^2K(y, s)\}K(y, t)].$$

Comparing with (7.2) and using (7.1b), one sees that the l.h.s. of (7.4) in fact vanishes identically. Equation (7.4) is therefore a linear integral equation of Volterra type, since K is continuous and $c>0$, for $D_2^2K-c^2D_1^2K$. It follows that this expression vanishes in the domain covered by the limits of integration, that is, in the light cone.

To obtain this result for $F\in C^1$, hence $c\in C^2$, $K\in C^1$, we must resort to a limiting argument, since $\Box_c K$ is defined only as a distribution in the interior of the light cone. We assume that F obeys an estimate (4.3). Let $\{F_n\}\subset C^2$ have F as C^1 limit. Then an easy estimate (like (5.7) in [6]) shows that $\{F_n\}$ obeys an estimate (4.3) with ε replaced by $\frac{1}{2}\varepsilon$. Therefore the GL equations for F_n have solutions $c_n\in C^3$ and $K_n\in C^2$, and $c_n\to c$ in C^2, $K_n\to K$ in C^1. By the result just proved, $\Box_{c_n}K_n=0$ for all n. For any test function ϕ supported in the interior of the light cone for c,

$$\begin{aligned}\langle\Box_c K,\phi\rangle &= \int -D_2KD_2\phi+D_1KD_1(c^2\phi)\\ &= \lim_{n\to\infty}\int -D_2K_nD_2\phi+D_1K_nD_1(c_n^2\phi)\\ &= \lim_{n\to\infty}\langle\Box_{c_n}K_n,\phi\rangle=0.\end{aligned}$$

The second equality is valid because the light cone for c_n tends to the light cone for c, and supp ϕ must have a nonzero distance to the boundary rays.

The result shows that the solution K of the GL equation solves the Chudov system also. Since we showed in §§ 2 and 3 that the solution of the Chudov system solves the GL system, it follows that these two systems are *completely equivalent.* It also follows that c, as part of the solution of the GL system for prescribed F, solves the inverse problem stated in § 1. Together with the stability statement of § 5, this constitutes a proof of Theorem B.

Acknowledgments. I would like to thank Ken Driessel, Mike Gage and the referee for a number of helpful suggestions and comments. This paper appeared in preliminary form as Technical Summary Report No. 2007 of the Mathematics Research Center, University of Wisconsin Madison. I would like to thank the staff of MRC for their support and skillful assistance, and the secretarial staff at MSU for a superb typing job.

REFERENCES

[1] J. BERRYMAN, *Inverse Methods for Elastic waves in Stratified Media*, Bell Laboratories, (preprint, 1979).

[2] J. WARE AND K. AKI, *Continuous and discrete inverse-scattering problems in a stratified elastic medium* I: *Plane waves at normal incidence*, J. Acoust. Soc. Amer., 45 (1969), pp. 911–921.

[3] D. COHEN AND N. BLEISTEIN, *An inverse method for determining small variations in propagation speed*, SIAM J. Appl. Math., 32 (1977), pp. 784–799.

[4] K. CHADAN AND P. SABATIER, *Inverse Problems in Quantum Scattering Theory*, Springer, New York, 1977.

[5] P. GOPILLAUD, *An approach to inverse filtering of near-surface layer effects from seismic records*, Geophys., 26 (1961), pp. 754–760.

[6] W. SYMES, *Inverse boundary value problems and a theorem of Gel'fand and Levitan*, J. Math. Anal. Appl., 71 (1979), pp. 378–402.

[7] ———, *Numerical stability in an inverse scattering problem*, MRC TSR 1990, University of Wisconsin, Madison, Wisconsin, 1979; SIAM J. Numer. Anal., 17 (1980), pp. 707–732.

[8] O. HALD, *The inverse Sturm–Liouville problem for symmetric potentials*, Acta Math. (Sweden), 141 (1979), pp. 264–291.

[9] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121–251.

[10] P. LAX, *Notes on Hyperbolic P.D.E.*, Stanford University, 1963.
[11] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Wiley-Interscience, New York, 1962.
[12] W. SYMES, *Invertibility of the trace operator and local energy decay for the string with* BV *coefficients*; *Sharp bounds in the* 1-*D inverse reflection problem* (preprints, 1980).
[13] I. M. GEL'FAND AND B. M. LEVITAN, *On the determination of a differential equation from its spectral function*, Isv. Akad. Nauk. SSSR Ser. Matem., 15 (1951), pp. 309–360, (in Russian) = AMS Trans. Ser. 2, 1 (1955), pp. 253–304.
[14] L. D. FADEEV, *The inverse problem in the quantum theory of scattering*, J. Math. Phys., 4 (1963), pp. 72–104.
[15] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions* I, Academic Press, New York, 1964.
[16] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer, New York, 1964.
[17] R. KREUGER, *An inverse problem for an absorbing medium with multiple discontinuities*, Quart. Appl. Math., (1978), pp. 235–253.
[18] I. KAY, *The inverse problem when the coefficient is a rational function*, Comm. Pure. Appl. Math., 13 (1960), pp. 371–393.
[19] A. BAMBERGER, G. CHAVENT AND P. LAILLY, *About the stability of the inverse problem in* 1-D *wave equations—application to the interpretation of seismic profiles*, Appl. Math. Optim., 5 (1979), pp. 1–47.
[20] M. GERVER, *Inverse problem for the one-dimensional wave equation*, Geophys. J. Roy. Astronom. Soc., 21 (1970), pp. 337–357.
[21] M. SONDHI AND B. GOPINATH, *Determination of the shape of the human vocal tract from acoustical measurements*, J. Acoust. Soc. Amer., 49 (1971), pp. 1867–1873.
[22] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0012 $01.00/0

THE BASIC BESSEL FUNCTIONS AND POLYNOMIALS*

MOURAD E. H. ISMAIL†

This paper is dedicated to my parents on the occasion of their 40th anniversary

Abstract. Basic analogues of the Bessel polynomials and their generalization are introduced. These polynomials are orthogonal on the unit circle $|z| = 1$ with respect to a complex weight function. They satisfy a three-term recurrence relation, and the associated continued fraction is computed. Similar results are also established for the little q-Jacobi polynomials. Integral representations for the modified basic Bessel functions are also established.

Introduction and notation. The present work is a detailed study of a basic analogue of the Bessel polynomials

$$y_n(x) = {}_2F_0\left(-n, n+1; -; -\frac{x}{2}\right) \tag{1.1}$$

and the generalized Bessel polynomials

$$y_n(x; a) = {}_2F_0\left(-n, n+a-1; -; -\frac{x}{2}\right); \tag{1.2}$$

see Grosswald [10, (8) p. 6, (10) p. 7, (27) p. 13]. The hypergeometric function ${}_2F_0$ is defined by

$${}_2F_0(a, b; -; x) = \sum_0^\infty \frac{(a)_n (b)_n}{n!} x^n, \tag{1.3}$$

where

$$(a)_0 = 1, \qquad (a)_n = a(a+1)\cdots(a+n-1), \qquad n > 0. \tag{1.4}$$

Another notation for the generalized Bessel polynomials is

$$y_n(x; a, b) = {}_2F_0\left(-n, n+a-1; -; -\frac{x}{b}\right). \tag{1.5}$$

The generalized Bessel polynomials are orthogonal on the unit circle with respect to a complex "weight" function. The orthogonality relation is (Grosswald [10, p. 30])

$$\frac{1}{2\pi i}\int_{|z|=1} y_n(z; a) y_m(z; a)\left\{\sum_{k=0}^\infty \frac{(-2/z)^k}{(a-1)_k}\right\} dz = \frac{2(-1)^{n+1} n!}{(2n+a-1)(a-1)_n}\delta_{m,n}. \tag{1.6}$$

The zeros of the Bessel polynomials have very interesting properties; see Burchnal [6], Grosswald [9], [10] and Ismail and Kelker [14]. The recent interesting article of De Bruin, Saff and Varga [8] contains refinements of the asymptotic results collected in [10] and settles a conjecture of Luke.

We recall that the basic hypergeometric function ${}_{r+1}\phi_r$ is

$${}_{r+1}\phi_r(a_1, \cdots; a_{r+1}; b_1, \cdots, b_r; q; x) = \sum_0^\infty \frac{(a_1; q)_n \cdots (a_{r+1}; q)_n}{(b_1; q)_n \cdots (b_r; q)_n} \frac{x^n}{(q; q)_n}, \tag{1.7}$$

* Received by the editors June 17, 1980 and in revised form September 16, 1980. This research was partially supported by the National Science Foundation under grant MCS-7903518 and the Natural Science and Engineering Research Council of Canada under grant A4522.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85281.

with

$$(1.8)\qquad (a;q)_0=1,\quad (a;q)_n=(1-a)\cdots(1-aq^{n-1}),\quad (a;q)_\infty=\prod_0^\infty(1-aq^n).$$

Abdi [1] defines a basic analogue of $y_n(x;a)$ as

$$(1.9)\qquad J(q;a-1;n;x)=\frac{(q^{a-1};q)_n}{(q;q)_n}{}_2\phi_1(q^{-n},q^{a+n-1};0;q,x),$$

which is a q-analogue of the ${}_2F_0$ representation (1.2). Grosswald [10, p. 151] says that, in the spirit of a "basic" function, $\lim_{q\to1^-}J(q;a-1;n;x)$ is $((a-1)_n/n!)y_n(x;a)$. He then adds "otherwise, the connection of $J(q;a-1;n;x)$ to $y_n(x;a,b)$ is rather tenuous". Our basic analogue is different from Abdi's. Our polynomials are orthogonal on the unit circle and arise from the basic Bessel functions in the same way the Bessel polynomials and their generalization arise from the ordinary Bessel functions.

The paper is arranged as follows. Section 2 discusses the relationship between the Bessel and Lommel polynomials and the Bessel functions. When we apply the same procedure to basic Bessel functions we are naturally led to basic Lommel polynomials (Ismail [13]) and basic Bessel polynomials. The simple basic Bessel polynomials turn out to be

$$(1.10)\qquad y_n(x|q^2)=q^{n(n-1)/2}{}_2\phi_1(q^{-n},q^{n+1};-q;q,-2xq),$$

and are generalized to

$$(1.11)\qquad y_n(x;a|q^2)=q^{n(n-1)/2}{}_2\phi_1(q^{-n},q^{n+a-1};-q;q,-2xq),$$

bearing the same analogy to (1.1) and (1.2). In § 3 we establish the orthogonality relation for the polynomials in (1.10) and (1.11). Both our polynomials and Abdi's polynomials are special cases of the little q-Jacobi polynomials of Hahn,

$$(1.12)\qquad p_n(x;\alpha,\beta|q)={}_2\phi_1(q^{-n},\alpha\beta q^{n+1};\alpha q;q,qx),$$

studied by Andrews and Askey [3], [4] and Hahn [11]. They proved the orthogonality relation

$$(1.13)\qquad \sum_0^\infty\frac{\alpha^jq^j(q^{j+1};q)_\infty}{(\beta q^{j+1};q)_\infty}p_n(q^j;\alpha,\beta|q)p_m(q^j,\alpha,\beta|q)=\frac{\alpha^nq^n(\alpha\beta q^{n+1};q)_\infty(q;q)_n}{(\beta q^{n+1};q)_\infty(\alpha q;q)_\infty(\alpha q;q)_n(1-\alpha\beta q^{2n+1})}\delta_{m,n}.$$

Following the ideas in [12], Al-Salam and Ismail [2] derived reproducing kernels for the little q-Jacobi polynomials; also see Stanton's very interesting work [18]. Our polynomials correspond to $\alpha=-1$, $\beta=-q^{a-2}$, while Abdi's correspond to the limiting case $\beta=q^{a-2}/\alpha$ and $\alpha\to0$. In § 3 we also establish the orthogonality of the little q-Jacobi polynomials on a circle with respect to a complex weight function. To the best of my knowledge this is a new result. It is interesting to note that (1.13) follows from our complex orthogonality relation, as will be indicated in § 3. In § 4 the continued fraction associated with $y_n(x|q^2)$ is computed using the asymptotic results of [13]. A generating function and an asymptotic formula are also established. Using a theorem of Markoff and analytic continuation arguments we compute the continued fraction associated with the little q-Jacobi polynomials. The last section, § 5, contains a basic analogue of the Hankel integral representation for $1/\Gamma(z)$. This integral is related to some integrals of Ramanujan; see Askey [5]. Using the basic analogue of the Hankel integral we derive an integral representation for a modified basic Bessel function.

2. The basic Bessel polynomials. There are two sets of orthogonal polynomials associated with Bessel functions. The first is the set of Lommel polynomials $\{R_{n,\nu}(1/z)\}$ which arises when we iterate the recursion (Watson [19, p. 294)]).

$$J_{\nu+1}(z)=\frac{2\nu}{z}J_\nu(z)-J_{\nu-1}(z), \tag{2.1}$$

to

$$J_{\nu+m}(z)=R_{m,\nu}(z)J_\nu(z)-R_{m-1,\nu+1}(z)J_{\nu-1}(z). \tag{2.2}$$

The Lommel polynomials are orthogonal on a finite interval; see [13] for references and details. The Bessel function of the first kind $J_\nu(z)$ has two basic analogues, namely

$$\begin{aligned} J_\nu^{(1)}(z;q)&=\frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\sum_0^\infty \frac{(-1)^n(z/2)^{\nu+2n}}{(q;q)_n(q^{\nu+1};q)_n},\\ J_\nu^{(2)}(z;q)&=\frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\sum_0^\infty \frac{(-1)^n(z/2)^{\nu+2n}}{(q;q)_n(q^{\nu+1};q)_n}q^{n(\nu+n)}, \end{aligned} \tag{2.3}$$

for $0<q<1$. Jackson [15] introduced the basic Bessel functions using a different notation. The above notation is due to us in [13]. The function $J_\nu^{(2)}(z;q)$ is an entire transcendental function. It is clear that

$$\lim_{q\to 1^-} J_\nu^{(k)}(z(1-q);q)=J_\nu(z), \qquad k=1,2.$$

The basic Bessel functions satisfy the recursion [13]

$$q^\nu J_{\nu+1}^{(j)}(z;q)=2(1-q^\nu)z^{-1}J_\nu^{(j)}(z;q)-J_{\nu-1}^{(j)}(z;q), \qquad j=1,2. \tag{2.4}$$

Iterating the recursion (2.5) leads to the basic Lommel polynomials [13]. We now introduce the modified basic Bessel functions $I_\nu^{(j)}(z;q)$ and $K_\nu^{(j)}(z;q)$, $j=1,2$ in the following manner:

$$I_\nu^{(j)}(z;q)=e^{-i\pi\nu/2}J_\nu^{(j)}(z\,e^{i\pi/2};q) \tag{2.5}$$

and

$$\begin{aligned} K_\nu^{(j)}(z;q)&=\frac{\pi}{2\sin(\pi\nu)}\{I_{-\nu}^{(j)}(z;q)-I_\nu^{(j)}(z;q)\}, \qquad \nu\neq 0,\pm1,\pm2,\cdots,\\ K_n^{(j)}(z;q)&=\lim_{\nu\to n}K_\nu^{(j)}(z;q), \qquad n=0,\pm1,\pm2,\cdots, \end{aligned} \tag{2.6}$$

$j=1,2$. These definitions are analogues to the case $q=1$ (Watson [19, pp. 77, 78)]. Surprisingly, the modified basic Bessel functions have not been studied earlier.

The recurrence relation

$$q^\nu K_{\nu+1}^{(j)}(z;q)=2(1-q^\nu)z^{-1}K_\nu^{(j)}(z;q)+K_{\nu-1}^{(j)}(z;q), \qquad j=1,2 \tag{2.7}$$

follows from (2.4), (2.5) and (2.6). The $I_\nu^{(j)}(z;q)$ also satisfies (2.7). We now iterate (2.7) to obtain

$$q^{\nu m+m(m-1)/2}K_{\nu+m}^{(j)}(z;q)=\phi_{m,\nu}(z)K_\nu^{(j)}(z;q)+\phi_{m-1,\nu+1}(z)K_{\nu-1}^{(j)}(z;q), \tag{2.8}$$

where the functions $\phi_{m,\nu}(z)$ depend on q, of course, and are generated by

$$\phi_{m+1,\nu}(z)=2(1-q^{m+\nu})z^{-1}\phi_{m,\nu}(z)+q^{m+\nu-1}\phi_{m-1,\nu}(z), \qquad m=1,2,\cdots, \tag{2.9}$$

and the initial conditions

$$\phi_{0,\nu}(z)=1, \qquad \phi_{1,\nu}(z)=2(1-q^{\nu})z^{-1}. \tag{2.10}$$

When $\nu=\frac{1}{2}$ we get

$$K^{(j)}_{m+1/2}(z;q)=q^{-m^2/2}\{\phi_{m,1/2}(z)+\phi_{m-1,3/2}(z)\}K^{(j)}_{1/2}(z;q), \tag{2.11}$$

because $K^{(j)}_{\nu}(z;q)$ is an even function of ν; see (2.6). The polynomial

$$y_m(z|q)=\phi_{m,1/2}\left(\frac{1}{z}\right)+\phi_{m-1,3/2}\left(\frac{1}{z}\right) \tag{2.12}$$

is the basic analogue of the simple polynomial $y_n(z)$. The polynomials $\{y_n(x)\}$ are similarly related to the modified Bessel functions.

We now derive the explicit formula

$$y_n(z|q)=\sum_{k=0}^{n}\frac{(q^{1/2};q^{1/2})_{n+k}(2z)^k}{(q;q)_k(q^{1/2};q^{1/2})_{n-k}}q^{(n-k)(n-k-1)/4}. \tag{2.13}$$

We first go back to (2.9) and (2.10), and replace z by iz to get

$$i^{n+1}\phi_{n+1,\nu}(iz)=2i^n(1-q^{n+\nu})z^{-1}\phi_{n,\nu}(iz)-i^{n-1}q^{n+\nu-1}\phi_{n-1,\nu}(iz),$$

which when compared with the three-term recurrence relation for the q-Lommel polynomials $\{R_{n,\nu}(z;q)\}$ (Ismail [13]) identifies $\phi_{n,\nu}(z)$ as

$$\phi_{n,\nu}(z)=i^{-n}R_{n,\nu}(-iz;q). \tag{2.14}$$

We now use this identification, the explicit formula [13],

$$R_{n,\nu}(z;q)=\sum_{j=0}^{[n/2]}\frac{(-1)^j(q^{\nu};q)_{n-j}(q;q)_{n-j}}{(q;q)_j(q^{\nu};q)_j(q;q)_{n-2j}}\left(\frac{z}{2}\right)^{2j-n}q^{j(j+\nu-1)}, \tag{2.15}$$

and the definition (2.12) to get

$$\begin{aligned} y_n(z|q)=i^{-n}&\sum_{0}^{[n/2]}\frac{(2iz)^{n-2j}(-1)^j(q^{1/2};q)_{n-j}(q;q)_{n-j}}{(q;q)_j(q^{1/2};q)_j(q;q)_{n-2j}}q^{j(j-1/2)}\\ &+i^{-n}\sum_{0}^{[(n-1)/2]}\frac{(2iz)^{n-1-2j}(-1)^j(q^{3/2};q)_{n-1-j}(q;q)_{n-1-j}}{(q;q)_j(q^{3/2};q)_j(q;q)_{n-2j-1}}q^{j(j+1/2)}. \end{aligned} \tag{2.16}$$

When n is even, (2.16) and the observations

$$(q^{1/2};q)_m(q;q)_m=(q^{1/2};q^{1/2})_{2m}, \qquad (q^{3/2};q)_m(q;q)_m=(q;q^{1/2})2m$$

imply (2.13). Similarly the case of odd n can be handled.

Another way of proving (2.13) is to first establish the three-term recurrence relation

$$y_{n+1}(z|q)=2(1-q^{n+1/2})zy_n(z|q)+q^{n-1/2}y_{n-1}(z|q), \qquad n\geqq 1, \tag{2.17}$$

by letting $\nu=n+\frac{1}{2}$ in (2.7) and making use of

$$K^{(j)}_{n+1/2}(z;q)=q^{-n^2/2}y_n\left(\frac{1}{z}\right)K^{(j)}_{1/2}(z;q), \tag{2.18}$$

which follows from (2.11) and (2.12). Then use (2.18) to get

$$y_0(z)=1, \qquad y_1(z)=1+2(1-q^{1/2})z. \tag{2.19}$$

Now it is a routine matter to see that (2.13) satisfies the recursion (2.17) and the initial conditions (2.19).

A basic hypergeometric representation of $y_n(z|q)$ follows from (2.13) and

$$\frac{(q;q)_n}{(q;q)_{n-k}} = (q^{-n};q)_k(-1)^k q^{k(2n-k+1)/2}$$

(Slater [17, p. 241]). The result is

$$y_n(x|q^2) = q^{n(n-1)/2}{}_2\phi_1(q^{-n}, q^{n+1}; -q; q, -2qx). \tag{2.20}$$

Clearly as $q \to 1$, $y_n(x/(1-q)|q) \to {}_2F_0(-n, n+1; -; x/2)$ as one expected. Formula (2.20) suggests defining the generalized Bessel polynomials $y_n(x; a|q^2)$ by

$$y_n(x; a|q^2) = q^{n(n-1)/2}{}_2\phi_1(q^{-n}, q^{n+a-1}; -q; q, -2qx), \tag{2.21}$$

which is radically different from Abdi's (1.9) although as $q \to 1$ both essentially tend to the common limit ${}_2F_0(-n, n+a-1; -; x/2)$. We shall assume that a is neither a negative integer nor zero in order to ensure that $y_n(x; a|q^2)$ has degree n.

3. Orthogonality. Let $w(z)$ be a weight function, possibly complex-valued, with moments w_k; that is,

$$w_k = \int_U z^k w(z)\, dz, \qquad k \geqq 0, \tag{3.1}$$

where $w(z)$ and the contour U are to be determined from the orthogonality relation

$$\int_U w(z) z^k p_n(z; \alpha, \beta|q)\, dz = 0, \qquad k = 0, 1, \cdots, n-1. \tag{3.2}$$

Therefore

$$\sum_{j=0}^{n} \frac{(q^{-n};q)_j(\alpha\beta q^{n+1};q)_j}{(q;q)_j(\alpha q;q)_j} q^j w_{j+k} = 0, \qquad k = 0, 1, \cdots, n-1. \tag{3.3}$$

Since all the known identities of this type are basic hypergeometric function identities, we attempt

$$w_k = \frac{(c;q)_k}{(d;q)_k}, \tag{3.4}$$

and proceed to compute c and d from (3.3). This leads to

$${}_3\phi_2(q^{-n}, \alpha\beta q^{n+1}, cq^k; \alpha q, dq^k; q, q) = 0. \tag{3.5}$$

We used $(c;q)_{k+j} = (c;q)_k(cq^k;q)_j$. We now go through the list of summable ${}_3\phi_2$'s in Slater [17, p. 247] and realize that the ${}_3\phi_2$ in (3.5) must be balanced (Saalschutzian). A basic hypergeometric function (1.7) is balanced if

$$q \prod_1^{r+1} a_j = \prod_1^r b_j.$$

The limiting case $\alpha \to 0, \beta \to \infty$ while $\alpha\beta$ has a limit will be treated separately. The basic hypergeometric function in (3.5) is balanced if $\alpha = 0$ or $d = \beta cq$ when $\alpha \neq 0$. If $\alpha = 0$, the left side of (3.5) is $(cq^k)^n(d/c;q)_n/(dq^k;q)_n$ and it is clear that no choice of c and d will make it vanish when $k = 0, 1, \cdots, n-1$. Thus $\alpha \neq 0$ and d is βqc. With this choice of d

the left side of (3.5) becomes

$$\frac{(q^{-n}/\beta;q)_n(\alpha q^{1-k}/c;q)_n}{(\alpha q;q)_n(q^{-n-k}c^{-1}\beta^{-1};q)_n},$$

in view of [17, p. 247]. The only way for the above expression to vanish when $0 \leqq k < n$ is if $c = \alpha q$. This leads to

$$w_k = \frac{(\alpha q;q)_k}{(\alpha\beta q^2;q)_k}. \tag{3.6}$$

Set

$$\rho(z;\alpha,\beta|q) = \sum_0^\infty z^{-k-1}\frac{(\alpha q;q)_k}{(\alpha\beta q^2;q)_k}. \tag{3.7}$$

Clearly,

$$\frac{1}{2\pi i}\int_{|z|=r} z^k\rho(z;\alpha,\beta|q)\,dz = w_k$$

holds for any $r>1$. Therefore

$$\frac{1}{2\pi i}\int_{|z|=r} p_n(z;\alpha,\beta|q)p_m(z;\alpha,\beta|q)\rho(z;\alpha,\beta|q)\,dz = \lambda_n\delta_{m,n}.$$

We now evaluate λ_n. It is clear that

$$\begin{aligned}
\lambda_n &= \frac{1}{2\pi i}\int_{|z|=r} p_n^2(z;\alpha,\beta|q)\rho(z;\alpha,\beta|q)\,dz \\
&= \frac{(q^{-n};q)_n(\alpha\beta q^{n+1};q)_n}{(q;q)_n(\alpha q;q)_n}q^n\cdot\frac{1}{2\pi i}\int_{|z|=r} z^n p_n(z;\alpha,\beta|q)\rho(z;\alpha,\beta|q)\,dz \\
&= \frac{(q^{-n};q)_n(\alpha\beta q^{n+1};q)_n}{(q;q)_n(\alpha q;q)_n}\cdot\frac{(\alpha q;q)_n}{(\alpha\beta q^2;q)_n}q^n \\
&\quad\cdot {}_3\phi_2(q^{-n},\alpha\beta q^{n+1},\alpha q^{n+1};\alpha\beta q^{n+2},\alpha q;q,q) \\
&= q^n\frac{(q^{-n};q)_n(\alpha\beta q^{n+1};q)_n}{(q;q)_n(\alpha\beta q^2;q)_n}\cdot\frac{(q;q)_n(\beta q;q)_n}{(\alpha\beta q^{n+2};q)_n(q^{-n}/\alpha;q)_n},
\end{aligned}$$

by [17, IV-4, p. 247]. After some straightforward manipulations we establish the orthogonality relation

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{|z|=r>1} p_n(z;\alpha,\beta|q)p_m(z;\alpha,\beta|q)\rho(z;\alpha,\beta|q)\,dz \\
&\qquad = \alpha^n q^n\frac{(\beta q;q)_n(q;q)_n(1-\alpha\beta q)}{(\alpha q;q)_n(\alpha\beta q;q)_n(1-\alpha\beta q^{2n+1})}\delta_{m,n},
\end{aligned} \tag{3.8}$$

holding for $\alpha,\ \beta \neq q^{-n},\ \alpha\beta \neq q^{-n}$ for any positive integer n. For our q-Bessel polynomials (2.20) the orthogonality relation (3.8) reduces to

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{|z|=r>1/2} y_n(z;a|q^2)y_m(z;a|q^2)\left\{\sum_0^\infty\frac{(-1,q)_k}{(q^{a-1};q)_k}(-2z)^{-k}\right\}dz \\
&\qquad = \frac{(-1)^{n+1}q^{n^2}(-q^{a-1};q)_n(q;q)_n}{(-q;q)_n(q^{a-1};q)_n(1-q^{2n+a-1})}\delta_{m,n}.
\end{aligned} \tag{3.9}$$

In certain cases the weight function in (3.7), (3.8) and (3.9) can be simplified by adding an analytic function to it. The sum to be used is Ramanujan's ${}_1\psi_1$ sum (Slater [17, p. 248])

$$\sum_{-\infty}^{\infty} \frac{(c;q)_n}{(d;q)_n} z^n = \frac{(dc^{-1};q)_\infty (q;q)_\infty (qc^{-1}z^{-1};q)_\infty (cz;q)_\infty}{(d;q)_\infty (dc^{-1}z^{-1};q)_\infty (qc^{-1};q)_\infty (z;q)_\infty}, \tag{3.10}$$

where

$$(\lambda;q)_n = \prod_0^\infty \left\{ \frac{(1-\lambda q^j)}{(1-\lambda q^{n+j})} \right\}, \qquad n = 0, \pm 1, \cdots. \tag{3.11}$$

It is easy to see that (3.11) agrees with (1.8) when n is nonnegative. Furthermore

$$(\lambda;q)_{-n} = \frac{(-\lambda)^{-n} q^{n(n+1)/2}}{(q/\lambda;q)_n}. \tag{3.12}$$

The bilateral series in (3.10) converges if and only if $|z|<1$ and $|z|>|d/c|$, $c \neq 0$, $c \neq q^n$, $d \neq q^{-n}$ for any nonnegative integer n. Set

$$\rho_1(z;\alpha,\beta|q) = \sum_{-\infty}^{\infty} z^{-k} \frac{(\alpha;q)_k}{(\alpha\beta q;q)_k}. \tag{3.13}$$

Clearly, $\rho(z;\alpha,\beta|q) - (1-\alpha\beta q)(1-\alpha)^{-1}\rho_1(z;\alpha,\beta|q)$ is an analytic function of z when $\alpha \neq 1$, $\alpha\beta \neq 1$ and both ρ and ρ_1 converge; hence, all its moments vanish. We have, by (3.10),

$$\rho_1(z;\alpha,\beta|q) = \frac{(\beta q;q)_\infty (q;q)_\infty (qz/\alpha;q)_\infty (\alpha/z;q)_\infty}{(\alpha\beta q;q)_\infty (q\beta z;q)_\infty (q/\alpha;q)_\infty (1/z;q)_\infty}. \tag{3.14}$$

This proves the following theorem.

THEOREM 3.1. *When $|\beta q|<1$, $\alpha \neq 0$, $\alpha \neq q^n$, $\alpha\beta q \neq q^{-n}$, $n = 0, 1, \cdots$, the orthogonality relation*

$$\begin{aligned} &\frac{1}{2\pi i} \int_{|z|=r} p_n(z;\alpha,\beta|q) p_m(z;\alpha,\beta|q) \frac{(qz/\alpha;q)_\infty (\alpha/z;q)_\infty}{(q\beta z;q)_\infty (1/z;q)_\infty} dz \\ &\quad = \frac{\alpha^n q^n (\alpha\beta q^{n+1};q)_\infty (q/\alpha;q)_\infty (1-\alpha)}{(q^{n+1};q)_\infty (\beta q^{n+1};q)_\infty (\alpha;q)_n (1-\alpha\beta q^{2n+1})} \delta_{mn} \end{aligned} \tag{3.15}$$

holds for $1<r<|\beta q|^{-1}$.

We now illustrate how our orthogonality relation (3.15) implies Hahn's (1.13). The idea is to observe that if a sequence of polynomials $\{p_n(x)\}$ is orthogonal on a bounded interval (a, b) with respect to a Borel measure $d\mu$ then $\int_a^b d\mu(t)/(z-t)$ is a complex weight function (Pollaczek [16]). This is so because if U is a contour containing (a, b) in its interior then

$$\begin{aligned} \frac{1}{2\pi i} \int_U p_n(z) p_m(z) \int_a^b \frac{d\mu(t)}{z-t} dz &= \int_a^b d\mu(t) \int_U \frac{p_n(z) p_m(z)}{z-t} \frac{dz}{2\pi i} \\ &= \int_a^b p_n(t) p_m(t)\, d\mu(t). \end{aligned}$$

so there is a chance our complex weight function in (3.15) differs from $\int_a^b d\mu(t)/(z-t)$ by an analytic function. To recover $\mu(t)$ from its Stieltjes transform, we use the

Perron–Stieltjes inversion formula

$$F(z)=\int_{-\infty}^{\infty}\frac{d\mu(t)}{z-t} \quad \text{implies}$$

$$\mu(t_2)-\mu(t_1)=\lim_{\varepsilon\to 0^+}\int_{t_1}^{t_2}\frac{F(t-i\varepsilon)-F(t+i\varepsilon)}{2\pi i}\,dt.$$

The candidate measure $d\mu$ for the little q-Jacobi polynomials must be purely discrete in order to make $\int_a^b d\mu(t)/(z-t)$ single valued, since the complex weight is single valued. The point masses will occur at the poles of $\int_a^b d\mu(t)/(z-t)$. The only poles of the complex weight function in (3.5) that lie within $|z|=r$ are $z=q^j, j=0,1,\cdots$. The point masses obviously equal the respective residues. The residue R_j at $z=q^j$ of the complex weight function is

$$R_j=\frac{(q^{j+1}/\alpha;q)_\infty(\alpha q^{-j};q)_\infty}{(\beta q^{j+1};q)_\infty(q^{-j};q)_j(q;q)_\infty}\lim_{z\to q^j}\frac{z-q^j}{1-q^j/z},$$

which can be simplified to

$$R_j=\frac{\alpha^j q^j(q/\alpha;q)_\infty(\alpha;q)_\infty}{(\beta q^{j+1};q)_\infty(q;q)_\infty(q;q)_j}.$$

All that is left now is to show that the complex weight function and the Stieltjes transform of the constructed measure differ by a function analytic in $|z|\leqq r$. This follows from the Mittag-Leffler expansion. Note that we could have obtained the same result simply by evaluating the contour integral in (3.15) using the calculus of residues. One reason for including the above argument is that it relates the complex weight function to the Stieltjes transform of the real measure. This will be further explored in the next section when we compute the continued fraction associated with the little q-Jacobi polynomials.

In the case of our q-Bessel polynomials, the orthogonality relation (3.15) reduces to

$$\begin{aligned}\frac{1}{2\pi i}\int_{|z|=r}&y_n(z;a|q^2)y_m(z;a|q^2)\frac{(-2qz;q)_\infty(-1/2z;q)_\infty}{(-2q^{a-1}z;q)_\infty(1/2z;q)_\infty}\,dz\\ &=\frac{(-1)^{n+1}q^{n^2}(q^{n+a-1};q)_\infty(-q^{n+1};q)_\infty}{(-q^{a+n-1};q)_\infty(q^{n+1};q)_\infty(1-q^{2n+a-1})}\delta_{mn},\end{aligned}\tag{3.16}$$

where $\frac{1}{2}<r<\frac{1}{2}q^{1-a}$, $a>1$.

Finally we come to the case $\beta=\gamma/\alpha$ and $\alpha\to 0$. It is easy to repeat the aforementioned manipulations. The only difference is that the balanced ${}_3\phi_2$ sum is replaced by the q-Vandermonde sum [17, p. 247] and the ${}_1\psi_1$ sum no longer works, so in (3.8) ρ may be replaced by $\sum_{-1}^{\infty} z^{-k-1}/(\gamma q^2;q)_k$. This new complex weight can be summed only if $\gamma=1$ and the resulting real measure agrees with the limiting case of the measure constructed by Andrews and Askey and Hahn.

4. Continued fraction and asymptotic formulas. It is known that the nth convergent of the continued fraction

$$\frac{3}{1}+\frac{3z}{\lfloor 1}+\frac{5z}{\lfloor 1}+\cdots$$

of $(e^{2/z}+1)/(e^{2/z}-1)$ is $[y_n(z)+(-1)^n y_n(-z)]/[y_n(z)-(-1)^n y_n(-z)]$ (Grosswald [10, Chapt. 8]). We now derive a basic analogue of this result via the connection between the basic Bessel polynomials and the basic Lommel polynomials.

In [13] we proved that the basic Lommel polynomials $\{R_{n,\nu}(1/z;q)\}$ are orthogonal with respect to a purely discrete measure $d\mu$ of bounded support, and that

$$\int_{-\infty}^{\infty}\frac{d\mu(t)}{z-t}=2(1-q^{\nu})J_{\nu}^{(2)}\Big(\frac{1}{z};q\Big)\Big/J_{\nu-1}^{(2)}\Big(\frac{1}{z};q\Big),\qquad z\notin\operatorname{supp}\{d\mu\}. \tag{4.1}$$

The left side of (4.1) is the continued fraction

$$\chi_{\nu}(z)=\frac{2(1-q^{\nu})}{\big|2(1-q^{\nu})z}\Big|-\frac{q^{\nu}}{\big|2(1-q^{\nu+1})z}\Big|-\frac{q^{\nu+1}}{\big|2(1-q^{\nu+2})z}\Big|-\cdots, \tag{4.2}$$

because $R_{n,\nu}(z;q)$ satisfies

$$R_{n+1,\nu}(z;q)=2(1-q^{\nu+n})z^{-1}R_{n,\nu}(z;q)-q^{n+\nu-1}R_{n-1,\nu}(z;q); \tag{4.3}$$

see [13]. In [13] we also proved that

$$\chi_{\nu}(z)=\lim_{n\to\infty}\frac{2(1-q^{\nu})R_{n-1,\nu+1}(1/z;q)}{R_{n,\nu}(1/z;q)}. \tag{4.4}$$

Using this information, (2.5), (2.9) and (2.14) we obtain the continued fraction representation

$$\begin{aligned}\chi_{\nu}(z)&=\frac{1}{\big|2(1-q^{\nu})z}\Big|+\frac{q^{\nu}}{\big|2(1-q^{\nu+1})z}\Big|+\frac{q^{\nu+1}}{\big|2(1-q^{\nu+2})z}\Big|+\cdots\\&=\frac{I_{\nu}^{(2)}(1/z;q)}{I_{\nu-1}^{(2)}(1/z;q)}.\end{aligned} \tag{4.5}$$

When we replace z by $z/(1-q)$ and let $q\to 1$, the special case $\nu=\frac{1}{2}$ of (4.5) reduces to Lambert's continued fraction

$$\frac{1}{\big|z}\Big|+\frac{1}{\big|3z}\Big|+\frac{1}{\big|5z}\Big|+\cdots=\sinh\Big(\frac{1}{z}\Big)\Big/\cosh\Big(\frac{1}{z}\Big), \tag{4.6}$$

since, (Watson [19, pp. 54, 55])

$$\Gamma\Big(\frac{1}{2}\Big)\Big(\frac{z}{2}\Big)^{1/2}I_{1/2}(z)=\sinh z\quad\text{and}\quad\Gamma\Big(\frac{1}{2}\Big)\Big(\frac{z}{2}\Big)^{1/2}I_{-1/2}(z)=\cosh z. \tag{4.7}$$

The same analogy holds in the basic case. It is easy to see that

$$\Big(\frac{z}{2}\Big)^{1/2}I_{1/2}^{(2)}(z;q)=\frac{(q^{1/2};q)_{\infty}}{(q;q)_{\infty}}\sum_{0}^{\infty}\frac{(z/2)^{2n+1}}{(q^{1/2};q^{1/2})_{2n+1}}q^{n(n+1/2)} \tag{4.8}$$

and

$$\Big(\frac{z}{2}\Big)^{1/2}I_{-1/2}^{(2)}(z;q)=\frac{(q^{1/2};q)_{\infty}}{(q;q)_{\infty}}\sum_{0}^{\infty}\frac{(z/2)^{2n}}{(q^{1/2};q^{1/2})_{2n}}q^{n(n-1/2)}, \tag{4.9}$$

which, in view of Euler's formula (Slater [17, p. 93])

$$\sum_{0}^{\infty}\frac{(-1)^n z^n}{(q;q)_n}q^{n(n-1)/2}=(z;q)_{\infty}, \tag{4.10}$$

imply

$$\Big(\frac{z}{2}\Big)^{1/2}I_{-1/2}^{(2)}(z;q)+\Big(\frac{z}{2}\Big)^{1/2}I_{1/2}^{(2)}(z;q)=\frac{(q^{1/2};q)_{\infty}}{(q;q)_{\infty}}\Big(-\frac{z}{2};q^{1/2}\Big)_{\infty} \tag{4.11}$$

and

$$\left(\frac{z}{2}\right)^{1/2} I^{(2)}_{-1/2}(z;q) - \left(\frac{z}{2}\right)^{1/2} I^{(2)}_{1/2}(z,q) = \frac{(q^{1/2};q)_\infty}{(q;q)_\infty}\left(\frac{z}{2};q^{1/2}\right)_\infty. \tag{4.12}$$

The relationships (4.11) and (4.12) lead to

$$\frac{I^{(2)}_{-1/2}(z;q^2)}{I^{(2)}_{1/2}(z;q^2)} = \frac{(-z/2;q_\infty + (z/2;q)_\infty}{(-z/2;q)_\infty - (z/2;q)_\infty}. \tag{4.13}$$

Combining (4.5) and (4.13) we obtain the following generalization of (4.6):

$$\cfrac{(1-q)z\Big|}{q} + \cfrac{(1-q^3)z\Big|}{q^3} + \cfrac{(1-q^5)z\Big|}{q^5} + \cdots = \frac{(-z^{-1};q)_\infty + (z^{-1};q)_\infty}{(-z^{-1};q)_\infty - (z^{-1};q)_\infty}. \tag{4.14}$$

We now relate the convergents of (4.14) to the Bessel polynomials. Recall that the nth convergent in $\chi_{1/2}(z)$ is $\phi_{n-1,3/2}(1/z)/\phi_{n,1/2}(1/z)$; see (2.14) and (4.4). The polynomials $\phi_{n,\nu}(z)$ are symmetric, that is, $\phi_{n,\nu}(-z) = (-1)^n \phi_{n,\nu}(z)$, as can be seen from (2.14) and (2.15). Use this information and (2.12) to get

$$\phi_{m,1/2}\left(\frac{1}{z}\right) = y_m(z|q) + (-1)^m y_m(z|q) \tag{4.15}$$

and

$$\phi_{m-1,3/2}\left(\frac{1}{z}\right) = y_m(z|q) - (-1)^m y_m(z|q). \tag{4.16}$$

This establishes

THEOREM 4.1. *The* nth *convergent of the continued fraction* (4.14) *is*

$$\frac{y_n(z/2|q^2) + (-1)^n y_n(z/2|q^2)}{y_n(z/2|q^2) - (-1)^n y_n(z/2|q^2)},$$

which converges to the right side in (4.13).

Note that $(z(1-q);q)_\infty$ tends to e^{-z} as $q \to 1^-$, as can be seen from (4.10). Consequently the right side of (4.14) is

$$\left[1 + \frac{(-z^{-1};q)_\infty}{(z^{-1};q)_\infty}\right] \cdot \left[-1 + \frac{(-z^{-1};q)_\infty}{(z^{-1};q)_\infty}\right]^{-1},$$

which is the analogue of $(e^{2z}+1)(e^{2z}-1)^{-1}$. The function $(-z;q)_\infty/(z;q)_\infty$ has the power series expansion

$$(-z;q)_\infty/(z;q)_\infty = \sum_0^\infty \frac{(-1;q)_n}{(q;q)_n} z^n \tag{4.17}$$

(Slater [17, p. 248]).

We now establish a generating function and an asymptotic formula for the polynomials $\{y_n(x|q)\}$. We shall prove the following:

THEOREM 4.2. *We have*

$$y_n(z|q^2) \sim (2z)^n (q;q^2)_\infty \left(-\frac{z}{2};q\right)_\infty \quad \text{as } n \to \infty, \tag{4.18}$$

and

$$\sum_0^\infty y_n(z|q)t^n = \sum_0^\infty \frac{(t/2z;q)_l}{(2zt;q)_{l+1}}(-2zt)^l q^{l^2/2} + t\sum_0^\infty \frac{(t/2z;q)_l}{(2zt;q)_{l+1}}(-2zt)^l q^{l(l+1)/2} \tag{4.19}$$

Proof. Formula (4.18) follows from (2.12), (2.14) and the asymptotic formula [13]

$$R_{n,\nu+1}(z;q) \sim (q;q)_\infty \left(\frac{z}{2}\right)^{-n-\nu} J_\nu^{(2)}(z;q).$$

The reader can easily fill in the details. The second relationship (4.19) follows from (2.12), (2.14) and the generating function [13]

$$\sum_0^\infty R_{n,\nu}(z;q)t^n = \sum_0^\infty \frac{(-tz/2;q)_l}{(2t/z;q)_{l+1}}\left(\frac{-2tq^\nu}{z}\right)^l q^{l(l-1)/2}.$$

We conclude the present section by computing the continued fraction associated with the little q-Jacobi polynomials. They satisfy the three-term recurrence relation

$$-xp_n(x;\alpha,\beta|q) = a_n p_{n+1}(x;\alpha,\beta|q) - (a_n+b_n)p_n(x;\alpha,\beta|q) + b_n p_{n-1}(x;\alpha,\beta|q), \qquad n>0, \tag{4.20}$$

with

$$p_0(x;\alpha,\beta|q)=1, \qquad p_1(x;\alpha,\beta|q) = 1-(1-\alpha\beta q^2)x/(1-\alpha q),$$
$$a_n = q^n \frac{(1-\alpha q^{n+1})(1-\alpha\beta q^{n+1})}{(1-\alpha\beta q^{2n+1})(1-\alpha\beta q^{2n+2})}, \qquad b_n = \frac{\alpha q^n(1-q^n)(1-\beta q^n)}{(1-\alpha\beta q^{2n})(1-\alpha\beta q^{2n+1})}. \tag{4.21}$$

The following lemma identifies the continued fraction associated with a sequence of orthogonal polynomials (Chihara [7, pp. 89–90] and Pollaczek [16]).

LEMMA 4.3 (Markoff's theorem). *If $\{p_n(x)\}$ is a sequence of polynomials satisfying*

$$p_{n+1}(x) = (A_n x + B_n)p_n(x) - C_n p_{n-1}(x),\ p_0(x)=1, \qquad p_1(x) = A_0 x + B_0,$$

with $A_n C_{n+1} \neq 0$, $n \geqq 0$ and are orthogonal with respect to a positive Borel measure $d\mu$ on a bounded interval (a, b), then the continued fraction

$$\cfrac{A_0}{A_0 z + B_0} - \cfrac{C_1}{A_1 z + B_1} - \cfrac{C_2}{A_2 z + B_2} - \cdots \tag{4.22}$$

converges to $\int_a^b d\mu(t)/(z-t)$ for $z \notin (a, b)$ provided that $\int_a^b d\mu(t) = 1$.
We now prove

THEOREM 4.4. *The continued fraction* (4.22), where

$$A_n = -\frac{1}{a_n}, \qquad B_n = 1 + \frac{b_n}{a_n}, \qquad C_n = \frac{b_n}{a_n} \tag{4.23}$$

and a_n, b_n are as in (4.21) *(with b_0 interpreted as zero), is equal to*

$$F(z) = \frac{(\alpha q;q)_\infty}{(z-1)(\alpha\beta q^2;q)_\infty} {}_2\phi_1\left(\beta q, \frac{1}{z}; \frac{q}{z}; q, q\alpha\right), \qquad z \notin [0,1]. \tag{4.24}$$

Proof. In view of Lemma 4.3 all we need to show is that $F(z)$ is the Stieltjes transform of the real measure $d\mu(t)$ normalized by $\int_{-\infty}^{\infty} d\mu(t) = 1$. Let μ be a step

function with a point mass $\lambda\alpha^j q^{j^2}(\beta q;q)_j/(q;q)_j$ at $x=q^j$, $j=0,1,\cdots$ and λ to be determined from the normalization condition. Clearly,

$$\frac{1}{\lambda}=\sum_0^\infty \frac{\alpha^j q^{j^2}(\beta q;q)_j}{(q;q)_j}=\frac{(\alpha\beta q^2;q)_\infty}{(\alpha q;q)_\infty}, \tag{4.25}$$

by the q-binomial theorem [17, p. 248]. If $G(z)$ denotes the continued fraction under consideration, then

$$G(z)=\lambda\sum_0^\infty \frac{\alpha^j q^{j^2}(\beta q;q)_j}{(q;q)_j(z-q^j)}=\frac{\lambda}{z-1}\sum_0^\infty \frac{(\beta q;q)_j(1/z;q)_j}{(q;q)_j(q/z;q)_j}\alpha^j q^j$$

$$=\frac{\lambda}{z-1}{}_2\phi_1\left(\beta q,\frac{1}{z};\frac{q}{z};q,\alpha q\right),$$

which when combined with (4.25) imply (4.24). This completes the proof.

5. Integral representations. The Hankel integral representation of $1/\Gamma(z)$ is

$$\frac{1}{\Gamma(z)}=\frac{1}{2\pi i}\int_C e^t t^{-z-1}\,dt, \tag{5.1}$$

where C consists of the lower edge of the cut (along the negative real axis) from $-\infty$ to $-\rho$, the circle $t=\rho e^{i\theta}$, $-\pi\leqq\theta\leqq\pi$ and the upper edge of the cut from $-\rho$ to $-\infty$. We now derive a similar integral representation for $1/\Gamma_q(z)$, where

$$\Gamma_q(z)=\frac{(q;q)_\infty(1-q)^{1-z}}{(q^z;q)_\infty},\qquad 0<q<1,\quad z\neq 0,-1,-2,\cdots; \tag{5.2}$$

see Askey [2]. Askey [2] gave a proof of Ramanujan's formula

$$\int_0^\infty t^{z-1}\frac{(-at;q)_\infty}{(-t;q)_\infty}\,dt=\frac{(a;q)_\infty(q^{1-z};q)_\infty}{(q;q)_\infty(aq^{-z};q)_\infty}\frac{\pi}{\sin(\pi z)},\qquad \operatorname{Re} z>0, \tag{5.3}$$

and we shall always assume $0<q<1$. Taking $a=0$ in (5.3), we see that

$$\int_0^\infty \frac{t^{z-1}\,dt}{(-t;q)_\infty}=\frac{\pi}{\sin(\pi z)}\frac{(1-q)^z}{\Gamma_q(1-z)}.$$

In the above formula replace z by $1-z$ and t by $(1-q)t$ to get

$$\frac{1}{\Gamma_q(z)}=\frac{\sin(\pi z)}{\pi}\int_0^\infty \frac{t^{-z}\,dt}{(-t(1-q);q)_\infty}. \tag{5.4}$$

The basic analogue of the series expansion of e^x, namely (Slater [17, p. 92])

$$\sum_0^\infty \frac{z^n}{(q;q)_n}=\frac{1}{(z;q)_\infty}, \tag{5.5}$$

suggests the following basic analogue of the integral in (5.1):

$$I=\frac{1}{2\pi i}\int_C \frac{t^{-x}\,dt}{(t(1-q);q)_\infty},$$

where C is the same contour in (5.1). When $0<x<1$, I can be evaluated in the

following manner:

$$I=\frac{1}{2\pi i}\int_{\infty}^{0}\frac{(te^{-i\pi})^{-x}(-dt)}{(-t(1-q);q)_{\infty}}+\frac{1}{2\pi i}\int_{0}^{\infty}\frac{(te^{i\pi})^{-x}(-dt)}{(-t(1-q);q)_{\infty}}$$
$$=\frac{\sin(\pi x)}{\pi}\int_{0}^{\infty}\frac{t^{-x}\,dt}{(-t(1-q);q)_{\infty}}=\frac{1}{\Gamma_q(x)},$$

by (5.4). The integral representation

$$\frac{1}{\Gamma_q(z)}=\frac{1}{2\pi i}\int_C\frac{t^{-z}\,dt}{(t(1-q);q)_{\infty}} \tag{5.6}$$

then follows from the identity theorem for analytic functions. The representation (5.6) is a basic analogue of Hankel's formula (5.1).

We now turn to finding integral representations for the modified basic Bessel functions. Note that

$$I_{\nu}^{(1)}(z;q)=\left(\frac{z}{2}\right)^{\nu}\sum_{0}^{\infty}\frac{(z/2)^{2k}(1-q)^{-\nu-k}}{(q;q)_k\Gamma_q(\nu+k+1)} \tag{5.7}$$

follows from (2.3), (2.5) and (5.2). We substitute the integral representation (5.6) in (5.7) and "formally" change the integration and summation processes, then change the integration variable to $t/(1-q)$ to obtain

$$I_{\nu}^{(1)}(z;q)=\frac{(z/2)^{\nu}}{2\pi i}\int_C\frac{t^{-\nu-1}}{(t;q)_{\infty}(z^2/4t;q)_{\infty}}\,dt,\qquad |\arg z|<\frac{\pi}{2}, \tag{5.8}$$

using Euler's formula (5.5). The above derivation is indeed formal because the series in (5.5) converges only for $|z|<1$ and the integral in (5.8) extends to infinity. We now prove (5.8) by showing that both sides satisfy the same second order q-difference equation and have the same behavior at the origin. Denote the right side in (5.8) by $f_{\nu}(z)$. It suffices to consider only positive z, by analytic continuation. Clearly

$$f_{\nu}(\sqrt{q}z)=q^{\nu/2}\frac{(z/2)^{\nu}}{2\pi i}\int_C\frac{t^{-\nu-1}[1-(z^2/4t)]\,dt}{(t;q)_{\infty}(z^2/4t;q)_{\infty}}$$
$$=q^{\nu/2}f_{\nu}(z)-\frac{z}{2}q^{\nu/2}f_{\nu+1}(z). \tag{5.9}$$

In the definition of $f_{\nu}(z)$ replace t by $tz^2/4$ to get

$$f_{\nu}(z)=\frac{(z/2)^{-\nu}}{2\pi i}\int_C\frac{t^{-\nu-1}\,dt}{[tz^2/4;q]_{\infty}(t^{-1};q)_{\infty}},$$

which leads to

$$f_{\nu}(\sqrt{q}z)=q^{-\nu/2}f_{\nu}(z)-\frac{z}{2}q^{-\nu/2}f_{\nu-1}(z). \tag{5.10}$$

Combining (5.9) and (5.10) leads to

$$f_{\nu}(qz)-(q^{\nu/2}+q^{-\nu/2})f_{\nu}(\sqrt{q}z)+\left(1-\frac{z^2}{4}\right)f_{\nu}(z)=0,$$

which is the same q-difference equation satisfied by $I_{\nu}^{(1)}$, $I_{-\nu}^{(1)}$, $K_{\nu}^{(1)}$, as can be seen from [13, (2.2)] and (2.5). Since $I_{-\nu}^{(1)}$ and $K_{\nu}^{(1)}$ are $O(z^{-\nu})$ as $z\to 0$, for $\nu>0$ the function

$f_\nu(z)$ must be a constant multiple of $I_\nu^{(1)}(z;q)$. The constant is unity since the limits $\lim_{z\to 0}(2/z)^\nu I_\nu^{(1)}(z;q)$ and $\lim_{z\to 0}(2/z)^\nu f_\nu(z)$ have the common value $(1-q)^{-\nu}/\Gamma_q(\nu+1)$; see (5.7) and (5.6). This completes the proof of (5.8).

We now derive two more integral representations. Assume $z>0$ and set $t=ze^w/2$. The relationship (4.8) then becomes

$$I_\nu^{(1)}(z;q)=\frac{1}{2\pi i}\int_{\infty-\pi i}^{\infty+\pi i}\frac{e^{-\nu w}\,dw}{(e^w z/2;q)_\infty(e^{-w}z/2;q)_\infty}, \tag{5.11}$$

whose domain of validity can then be extended to $|\arg z|<\pi/2$. The representation (5.11) is a generalization of Watson [19, (4), p. 176]. In (5.11) take the contour to consist of three sides of a rectangle with vertices at $\infty-i\pi$, $-\pi i$, πi and $\infty+\pi i$. Write $t\pm i\pi$ for w on the sides parallel to the real axis, and $\pm i\theta$ for w on the lines joining p to $\pm i\pi$, to obtain

(5.12)

$$I_\nu^{(1)}(z;q)=\frac{1}{\pi}\int_0^\pi\frac{\cos\nu\theta\,d\theta}{(e^{i\theta}z/2;q)_\infty(e^{-i\theta}z/2;q)_\infty}-\frac{\sin\nu\pi}{\pi}\int_0^\infty\frac{e^{-\nu t}\,dt}{(-z\,e^t/2;q)_\infty(-z\,e^{-t}/2;q)_\infty},$$

a basic analogue of [19, (4), p. 18]. Finally we have, via (2.6),

$$K_\nu^{(1)}(z;q)=\int_0^\infty\frac{\cosh\nu t\,dt}{(-z\,e^{t/2}/2;q)_\infty(-z\,e^{-t/2}/2;q)_\infty}, \tag{5.13}$$

or

$$K_\nu^{(1)}(z;q)=\frac{1}{2}\int_0^\infty\frac{e^{-\nu t}\,dt}{(-z\,e^{t/2}/2;q)_\infty(-z\,e^{-t/2}/2;q)_\infty},\qquad |\arg z|<\frac{\pi}{2}. \tag{5.14}$$

Note that similar integral representations can be established for $I_\nu^{(2)}(z;q)$, since

$$I_\nu^{(2)}(z;q)=\left(\frac{z^2}{4};q\right)_\infty I_\nu^{(1)}(z;q), \tag{5.15}$$

which follows from the definitions of $I_\nu^{(1)}$ and $I_\nu^{(2)}$, Euler's formula (5.5) or (4.10) and Slater [17, (IV.1), p. 247].

REFERENCES

1. W. H. Abdi, *A basic analogue of the Bessel polynomial*, Math. Nachr., 30 (1966), pp. 209–219.
2. W. Al-Salam and M. E. H. Ismail, *Reproducing kernels for q-Jacobi polynomials*, Proc. Amer. Math. Soc., 57 (1977), pp. 105–110.
3. G. E. Andrews and R. A. Askey, *Enumeration of partitions: the role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Ainger, ed, Reidel, Dordrecht, 1977, pp. 3–26.
4. ———, *The classical and discrete orthogonal polynomials and their q-analogues*, in preparation.
5. R. A. Askey, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.
6. J. L. Burchnal, *The Bessel polynomials*, Canad. J. Math., 3 (1951), pp. 62–68.
7. T. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.
8. M. G. DeBruin, E. Saff and R. Varga, *On the zeros of generalized Bessel polynomials*, to appear.
9. E. Grosswald, *The student's t-distribution for odd degrees of freedom is infinitely divisible*, Ann. Probab., 4 (1976), pp. 680–693.
10. ———, *The Bessel Polynomials*, Lecture Notes in Mathematics, Springer-Verlag, New York, 1978.
11. W. Hahn, *Über orthogonal polynome, die q-differenzengleichungen genügen*, Math. Nachr., 2 (1949), pp. 4–34.

12. M. E. H. ISMAIL, *Connection relations and bilinear formulas for the classical orthogonal polynomials*, J. Math. Anal. Appl., 57 (1977), pp. 487–496.
13. ———, *The zeros of basic Bessel functions, the function $J_{\nu+ax}(x)$ and associated orthogonal polynomials*, J. Math. Anal. Applications, to appear.
14. M. E. H. ISMAIL AND D. H. KELKER, *The Bessel polynomials and the student t-distribution*, this Journal, 7 (1976), pp. 82–91.
15. F. H. JACKSON, *On generalized functions of Legendre and Bessel*, Trans. Royal Soc. Edinburgh, 41 (1903), pp. 1–28.
16. F. POLLACZEK, *Sur une généralisation des polynômes de Jacobi*, Memorial des Sciences Mathematique 121, Paris, 1956.
17. L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966
18. D. STANTON, *A short proof of a generating function for Jacobi polynomials*, Proc. Amer. Math. Soc., 80 (1980), pp. 398–400.
19. G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0013 $01.00/0

A NOTE ON STRICTLY CAUSAL OPERATORS*

VACLAV DOLEZAL†

Abstract. It is shown that every linear, bounded, strictly causal operator on a resolution space $[H, P^t]$ is in fact contractive under a certain inner product which generates a norm equivalent to the original norm on H.

As is known, strictly causal operators, introduced in [1], are generalizations of Volterra operators, and play an important role in system theory. In particular, they enjoy the following property: If $A: H \to H$ is a linear, bounded, strictly causal operator, then $I-A$ possesses a causal, bounded inverse and

$$S_n = \sum_{i=0}^{n} A^i \to (I-A)^{-1}$$

as $n \to \infty$ in the uniform operator topology.

In this note we show that every linear, bounded, strictly causal operator $A: H \to H$ is a contraction provided that the inner product $\langle \cdot, \cdot \rangle_0$ on H is replaced by a certain inner product $\langle \cdot, \cdot \rangle_1$ which generates a norm $|\cdot|_1$ equivalent to the original norm $|\cdot|_0$. As a consequence, we give an estimate for the speed of convergence of S_n to $(I-A)^{-1}$ in the original operator norm.

To state the result, we introduce the following notation. If H is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\alpha$, we denote by $|\cdot|_\alpha$ the norm generated by $\langle \cdot, \cdot \rangle_\alpha$. Moreover, if $A: H \to H$ is a linear, bounded operator, we put $\|A\|_\alpha = \sup\{|Ax|_\alpha : x \in H, |x|_\alpha = 1\}$.

THEOREM. *Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_0$, let $A: H \to H$ be a linear, bounded (in norm $|\cdot|_0$) operator having zero spectral radius and let $0 < \lambda < 1$. Then:*

(i) *There exists an inner product $\langle \cdot, \cdot \rangle_1$ on H such that the generated norm $|\cdot|_1$ on H is equivalent to $|\cdot|_0$, and*

$$\|A\|_1 \leqq \lambda. \tag{1}$$

(ii) *For every integer $n \geqq 0$,*

$$\left\| (I-A)^{-1} - \sum_{i=0}^{n} A^i \right\|_0 \leqq C_\lambda \lambda^n, \tag{2}$$

where the constant C_λ is independent of n.

COROLLARY. *If $[H, P^t]$ is a Hilbert resolution space with inner product $\langle \cdot, \cdot \rangle_0$, if $A: H \to H$ is strictly causal* [1] *and if $0 < \lambda < 1$, then* (i) *and* (ii) *hold.*

Proof of the theorem. By Gel'fand's theorem [2, p. 263], $\|A^n\|_0^{1/n} \to r(A) = 0$ as $n \to \infty$. Thus, select an integer $k \geqq 1$ so that

$$\|A^k\|_0^{1/k} \leqq \lambda. \tag{3}$$

Next, for each $x, y \in H$ put

$$\begin{aligned} \langle x, y \rangle_1 = {} & \lambda^{2k-2} \langle x, y \rangle_0 + \lambda^{2k-4} \langle Ax, Ay \rangle_0 + \cdots \\ & + \lambda^2 \langle A^{k-2}x, A^{k-2}y \rangle_0 + \langle A^{k-1}x, A^{k-1}y \rangle_0. \end{aligned} \tag{4}$$

* Received by the editors December 13, 1978, and in revised form September 22, 1980.

† Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, New York 11794. This research was supported by the National Science Foundation under grant MCS78-01992.

Clearly, $\langle \cdot, \cdot \rangle_1$ is an inner product on H, and for the corresponding norm $|\cdot|_1$ we have

$$|x|_1^2 = \lambda^{2k-2}|x|_0^2 + \lambda^{2k-4}|Ax|_0^2 + \cdots + \lambda^2|A^{k-2}x|_0^2 + |A^{k-1}x|_0^2. \tag{5}$$

From (5) it follows that

$$\alpha_\lambda |x|_0 \leqq |x|_1 \leqq \beta_\lambda |x|_0 \tag{6}$$

for every $x \in H$, where

$$\alpha_\lambda = \lambda^{k-1}, \qquad \beta_\lambda = (\lambda^{2k-2} + \lambda^{2k-4}\|A\|_0^2 + \cdots + \|A^{k-1}\|_0^2)^{1/2}. \tag{7}$$

Hence, the norms $|\cdot|_1$ and $|\cdot|_0$ are equivalent.

On the other hand, (5) and (3) yield, for any $x \in H$,

$$\begin{aligned} |Ax|_1^2 &= \lambda^{2k-2}|Ax|_0^2 + \lambda^{2k-4}|A^2x|_0^2 + \cdots + \lambda^2|A^{k-1}x|_0^2 + |A^kx|_0^2 \\ &\leqq \lambda^{2k}|x|_0^2 + \lambda^{2k-2}|Ax|_0^2 + \cdots + \lambda^2|A^{k-1}x|_0^2 \\ &= \lambda^2|x|_1^2. \end{aligned}$$

Hence, $\|A\|_1 \leqq \lambda$ and claim (i) is proven.

Furthermore, (6) shows that a linear operator $M: H \to H$ is bounded in $|\cdot|_1$ if and only if M is bounded in $|\cdot|_0$, and we have

$$\alpha_\lambda \beta_\lambda^{-1} \|M\|_0 \leqq \|M\|_1 \leqq \alpha_\lambda^{-1} \beta_\lambda \|M\|_0. \tag{8}$$

However, (i) implies that, for any integer $n \geqq 0$,

$$\left\| (I-A)^{-1} - \sum_{i=0}^{n} A^i \right\|_0 \leqq \sum_{i=n+1}^{\infty} \|A\|_1^i \leqq \sum_{i=n+1}^{\infty} \lambda^i = (1-\lambda)^{-1}\lambda^{n+1}.$$

Hence, by (8),

$$\left\| (I-A)^{-1} - \sum_{i=0}^{n} A^i \right\|_0 \leqq \alpha_\lambda^{-1}\beta_\lambda \left\| (I-A)^{-1} - \sum_{i=0}^{n} A^i \right\|_1 \leqq \alpha_\lambda^{-1}\beta_\lambda (1-\lambda)^{-1}\lambda^{n+1},$$

which proves (ii).

To prove the corollary, note that aA is strictly causal for any number a whenever A is strictly causal [1, Thm. 3.1]. Thus, by virtue of [1, Thm. 4.2], the series $\sum_{i=0}^{\infty} a^i A^i$ converges in the operator norm $\|\cdot\|_0$. Hence, $r(A) = 0$, which proves the claim.

Note that under the new inner product $\langle \cdot, \cdot \rangle_1$ the operators P^t from the resolution of identity $\{P^t\}$ remain bounded projections, but are no longer orthogonal projections except for the trivial case when $A = 0$.

REFERENCES

[1] R. M. DESANTIS, *Causality, strict causality and invertibility for systems in Hilbert resolution spaces*, SIAM J. Control, 12 (1974), pp. 536–553.

[2] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0014 $01.00/0

TWO RESULTS OF RAMANUJAN*

JACQUES DUTKA†

Abstract. New proofs are given of two results of Ramanujan, one of the latter of which has attracted a good deal of attention. For series associated with these results, inverse factorial expansions which converge rapidly for large n are derived. Included here is an improvement of a result of Watson [Quart. J. Math. Oxford, 1 (1930), pp. 310–318.]

1. About a half century ago, a result stated by Ramanujan [1, p. 351],

$$\frac{1}{n}+\left(\frac{1}{2}\right)^2\frac{1}{n+1}+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2\frac{1}{n+2}+\cdots = \left\{\frac{\Gamma(n)}{\Gamma(n+1/2)}\right\}^2\left\{1+\left(\frac{1}{2}\right)^2+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2+\cdots \text{ to } n \text{ terms}\right\}, \tag{1.1}$$

aroused considerable interest among British mathematicians. Proofs were given by Watson [2], Darling [3], and generalizations by Whipple [4], Bailey [5] and [7] and Hodgkinson [6]. A discussion of this is also given by Hardy [8, pp. 106–107, 112].

In 1957, a facsimile edition of Ramanujan's Notebooks was published [9]. On pages 237 and 239 of Volume I, there is a group of related results stated without proofs, including (in a different notation) (1.1) and the complementary formula

$$\frac{1}{2n+1}+\left(\frac{1}{2}\right)^2\frac{1}{2n+3}+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2\frac{1}{2n+5}+\cdots = \frac{2}{\pi^2}\left\{\frac{\Gamma(n+1/2)}{\Gamma(n+1)}\right\}^2\left\{2G+\left[1+\left(\frac{2}{3}\right)^2+\left(\frac{2\cdot 4}{3\cdot 5}\right)^2+\cdots \text{ to } n \text{ terms}\right]\right\}, \tag{1.2}$$

where $G=0.915956\cdots$ is Catalan's constant.

It does not appear to have been noticed previously that (1.1) and (1.2) can be obtained by equating two distinct evaluations of the integral $(2/\pi)\int_0^1 k^m K\,dk$ in the particular cases where m is an odd and an even integer respectively and K is the complete elliptic integral of the first kind. This is shown in § 2. In § 3, there are derived some inverse factorial expansions for series arising in (1.1) and (1.2) which converge rapidly for large n.

2. On expanding the integrand of

$$K=\int_0^{\pi/2}\frac{d\varphi}{\sqrt{1-k^2\sin^2\varphi}}, \qquad k^2<1 \tag{2.1}$$

and integrating term by term, one finds that

$$K=\frac{\pi}{2}\left\{1+\left(\frac{1}{2}\right)^2k^2+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2k^4+\cdots\right\}$$

and

$$\frac{2}{\pi}\int_0^1 k^m K\,dk=\frac{1}{m+1}+\left(\frac{1}{2}\right)^2\frac{1}{m+3}+\left(\frac{1\cdot 3}{2\cdot 5}\right)^2\frac{1}{m+5}+\cdots. \tag{2.2}$$

* Received by the editors August 4, 1980, and in revised form October 22, 1980.
† Audits and Surveys, Inc., One Park Avenue, New York, New York 10016.

On the other hand, by Byrd and Friedman [10, no. 615.12],

$$m^2(m+1)\int_0^1 k^m K\,dk = (m^2-1)(m-1)\int_0^1 k^{m-2}K\,dk + (m+1),$$

whence the reduction formula

$$\int_0^1 k^m K\,dk = \frac{1}{m^2} + \left(\frac{m-1}{m}\right)^2 \int_0^1 k^{m-2}K\,dk \tag{2.3}$$

follows.

At this point, it is convenient to consider the cases where m is even and m is odd separately. By Byrd and Friedman [10, no. 616.03], $\int_0^1 kK\,dk = 1$ Thus, from (2.2), for $n=1$, (1.1) holds. For $m=2n-1$, $n>1$, one gets, on applying (2.3) repeatedly,

$$\int_0^1 k^{2n-1}K\,dk = \frac{1}{(2n-1)^2} + \left[\frac{(2n-2)}{(2n-1)(2n-3)}\right]^2 + \left[\frac{(2n-2)(2n-4)}{(2n-1)(2n-3)(2n-5)}\right]^2$$
$$+\cdots+\left[\frac{(2n-2)(2n-4)\cdots(4)}{(2n-1)(2n-3)\cdots(3)}\right]^2 + \left[\frac{(2n-2)(2n-4)\cdots(2)}{(2n-1)(2n-3)\cdots(3)}\right]^2.$$

On rewriting the terms on the right-hand side in reverse order, one gets

$$\int_0^1 k^{2n-1}K\,dk = \left[\frac{(2n-2)(2n-4)\cdots(2)}{(2n-1)(2n-3)\cdots(3)}\right]^2$$
$$\times\left\{1+\left(\frac{1}{2}\right)^2+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2+\cdots+\left[\frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots(2n-2)}\right]^2\right\}$$

or

$$\frac{2}{\pi}\int_0^1 k^{2n-1}K\,dk = \frac{1}{2}\left\{\frac{\Gamma(n)}{\Gamma(n+1/2)}\right\}^2\left\{1+\left(\frac{1}{2}\right)^2+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2+\cdots\right.$$
$$\left.+\left[\frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots(2n-2)}\right]^2\right\}. \tag{2.4}$$

In (2.2), substitute $m=2n-1$ and compare with (2.4). Then (1.1) follows.

Similarly, for $m=2n$ one gets, on applying (2.3) repeatedly,

$$\int_0^1 k^{2n}K\,dk = \frac{1}{(2n)^2} + \left[\frac{(2n-1)}{(2n)(2n-2)}\right]^2 + \left[\frac{(2n-1)(2n-3)}{(2n)(2n-2)(2n-4)}\right]^2$$
$$+\cdots+\left[\frac{(2n-1)(2n-3)\cdots(3)}{(2n)(2n-2)\cdots(2)}\right]^2 + \left[\frac{(2n-1)(2n-3)\cdots(1)}{(2n)(2n-2)\cdots(2)}\right]^2\cdot\int_0^1 K\,dk.$$

By Byrd and Friedman [9, no. 615.01], $\int_0^1 K\,dk = 2G$. Thus on substituting this and proceeding as above, one gets

$$\int_0^1 k^{2n}K\,dk = \left[\frac{(2n-1)(2n-3)\cdots(1)}{(2n)(2n-2)\cdots(2)}\right]^2$$
$$\times\left\{2G+1+\left(\frac{2}{3}\right)^2+\left(\frac{2\cdot 4}{3\cdot 5}\right)^2+\cdots+\left[\frac{2\cdot 4\cdots(2n-2)}{3\cdot 5\cdots(2n-1)}\right]^2\right\}$$

or

$$\frac{2}{\pi}\int_0^1 k^{2n}K\,dk=\frac{2}{\pi^2}\left\{\frac{\Gamma(n+1/2)}{\Gamma(n+1)}\right\}^2\left\{2G+1+\left(\frac{2}{3}\right)^2+\left(\frac{2\cdot 4}{3\cdot 5}\right)^2+\cdots+\left[\frac{2\cdot 4\cdots(2n-2)}{3\cdot 5\cdots(2n-1)}\right]^2\right\}. \tag{2.5}$$

In (2.2), substitute $m=2n$ and compare with (2.5). Then (1.2) follows.

3. Expansions for series arising in (1.1) and (1.2) respectively,

$$S_n=1+\left(\frac{1}{2}\right)^2+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2+\cdots+\left(\frac{1\cdot 3\cdots(2n-3)}{2\cdot 4\cdots(2n-2)}\right)^2 \tag{3.1}$$

and

$$T_n=1+\left(\frac{2}{3}\right)^2+\left(\frac{2\cdot 4}{3\cdot 5}\right)^2+\cdots+\left(\frac{2\cdot 4\cdots(2n-2)}{3\cdot 5\cdots(2n-1)}\right)^2, \tag{3.2}$$

which are useful for large n, will now be obtained.

On rewriting S_n, one finds

$$S_n=\frac{1}{\pi}\sum_{m=0}^{n-1}\left[\frac{\Gamma(m+1/2)}{\Gamma(m+1)}\right]^2,\qquad n\geqq 1,$$

and by Nielsen [11, p. 288],

$$\left[\frac{\Gamma(x)}{\Gamma(x+1/2)}\right]^2=\frac{1}{x}F\left(\frac{1}{2},\frac{1}{2},x+1;1\right),\qquad R(x)>0, \tag{3.3}$$

where $F(a,b,c;z)$ is the hypergeometric function. Thus, on substituting, one gets

$$S_n=\frac{1}{\pi}\sum_{m=0}^{n-1}\frac{1}{m+1/2}F\left(\frac{1}{2},\frac{1}{2},m+\frac{3}{2};1\right),\qquad n\geqq 1,$$

where

$$\begin{aligned}&\frac{1}{m+1/2}F\left(\frac{1}{2},\frac{1}{2},m+\frac{3}{2};1\right)\\&=\frac{1}{m+1/2}+\left(\frac{1}{2}\right)^2\frac{1}{1!(m+1/2)(m+3/2)}+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!(m+1/2)(m+3/2)(m+5/2)}+\cdots\\&=\frac{1}{m+1/2}-\left[\left(\frac{1}{2}\right)^2\frac{1}{1!1}\Delta\left(\frac{1}{m+1/2}\right)+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!2}\Delta\left(\frac{1}{(m+1/2)(m+3/2)}\right)+\cdots\right]\end{aligned}$$

and $\Delta(f(m))=f(m+1)-f(m)$. On summing this series, one finds that

$$\pi S_n=\sum_{m=0}^{n-1}\frac{1}{m+1/2}+\left[\left(\frac{1}{2}\right)^2\frac{1}{1!1(m+1/2)}+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!2(m+1/2)(m+3/2)}+\left(\frac{1\cdot 3\cdot 5}{2\cdot 2\cdot 2}\right)^2\frac{1}{3!3(m+1/2)(m+3/2)(m+5/2)}+\cdots\right]\Bigg|_{m=n}^{m=0},$$

$$\pi S_n = \sum_{m=0}^{n-1} \frac{1}{m+1/2} + \left[\left(\frac{1}{2}\right)\frac{1}{1} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{1}{2} + \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)\frac{1}{3} + \cdots\right]$$
$$- \left[\frac{1}{2}\,\frac{1}{1}\,\frac{1}{2n+1} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{1\cdot 3}{2(2n+1)(2n+3)}\right.$$
$$\left. + \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)\frac{1\cdot 3\cdot 5}{3(2n+1)(2n+3)(2n+5)} + \cdots\right].$$

By Nielsen [11, p. 15, (2) and (4)],

$$\sum_{m=0}^{n-1} \frac{1}{m+1/2} = \Psi\left(n+\frac{1}{2}\right) - \Psi\left(\frac{1}{2}\right) = \Psi\left(n+\frac{1}{2}\right) + \gamma + 2\ln 2,$$

where $\Psi(z) = d\ln\Gamma(z)/dz$ and $\gamma = 0.577215\cdots$ is Euler's constant. By Jolley [12, no. (262)],

$$\left(\frac{1}{2}\right)\frac{1}{1} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{1}{2} + \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)\frac{1}{3} + \cdots = 2\ln 2.$$

Thus

$$S_n = \frac{1}{\pi}\left\{\Psi\left(n+\frac{1}{2}\right) + \gamma + 4\ln 2\right.$$
$$\left. - \left[\left(\frac{1}{2}\right)\frac{1}{1(2n+1)} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{1\cdot 3}{2(2n+1)(2n+3)} + \cdots\right]\right\}, \qquad n \geqq 1, \tag{3.4}$$

where, by Nörlund [13, p. 42],

$$\Psi(x+1) = \log(x+\alpha) + \sum_{s=1}^{\infty} \frac{(-1)^s}{s}\,\frac{B_s^{(s)}(\alpha)}{(x+1)(x+2)\cdots(x+s)}, \qquad R(x+\alpha) > 0, \tag{3.5}$$

and the coefficients $B_s^{(s)}(\alpha)$ are generated by the relation

$$\frac{t(1+t)^{\alpha-1}}{\log(1+t)} = \sum_{s=0}^{\infty} \frac{t^s}{s!} B_s^{(s)}(\alpha).$$

(For the cases $\alpha = 0$ and $\alpha = 1$, the series in (3.5) were obtained by J. Binet. The coefficients $B_s^{(s)}(\alpha)$, which arise in interpolation formulas, etc., have been tabulated for various values of s and α.)

This result for S_n is an improvement over that of Watson [14, pp. 314–315], who, using a different method, obtained the equivalent of the asymptotic expansion

$$S_n \sim \frac{1}{\pi}\left[\log n + \gamma + 4\ln 2 - \frac{1}{4n} + \frac{5}{192n^2} + \cdots\right]. \tag{3.6}$$

An expansion for T_n in (3.2) will be obtained similarly:

$$T_n = \frac{\pi}{4}\sum_{m=1}^{n}\left[\frac{\Gamma(m)}{\Gamma(m+1/2)}\right]^2 = \frac{\pi}{4}\sum_{m=1}^{n}\frac{1}{m}F\left(\frac{1}{2}, \frac{1}{2}, m+1; 1\right),$$

where

$$\frac{1}{m}F\left(\frac{1}{2}, \frac{1}{2}, m+1; 1\right) = \frac{1}{m} - \left[\left(\frac{1}{2}\right)^2\frac{1}{1!1}\Delta\left(\frac{1}{m}\right) + \left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1\Delta}{2!2}\left(\frac{1}{m(m+1)}\right) + \cdots\right].$$

On summing this series, one finds that

$$\frac{4}{\pi}T_n = \sum_{m=1}^{n}\frac{1}{m}+\left[\left(\frac{1}{2}\right)^2\frac{1}{1!1}\frac{1}{m}+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!2}\frac{1}{m(m+1)}\right.$$
$$\left.+\left(\frac{1\cdot 3\cdot 5}{2\cdot 2\cdot 2}\right)^2\frac{1}{3!3}\frac{1}{m(m+1)(m+2)}+\cdots\right]\Bigg|_{m=n+1}^{m=1},$$

$$\frac{4}{\pi}T_n = \sum_{m=1}^{n}\frac{1}{m}+\left[\left(\frac{1}{2}\right)^2\frac{1}{1}+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2\frac{1}{2}+\left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2\frac{1}{3}+\cdots\right]$$
$$-\left[\left(\frac{1}{2}\right)^2\frac{1}{1!1(n+1)}+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!2(n+1)(n+2)}\right.$$
$$\left.+\left(\frac{1\cdot 3\cdot 5}{2\cdot 2\cdot 2}\right)^2\frac{1}{3!3(n+1)(n+2)(n+3)}+\cdots\right].$$

By Nielsen [11, p. 15, (2) and (5)],

$$\sum_{m=1}^{n}\frac{1}{m} = \Psi(n+1)-\Psi(1) = \Psi(n+1)+\gamma.$$

By Jolley [11, no. (385)],

$$\left(\frac{1}{2}\right)^2\frac{1}{1}+\left(\frac{1\cdot 3}{2\cdot 4}\right)^2\frac{1}{2}+\left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2\frac{1}{3}+\cdots = 4\ln 2-\frac{8}{\pi}G.$$

Thus

$$T_n = \frac{\pi}{4}\left\{\Psi(n+1)+\gamma+4\ln 2-\frac{8}{\pi}G\right.$$
$$\left.-\left[\left(\frac{1}{2}\right)^2\frac{1}{1!1(n+1)}+\left(\frac{1\cdot 3}{2\cdot 2}\right)^2\frac{1}{2!2(n+1)(n+2)}+\cdots\right]\right\} \tag{3.7}$$

where $\Psi(n+1)$ can be obtained from (3.5).

REFERENCES

[1] G. H. HARDY ET AL., eds., *Collected Papers of Srinivasa Ramanujan*, Cambridge University Press, 1927, reprinted by Chelsea, New York, 1962.

[2] G. N. WATSON, *Theorems stated by Ramanujan (VIII): theorems on divergent series*, J. London Math. Soc., 4 (1929), pp. 82–86.

[3] H. B. C. DARLING, *On a proof of one of Ramanujan's theorems*, J. London Math. Soc., 5 (1930), pp. 8–9.

[4] F. J. W. WHIPPLE, *The sum of the coefficients of a hypergeometric series*, J. London Math. Soc., 5 (1930), p. 192.

[5] W. N. BAILEY, *The partial sum of the coefficients of the hypergeometric series*, J. London Math. Soc., 6 (1931), pp. 40–41.

[6] J. HODGKINSON, *Note on one of Ramanujan's theorems*, J. London Math. Soc., 6 (1931), pp. 42–43.

[7] W. N. BAILEY, *On one of Ramanujan's theorems*, J. London Math. Soc., 7 (1932), pp. 34–36.

[8] G. H. HARDY, *Ramanujan . . .*, Cambridge University Press, 1940, reprinted by Chelsea, New York.

[9] *Notebooks of Srinivasa Ramanujan*, two volumes, Tata Institute of Fundamental Research, Bombay, 1957.

[10] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Scientists*, second edition, revised, Springer-Verlag, New York, 1971.

[11] N. NIELSEN, *Handbuch der Gammafunktion*, 1906, reprinted by Chelsea, New York, 1965.

[12] L. B. W. JOLLEY, *Summation of Series*, second revised edition, Dover, New York, 1961.
[13] N. E. NÖRLUND, *Sur les valeurs asymptotiques des nombres et des polynômes de Bernoulli*, Rend. Circ. Mat. Palermo, Serie II, X (1961), pp. 27–44.
[14] G. N. WATSON, *The constants of Landau and Lebesgue*, Quart. J. Math., Oxford Ser., 1 (1930), pp. 310–318.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

0036-1410/81/1203-0015 $01.00/0

AN APPLICATION OF GLICKSBERG'S THEOREM TO SET-VALUED INTEGRAL EQUATIONS ARISING IN THE THEORY OF THERMOSTATS*

K. GLASHOFF AND J. SPREKELS†

Abstract. In this paper we model a heat conduction problem arising in the theory of temperature regulation by thermostats. The corresponding mathematical problem consists of a parabolic initial-boundary value problem with nonlinear discontinuous boundary conditions. The problem is transformed into an equivalent set-valued integrodifferential equation, a solution of which is shown to exist by an application of a theorem due to Glicksberg [Proc. Amer. Math. Soc., 3 (1952), pp. 170–174].

1. Introduction. Many heat conduction processes are regulated by a thermostat, i.e., by a temperature-regulated switch that electrically activates the burner of a heater, such as that used for the central heating of a building.

In a recent paper [2] the authors discussed a one-dimensional model consisting of a rod which is heated at both ends. The thermostat responded to the temperature at the midpoint of the rod. In this paper we generalize the results of [2] to general N-dimensional domains; from the practical point of view we think of a solid heated at its surface. The thermostat responds to the temperature measured at a fixed point of the solid. The corresponding mathematical model is a parabolic initial-boundary value problem with discontinuous, nonlinear boundary conditions.

In § 2 we formulate the model and the equations involved. Moreover, we motivate the transformation into an equivalent set-valued integrodifferential equation of the form

$$y(t) \in \int_0^t K(t,\tau) F(y(\tau), \dot{y}(\tau))\, d\tau, \qquad 0 \leqq t \leqq T. \tag{1.1}$$

Here $F: \mathbb{R}^2 \to 2^{\mathbb{R}}$ denotes a set-valued function.

In § 3 we prove an extension of an existence result established in [2] for systems of set-valued integral equations of the form

$$y(t) \in \int_0^T K(t,\tau) F(\tau, y(\tau))\, d\tau, \qquad 0 \leqq t \leqq T, \quad y(\tau) \in \mathbb{R}^n. \tag{1.2}$$

Since more general kernels $K(t,\tau)$ are admitted than in [2], we are forced to use Glicksberg's fixed-point theorem as main tool instead of Bohnenblust–Karlin's (see [2]).

Section 4, the main section of this paper, brings a detailed study of the thermostat equation (1.1). Since the kernel K of the thermostat equation turns out to have exactly the properties needed for the application of the results of § 3, an existence result for the thermostat problem can be established.

We remark at this point that numerical calculations (see [2] for the one-dimensional case) suggest that the temperature distribution approaches a periodic state very rapidly. A proof of a corresponding result has not yet been given.

2. The mathematical model. The thermostat in question is assumed to have the following characteristic. There exist upper (lower) threshold values $\tau_0(\tau_1)$ for the

* Received by the editors February 8, 1980 and in revised form September 15, 1980.
† Universität Hamburg, Institut für Angewandte Mathematik, Hamburg, West Germany.

temperature such that the burner of the heater is switched

—"off", if the temperature rises and exceeds τ_0,
—"on", if the temperature falls and drops below τ_1.

This behavior can be interpreted as a hysteresis; see Fig. 1.

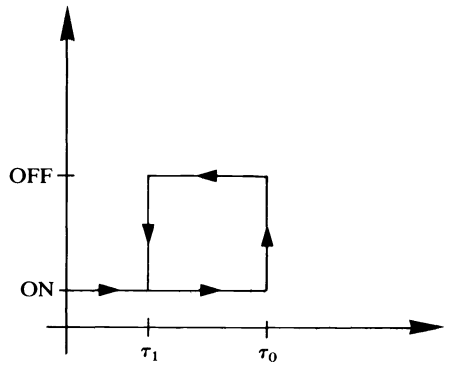
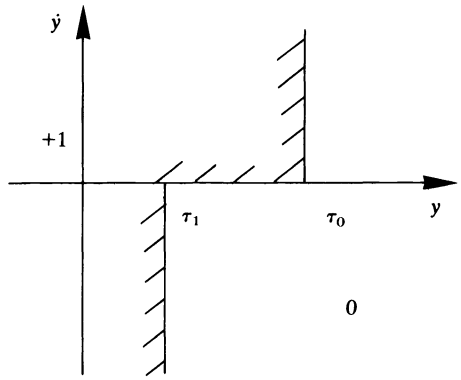


FIG. 1. *Hysteresis loop of the thermostat.*

Let $y(t)$ denote the temperature of the surrounding medium. Then the characteristic may be described by a function $f(y(t), \dot{y}(t))$, the behavior of which can be read from the phase-plane picture in Fig. 2.



$$f(y, \dot{y}) = \begin{cases} +1 & \textit{in the hatched region,} \\ 0 & \textit{elsewhere.} \end{cases}$$

FIG. 2

Analytically, f is given by

$$f(y, \dot{y}) = \tfrac{1}{4}(1 - \operatorname{sgn}(y - \tau_1))(1 - \operatorname{sgn} \dot{y}) + \tfrac{1}{4}(1 - \operatorname{sgn}(y - \tau_0))(1 + \operatorname{sgn} \dot{y}). \tag{2.1}$$

The discontinuity of f would give rise to a discontinuous single-valued integrodifferential operator. In order to overcome this discontinuity, we switch from the single-valued f

to the set-valued convexification of f; i.e., we introduce the function $F: \mathbb{R}^2 \to 2^{\mathbb{R}}$:

$$F(y, \dot{y}) = \begin{cases} [0, 1] & \text{whenever} \begin{cases} y = \tau_1, & \dot{y} \leqq 0, \\ \tau_1 < y < \tau_0, & \dot{y} = 0, \\ y = \tau_0, & \dot{y} \geqq 0, \end{cases} \\ \{f(y, \dot{y})\} & \text{otherwise.} \end{cases} \tag{2.2}$$

We now formulate our mathematical model. Let $\Omega \subset \mathbb{R}^N$ be a bounded, open domain, where $N \geqq 1$ is an integer. We assume that $\partial\Omega$ is a C^∞-manifold.

For fixed $T > 0$ and arbitrary $t \in (0, T)$ we set

$$G := \Omega \times (0, T), \quad \Gamma := \partial\Omega \times (0, T), \quad \Gamma_t := \partial\Omega \times (0, t). \tag{2.3}$$

The heat conduction process in Ω is governed by a parabolic equation of the form

$$\frac{\partial}{\partial t} y(x, t) - Ly(x, t) = 0 \quad \text{in } G, \tag{2.4}$$

where L denotes a symmetric, uniformly elliptic operator of second order on Ω:

$$Ly(x) := \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial y}{\partial x_j}(x)\right) + a(x)y(x),$$

$$a, a_{ij} \in C^\infty(\bar{\Omega}), \qquad a_{ij} = a_{ji}, \qquad i, j = 1, \cdots, N,$$

$$\sum_{i,j=1}^{N} a_{ij}(x) p_i p_j \geqq c_0 |p|^2, \qquad c_0 > 0.$$

For the initial temperature distribution we assume for convenience that

$$y(x, 0) = 0, \qquad x \in \bar{\Omega}. \tag{2.5}$$

The heating process is modeled by a third boundary condition of the form

$$By(\xi, t) = g(\xi)u(t), \qquad (\xi, t) \in \Gamma. \tag{2.6}$$

Here we have set (with a fixed constant $\alpha > 0$)

$$By(\xi) := \alpha\, \partial y(\xi) + y(\xi), \qquad \xi \in \partial\Omega.$$

If $n(\xi) = (n_1(\xi), \cdots, n_N(\xi))$ denotes the outer normal at $\xi \in \partial\Omega$, the outer conormal derivative $\partial y(\xi)$ at ξ is given by

$$\partial y(\xi) := \sum_{i,j=1}^{N} n_i(\xi) a_{ij}(\xi) \frac{\partial y}{\partial x_j}(\xi).$$

Finally, the function $g \in L^\infty(\partial\Omega)$ is assumed to be nonnegative on $\partial\Omega$ and positive on a subset of $\partial\Omega$ with positive measure. We assume that $\|g\|_\infty \leqq 1$.

In terms of our model, the solid Ω is heated at its surface; the time-dependent "temperature input" $u(t)$ is "distributed" by $g(\xi)$ over the whole surface $\partial\Omega$. For given "temperature-input-functions" $u(t)$ the system (2.4)–(2.6) is a well-posed parabolic initial-boundary value problem (in suitable function spaces).

We now have to take into account, however, that $u(t)$ is controlled by a thermostat. Let the thermostat be positioned at the (fixed) point $\bar{x} \in \Omega$. We now describe how $u(t)$ is determined by a heater. Here we choose a simple model to describe the relationship between the control u and fuel supply v (see [5] for the open-loop case). To this end, let $v(t)$, $0 \leqq v(t) \leqq 1$ denote the supply of fuel. We assume that with some fixed constant

$\beta > 0$ we have

$$\beta \dot{u}(t) + u(t) = v(t), \qquad t > 0, \quad u(0) = 0 \tag{2.7}$$

("$\cdot$" denotes differentiation with respect to t). The whole analysis of this paper remains valid if (2.7) is replaced by a higher order linear differential equation, $Du = v$, with appropriate boundary conditions.

The supply of fuel is controlled by the thermostat which in turn responds to the temperature $y(\bar{x}, t)$. Hence we obtain, with f as given by (2.1),

$$v(t) = f(y(\bar{x}, t), \dot{y}(\bar{x}, t)), \qquad t \geqq 0. \tag{2.8}$$

We thus have arrived at the system (2.4)–(2.8) as our final model.

As motivation for the following sections we briefly indicate how the given system can be transformed into a nonlinear integrodifferential equation; for details we refer to § 4. Let $K(x, t-\tau)$ denote the Green's kernel of the initial-boundary value problem (2.4)–(2.6); that is, let

$$y(x, t) = \int_0^t K(x, t-\tau) u(\tau)\, d\tau. \tag{2.9}$$

By (2.7) we obtain

$$u(\tau) = \int_0^\tau k(\tau - s) v(s)\, ds, \tag{2.10}$$

where k can easily be determined. A convolution of (2.9) and (2.10) yields for $y(t) := y(\bar{x}, t)$ an integral equation of the form

$$y(t) = \int_0^t G(t, \tau) f(y(\tau), \dot{y}(\tau))\, d\tau. \tag{2.11}$$

The integral operator on the right-hand side is not continuous which forces us to change to its set-valued analogue as induced by the mapping F. Such integral equations will be discussed in the next section. In order to stress the general structures we restrict our attention to a more special case. The corresponding results for the thermostat equation are established in § 4.

3. Nonlinear set-valued integral equations. Let $T > 0$ be fixed. We consider the system

$$y_i(t) \in \int_0^T G_i(t, \tau) F_i(\tau, y_1(\tau), \cdots, y_n(\tau))\, d\tau, \qquad i = 1, \cdots, n,$$

or for short,

$$y(t) \in \int_0^T G(t, \tau) F(\tau, y(\tau))\, d\tau. \tag{3.1}$$

Here $F: \mathbb{R}^{n+1} \to 2^{\mathbb{R}^n}$ is a set-valued mapping, and we assume that the bounded operators $\Delta_i : L^\infty[0, T] \to C[0, T]$,

$$(\Delta_i u)(t) := \int_0^T G_i(t, \tau) u(\tau)\, d\tau, \qquad i = 1, \cdots, n \tag{3.2}$$

satisfy the hypothesis

(H1) Whenever $\{v_m\}_{m \in \mathbb{N}} \subset L^\infty[0, T]$ is a bounded sequence with $v_m \overset{*}{\rightharpoonup} v$ for some $v \in L^\infty[0, T]$, then $\Delta_i v_m \rightharpoonup \Delta_i v$ in $C[0, T]$.

Let $C_n[0, T]$ denote the space of continuous n-vector functions on $[0, T]$; $L_n^\infty[0, T]$ is the space of essentially bounded measurable n-vector functions. The norm of either space is denoted by $\|\cdot\|_\infty$.

By a solution of (3.1) we mean a function $y \in C_n[0, T]$ such that there is a $v \in L^\infty[0, T]$ with

$$y(t) = \int_0^T G(t, \tau)v(\tau)\, d\tau, \qquad 0 \leqq t \leqq T, \tag{3.3a}$$

$$v(\tau) \in F(\tau, y(\tau)) \quad \text{a.e. (almost everywhere) on } [0, T]. \tag{3.3b}$$

Now let $\Delta: L_n^\infty[0, T] \to C_n[0, T]$, $\Delta := (\Delta_1, \cdots, \Delta_n)$, be defined by (3.2), and let

$$A := \Delta\phi, \tag{3.4}$$

where $\phi: C_n[0, T] \to 2^{L_n^\infty[0,T]}$ is defined by

$$v \in \phi(y) \quad \text{iff} \quad v(\tau) \in F(\tau, y(\tau)) \quad \text{a.e.} \tag{3.5}$$

Then we may rewrite (3.1) as a fixed-point equation for the set-valued operator A,

$$y \in Ay. \tag{3.6}$$

We want to apply the following result:

THEOREM 3.1 (Glicksberg [4]). *Let X be a locally convex topological vector space, and let $M \subset X$ be nonempty, convex and compact. Let the mapping $A: M \to 2^M$ satisfy the assumptions*:

(3.7) *Ay is nonempty, convex and closed for every $y \in M$.*

(3.8) *The graph $G(A) := \{(y, \xi): y \in M, \xi \in Ay\}$ of A is closed.*

Then there exists a $y \in M$ such that $y \in Ay$.

Theorem 3.1 yields as consequence:

COROLLARY 3.2. *Let $\Delta: L_n^\infty[0, T] \to C_n[0, T]$ be as above, and let ϕ satisfy the assumptions*:

(3.9) *$\phi(y) \subset L_n^\infty[0, T]$ is for every $y \in C_n[0, T]$ a nonempty and convex set.*

(3.10) *$\|v\|_\infty \leqq C$, whenever $v \in \phi(y)$ and $y \in C_n[0, T]$ with a fixed constant $C > 0$.*

(3.11) *ϕ is weakly closed; i.e., the graph of ϕ is closed with respect to the weak-star topology of $L_n^\infty[0, T]$ and the weak topology of $C_n[0, T]$.*

Then there is a $y \in C_n[0, T]$ such that $y \in \Delta\phi(y)$.

Proof. Let X be the space $C_n[0, T]$ equipped with the weak topology. We set $M := \Delta(K)$, where $K := \{v \in L_n^\infty[0, T]: \|v\|_\infty \leqq C\}$. By the linearity of Δ, M is convex. Moreover, K is weak-star sequentially compact in $L_n^\infty[0, T]$ and bounded. By hypothesis (H1), $\Delta(K)$ is compact with respect to the weak topology of $C_n[0, T]$. (3.10) yields $AM \subset M$, and by (3.9) Ay is convex and nonempty for every $y \in M$.

According to Theorem 3.1, we are left to show that the graph of A is closed. To this end, let $\{y_\lambda\}_{\lambda \in \Lambda}$ and $\{\xi_\lambda\}_{\lambda \in \Lambda}$ be Moore–Smith sequences in $C_n[0, T]$ such that $\xi_\lambda \in Ay_\lambda$, $\lambda \in \Lambda$, and $\xi_\lambda \to \xi$, $y_\lambda \to y$, for some $\xi, y \in C_n[0, T]$.

Then $\xi_\lambda = \Delta v_\lambda$, with suitable $v_\lambda \in \phi(y_\lambda)$, $\lambda \in \Lambda$. By (3.10) $\{v_\lambda\}_{\lambda \in \Lambda}$ is bounded in $L_n^\infty[0, T]$. Hence there exists a subsequence $\{v_m\}_{m \in \mathbb{N}}$ of $\{v_\lambda\}_{\lambda \in \Lambda}$ such that $v_m \to^* v$ for some $v \in L_n^\infty[0, T]$. By (3.11) we have $v \in \phi(y)$, and (H1) implies $\Delta v_m \to \Delta v$ in $C_n[0, T]$. The uniqueness of the weak limit yields $\xi = \Delta v$, i.e., $\xi \in Ay$. □

We now assume:

(H2) The set-valued mapping $F: [0, T]\times\mathbb{R}^n \to 2^{\mathbb{R}^n}$ satisfies the conditions:

(i) $F(t, z)\subset\mathbb{R}^n$ is nonempty, closed and convex, for every $(t, z)\in [0, T]\times\mathbb{R}^n$.

(ii) F is upper semicontinuous: To every $y=(t, z)\in[0, T]\times\mathbb{R}^n$ and every open set $\mathcal{O}$ with $F(y)\subset\mathcal{O}$ there is a neighborhood $U(y)$ of y such that $F(U(y)\cap([0, T]\times\mathbb{R}^n))\subset\mathcal{O}$.

(iii) $|v|\leqq M$, with a fixed $M>0$, whenever $v\in F(t, z)$ for some $(t, z)\in[0, T]\times\mathbb{R}^n$ ($|\cdot|$ denotes the maximum-norm in $\mathbb{R}^n$).

In [2, Lemma 3.4] we have shown that the operator ϕ as given by (3.5) satisfies (3.9) under the assumption (H2); (3.10) is an easy consequence of (H2, iii). Hence in order to apply Corollary 3.2 to the given situation, we are left to show that ϕ is weakly closed. We need a result stated in [2]:

LEMMA 3.3. *Let* (H2) *be satisfied. Then for every* $y=(t, z)\in[0, T]\times\mathbb{R}^n$ *it follows that*

$$F(y)=\bigcap_{\delta>0}\overline{\text{co}}\,F(U_\delta(y)\cap([0, T]\times\mathbb{R}^n)). \tag{3.12}$$

(*Here, for* $\Omega\subset\mathbb{R}^n$, *the closed convex hull of* Ω *is denoted by* $\overline{\text{co}}\,\Omega$).

Proof. See [2, Lemma 3.4]. □

THEOREM 3.4. *Let* (H2) *be satisfied. Then the operator* ϕ *as given by* (3.5) *is a weakly closed operator*; *i.e.*, ϕ *satisfies* (3.11).

Proof. Let $\{y_\lambda\}_{\lambda\in\Lambda}\subset C_n[0, T]$ be a Moore–Smith sequence such that $y_\lambda\rightharpoonup y$ for some $y\in C_n[0, T]$; moreover, let $v_\lambda\in\phi(y_\lambda)$, $\lambda\in\Lambda$ and $v_\lambda\rightharpoonup^* v$ for some $v\in L_n^\infty[0, T]$. We have to show that $v\in\phi(y)$. Since $\{v_\lambda\}$ is a bounded subset of $L_n^\infty[0, T]$, a subsequence v_m, $m\in\mathbb{N}$, of $\{v_\lambda\}$ converges in the weak-star topology of $L_n^\infty[0, T]$. By the uniqueness of the limit, $v_m\rightharpoonup^* v$. Let y_m be the element of $\{y_\lambda\}$ associated with v_m, $m\in\mathbb{N}$.

Let $\delta>0$ be given, and let $t\in(0, T)$ be fixed. Since $y_m\rightharpoonup y$, $y_m(s)\to y(s)$ for every $s\in[0, T]$. Hence there is an $n_0(t, \delta)$ such that

$$|y_m(t)-y(t)|<\frac{\delta}{2} \tag{3.13}$$

whenever $m\geqq n_0(t, \delta)$. Now y_m is uniformly continuous on $[0, T]$, which implies that there is an $h_0(m, \delta)$ such that

$$|y_m(t+h)-y_m(t)|<\frac{\delta}{2},\qquad m\geqq 1, \tag{3.14}$$

whenever $0<|h|<h_0(m, \delta)$, $t+h\in[0, T]$. Without loss of generality we may assume that

$$0<h_0(m, \delta)<\delta \quad \text{for all } \delta>0 \text{ and all } m, \tag{3.15}$$

$$h_0(m, \delta)\searrow 0 \quad \text{for } m\to\infty, \quad \delta>0 \text{ fixed}, \tag{3.16}$$

$$t\pm h_0(m, \delta)\in[0, T]. \tag{3.17}$$

(3.13) and (3.14) yield

$$|y_m(t+h)-y(t)|<\delta$$

whenever $m\geqq n_0(t, \delta)$ and $0<|h|<h_0(m, \delta)$. Hence, for such m and h,

$$v_m(t+h)\in F(t+h, y_m(t+h))\subset F(t+h, U_\delta(y(t)))\subset F(U_\delta(t, y(t))). \tag{3.18}$$

For fixed $\delta > 0$, let $h_m := h_0(m, \delta)$. From standard arguments we can conclude from (3.18) that

$$\frac{1}{2h_m} \int_{t-h_m}^{t+h_m} v_m(s)\, ds \in \overline{\text{co}}\, F(U_\delta(t, y(t))). \tag{3.19}$$

From $v_m \rightharpoonup^* v$ we obtain for $m \to \infty$

$$v(t) \in \overline{\text{co}}\, F(U_\delta(t, y(t))) \quad \text{a.e. on } [0, T];$$

hence

$$v(t) \in \bigcap_{\delta > 0} \overline{\text{co}}\, F(U_\delta(t, y(t))).$$

Lemma 3.3 proves the assertion. □

Remark. Theorem 3.4 seems to be of some interest by itself. It should be noted that the same proof works if we merely assume that $\{y_\lambda\}_{\lambda \in \Lambda}$ converges pointwise, instead of weakly, to y.

Combining the results of Corollary 3.2 and Theorem 3.4, we have established the following result for the set-valued fixed-point equation (3.6).

THEOREM 3.5. *Let the set-valued integral equation* (3.1) *be given, and let*

(3.20) *the set-valued mapping* $F : [0, T] \times \mathbb{R}^n \to 2^{\mathbb{R}n}$ *satisfy* (H2), *and*

(3.21) *the operators* (3.2) *satisfy* (H1).

Then (3.1) *has a solution in the sense of* (3.3).

4. Existence of a solution for the thermostat system. We now discuss the feedback-system (2.4)–(2.8) which comes from the thermostat problem. We transform the system into an equivalent set-valued integrodifferential equation.

LEMMA 4.1. *For every* $v \in L^\infty[0, T]$ *the initial-value problem*

$$\beta \dot{u}(t) + u(t) = v(t), \qquad t > 0, \quad u(0) = 0 \tag{4.1}$$

has a unique solution $u \in H_1^\infty[0, T]$ *which is given by*

$$u(t) = \beta^{-1} e^{-t/\beta} \int_0^t e^{s/\beta} v(s)\, ds. \tag{4.2}$$

The linear, bounded operator $S : L^\infty[0, T] \to H_1^\infty[0, T]$, $Sv = u$ *is continuous with respect to the weak-star topologies of* $L^\infty[0, T]$ *and* $H_1^\infty[0, T]$, *respectively.*

Next we consider the linear operator T which assigns to every $u \in H_1^\infty[0, T]$ the (generalized) solution of the following initial-boundary value problem at $x = \bar{x}$:

$$\frac{\partial}{\partial t} y(x, t) - Ly(x, t) = 0 \quad \text{in } G, \tag{4.3}$$

$$y(x, 0) = 0 \quad \text{in } \bar{\Omega}, \tag{4.4}$$

$$By(\xi, t) = g(\xi) u(t), \qquad (\xi, t) \in \Gamma. \tag{4.5}$$

The denotations have the same meaning as in § 2.

Now let $\lambda_m (\psi_m)$ denote the eigenvalues (normalized eigenfunctions) of the elliptic problem

$$L\psi(x) + \lambda \psi(x) = 0, \qquad x \in \Omega, \quad B\psi(\xi) = 0, \quad \xi \in \partial\Omega. \tag{4.6}$$

By well-known results stated in Agmon [1, Thm. 14.6, pp. 103ff.] we have

$$\{\psi_m\} \text{ is a complete orthonormal system in } L_2(\Omega), \tag{4.7}$$

$$\lambda_m \to +\infty, \quad \lambda_m \sim c \cdot m^{2/N}, \tag{4.8}$$

$$\|\psi_m\|_{C(\bar{\Omega})} = O(m^l) \quad \text{for some } l \in \mathbb{N}, \tag{4.9}$$

$$\psi_m \in C^\infty(\bar{\Omega}), \qquad (4.9) \text{ holds for any derivative of } \psi_m. \tag{4.10}$$

Let us introduce the Green's function

$$K(t, x; \tau, \xi) := \sum_{m=1}^{\infty} e^{-\lambda m(t-\tau)} \psi_m(x) \psi_m(\xi), \tag{4.11}$$

$$\text{for } 0 \leqq \tau < t \leqq T,\ x \in \bar{\Omega},\ \xi \in \partial\Omega.$$

To every $u \in C^\infty[0, T]$, the unique solution of (4.3)–(4.5) is given by (see Glashoff–Weck [3])

$$y(x, t; u) := \sum_{m=1}^{\infty} \int_{\Gamma_t} e^{-\lambda_m(t-\tau)} \psi_m(\xi) g(\xi) u(\tau)\, d\tau\, d\xi\, \psi_m(x). \tag{4.12}$$

By (4.8), (4.9) summation and integration may be interchanged, and we obtain

$$y(x, t; u) = \int_{\Gamma_t} K(t, x; \tau, \xi) u(\tau) g(\xi)\, d\tau\, d\xi. \tag{4.13}$$

We now define the operator T which assigns to every $u \in L^\infty[0, T]$ the generalized solution of (4.3)–(4.5) at $x = \bar{x}$:

$$\begin{aligned} (Tu)(t) := y(\bar{x}, t; u) &= \int_{\Gamma_t} K(t, \bar{x}; \tau, \xi) u(\tau) g(\xi)\, d\tau\, d\xi \\ &= \int_0^t k(t-\tau) u(\tau)\, d\tau, \end{aligned} \tag{4.14}$$

where

$$k(s) := \int_{\partial\Omega} K(s, \bar{x}; 0, \xi) g(\xi)\, d\xi.$$

The maximum principle yields (see [3])

$$k(s) \geqq 0, \qquad s \in [0, T], \tag{4.15}$$

$$k(\cdot) \in L^1[0, T], \qquad \int_0^t k(\tau)\, d\tau \leqq 1, \quad t \in [0, T]. \tag{4.16}$$

LEMMA 4.2. *T maps $L^\infty[0, T]$ into $C[0, T]$.*

Proof. Let $0 \leqq s < t \leqq T$. Then we have

$$Tu(t) - Tu(s) = \int_0^s (k(t-\tau) - k(s-\tau)) u(\tau)\, d\tau + \int_s^t k(t-\tau) u(\tau)\, d\tau.$$

Since $k(\cdot) \in L^1[0, T]$, the second integral converges to zero as s approaches t. Moreover we have $\lim_{s \to t-0} k(s-\tau) = k(t-\tau)$ for $t > \tau$. Thus by Lebesgue's theorem, the first integral also converges to zero for $s \to t$. □

The following lemma shows that the operator TS maps $L^\infty[0, T]$ into $C^1[0, T]$:

LEMMA 4.3. *$(T\dot{u}) = T\dot{u}$, for every $u \in H_1^\infty[0, T]$ with $u(0) = 0$.*

Proof. Let $u \in H_1^\infty[0, T]$ with $u(0) = 0$ be given, and let $t \in [0, T]$ and $h > 0$ be such

that $t+h \leqq T$. Then we obtain

$$\begin{aligned} h^{-1}&((Tu)(t+h)-(Tu)(t)) \\ &= h^{-1}\int_0^{t+h} k(t+h-\tau)u(\tau)\,d\tau - h^{-1}\int_0^t k(t-\tau)u(\tau)\,d\tau \\ &= h^{-1}\int_0^h k(t+h-\tau)u(\tau)\,d\tau \\ &\quad + h^{-1}\left\{\int_h^{t+h} k(t+h-\tau)u(\tau)\,d\tau - \int_0^t k(t-\tau)u(\tau)\,d\tau\right\} \\ &=: J_1(h)+J_2(h). \end{aligned}$$

The first summand $J_1(h)$ converges to zero for $h \to 0$ since $u(0)=0$ and (4.16):

$$\lim_{h\to 0} J_1(h) = k(t)u(0) = 0 \quad \text{a.e. on } [0, T].$$

For the second expression we obtain

$$\begin{aligned} J_2(h) &= h^{-1}\left\{\int_0^t k(t-s)u(s+h)\,ds - \int_0^t k(t-\tau)u(\tau)\,d\tau\right\} \\ &= \int_0^t k(t-\tau)\frac{u(\tau+h)-u(\tau)}{h}\,d\tau. \end{aligned}$$

Since $u \in H_1^\infty[0, T]$, we have a.e. on $[0, T]$

$$\lim_{h\to 0} h^{-1}\{u(\tau+h)-u(\tau)\} = \dot{u}(\tau).$$

Equation (4.16) and Lebesgue's theorem yield

$$\lim_{h\to 0} J_2(h) = \int_0^t k(t-\tau)\dot{u}(\tau)\,d\tau.$$

The lemma is proved. □

Now let us consider the set-valued mapping $F: \mathbb{R}^2 \to 2^{\mathbb{R}}$ as defined by (2.2). F induces the mapping

$$\phi : C^1[0, T] \to 2^{L^\infty[0,T]} \tag{4.17}$$

by

$$v \in \phi(y) \quad \text{iff} \quad v(\tau) \in F(y(\tau), \dot{y}(\tau)) \quad \text{a.e. on } [0, T]. \tag{4.18}$$

We now define:

DEFINITION. $\hat{y}(x, t)$ is called a *solution of the thermostat problem* (2.4)–(2.8) if and only if $y := \hat{y}(\bar{x}, \cdot) \in C^1[0, T]$ is a solution of $y \in TS\phi(y)$, i.e., if there exists a $v \in L^\infty[0, T]$ with (4.18) and $y = TSv$.

The main result of this paper is:

THEOREM 4.4. *The thermostat problem has a solution.*

Proof. We want to apply Theorem 3.1. The method of the proof is close to that of Theorem 3.5; we thus may be brief and refer to § 3.

Let X be the locally convex topological vector space $C^1[0, T]$ with the weak topology. We set $M := TS(V)$, where

$$V := \{v \in L^\infty[0, T]: \|v\|_\infty \leqq 1\}.$$

M is convex by the linearity of both T and S, and Lemmas 4.1–4.3 yield $M \subset X$. We show that $A(X) \subset M$, hence in particular $A(M) \subset M$. In fact, by definition of F we have $\phi(X) \subset V$ and thus $A(X) = TS\phi(X) \subset TS(V) = M$.

LEMMA 4.5. *Let $\{v_m\}_{m\in\mathbb{N}}$ be a bounded sequence in $L^\infty[0, T]$ such that $v_m \rightharpoonup^* v$ for some $v \in L^\infty[0, T]$. Then it follows that $Tv_m \to Tv$.*

Once Lemma 4.5 is proved it follows that M is weakly relatively compact in $C^1[0, T]$: If $\{\xi_\lambda\}_{\lambda\in\Lambda}$ is a Moore–Smith sequence in M, we have $\xi_\lambda = TSv_\lambda$, with $v_\lambda \in V$, $\lambda \in \Lambda$.

The bounded subset $\{v_\lambda\}_{\lambda\in\Lambda}$ of $L^\infty[0, T]$ contains a bounded subsequence $\{v_m\}_{m\in\mathbb{N}}$ such that $v_m \rightharpoonup^* v$ for some $v \in L^\infty[0, T]$. Then it follows that $Sv_m \rightharpoonup^* Sv$ in $H_1^\infty[0, T]$, and Lemmas 4.3 and 4.5 show that $\xi_m = TS\, v_m \to TSv$ in $C^1[0, T]$.

The other assumptions of the fixed-point theorem can be verified as in § 3, since F is obviously upper semicontinuous. Hence with Lemma 4.5 the proof of Theorem 4.4 is complete. □ □

Proof. Let $\{v_m\}_{m\in\mathbb{N}}$ be bounded in $L^\infty[0, T]$ and $v_m \to^* v$ for some $v \in L^\infty[0, T]$. We have to show that $Tv_m \to Tv$, i.e., that with $w_m := v_m - v$, $m \in \mathbb{N}$:

$$\lim_{m\to\infty} \int_0^T \int_0^t k(t-\tau) w_m(\tau)\, d\tau\, d\alpha(t) = 0,$$

for every normalized function α of bounded variation on $[0, T]$. Now, since $k(\cdot) \in L^1[0, T]$ and $w_m \rightharpoonup^* 0$, we have pointwise on $[0, T]$

$$z_m(t) := \int_0^t k(t-\tau) w_m(\tau)\, d\tau = \int_0^t k(s) w_m(t-s)\, ds \to 0,$$

for $m \to \infty$. Moreover, we have, with a constant $\gamma > 0$, $\|w_m\|_\infty \leqq \gamma$, since $\{v_m\}$ is bounded.

(4.15) and (4.16) yield pointwise on $[0, T]$ for every $m \in \mathbb{N}$

$$|z_m(t)| \leqq \|w_m\|_\infty \int_0^t k(s)\, ds \leqq \gamma.$$

We thus can conclude by Lebesgue's theorem that for every normalized function α of bounded variation

$$\lim_{m\to\infty} \int_0^T z_m(t)\, d\alpha(t) = 0. \qquad \square$$

REFERENCES

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

[2] K. GLASHOFF AND J. SPREKELS, *The regulation of temperature by thermostats and set-valued integral equations*, Integral Equations, to appear.

[3] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimensions: supremum-norm problems*, SIAM J. Control Optim., 14 (1976), pp. 662–681.

[4] I. L. GLICKSBERG, *A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points*, Proc. Amer. Math. Soc., 3 (1952), pp. 170–174.

[5] Y. SAKAWA, *Solution of an optimal control problem in a distributed parameter system*, IEEE Trans. Automat. Control, 9 (1964), pp. 420–426.

SIAM J. MATH. ANAL.
Vol. 12, No. 3, May 1981

ERRATUM:

UNIFORMLY VALID EXPANSIONS FOR LAPLACE INTEGRALS*

L. A. SKINNER†

Replace the first occurrence of (2.12) on page 1061 with

$$\psi_m(r, R, t) = g(r, t) \exp[-R^m v_m(r, t)]. \tag{2.11}$$

* This Journal, 11 (1980), pp. 1058–1067.

† Department of Mathematical Sciences, University of Wisconsin, Milwaukee, Wisconsin 53201.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0001 $01.00/0

THE RADIAL WAVE AND EULER-POISSON-DARBOUX EQUATIONS WITH SINGULAR DATA*

L. R. BRAGG†

Abstract. Representations of generalized dissipative solutions of the radial wave and radial Euler-Poisson-Darboux equations are obtained corresponding to singular initial data. The data have the structure $\mathcal{S}(r) \cdot \psi(r)$, in which $\mathcal{S}(r)$ involves a pole or logarithmic singularity at $r = 0$ and $\psi(r)$ is entire in r^2. Uniqueness is discussed. Differentiability of solutions in the neighborhood of the characteristic line is also treated.

1. Introduction. Let μ and k be real parameters with $\mu \geqq 1$ and $k \geqq 0$ and let $\Delta_\mu = D_r^2 + ((\mu - 1)/r)D_r$ denote the radial Laplacian operator. We shall be concerned with the structure of dissipative solutions of the following hyperbolic problems:

$$\begin{aligned} &\frac{\partial^2 W(r,t)}{\partial t^2} = \Delta_\mu W(r,t), \qquad r, t > 0, \\ &W(r,0+) = 0, \qquad W_t(r,0+) = \phi(r), \qquad r > 0, \end{aligned} \tag{1.1}$$

and

$$\begin{aligned} &\frac{\partial^2 E(r,t)}{\partial t^2} + \frac{k}{t}\frac{\partial E(r,t)}{\partial t} = \Delta_\mu E(r,t), \qquad r, t > 0, \\ &E(r,0+) = \phi(r), \qquad E_t(r,0+) = 0, \qquad r > 0, \end{aligned} \tag{1.2}$$

in which the real-valued function $\phi(r)$ has a singularity at $r = 0$. More precisely, $\phi(r)$ has the form $\phi(r) = \mathcal{S}(r) \cdot \psi(r)$, in which $\psi(r)$ is entire in r^2 and $\mathcal{S}(r)$ contains a pole, logarithmic singularity or both at $r = 0$. The equation in (1.1) is the radial wave equation while the one in (1.2) is the radial Euler-Poisson-Darboux equation. The solution of the equation in (1.1) in which $W(r,0+) = \phi(r)$, $W_t(r,0+) = 0$ is, of course, a solution of (1.2) corresponding to $k = 0$. In this, a solution, say $W^\mu(r,t)$, will be called *dissipative* if, for $t > 0$, $\lim_{r \to 0} W^\mu(r,t)$ exists and is finite. Closely related to (1.1) and (1.2) is the radial heat problem

$$\begin{aligned} &\frac{\partial u(r,t)}{\partial t} = \Delta_\mu u(r,t), \qquad r, t > 0, \\ &u(r,0+) = \phi(r), \qquad r > 0. \end{aligned} \tag{1.3}$$

There are a number of reasons for carrying out this study. In [8], [9] the author and J. W. Dettman developed representations of solutions of (1.1) and (1.2) (and their generalizations) in terms of Jacobi polynomials under the assumption that $\phi(r)$ is analytic in r^2. Our purpose here is to extend the treatment to handle data containing singularities. We shall be concerned with (i) determining how a singularity in the data propagates throughout the hyperbolic solution function and (ii) determining the differentiability properties of these solutions near the characteristics. A solution here will be taken in the following generalized sense: continuous and piecewise C^2. For example, if $\mu = 3$ and $\phi(r) = r^{-1}$, then the function $W(r,t) = t/r$ for $t < r$ and 1 for $t \geqq r$ is a generalized solution of (1.1). A singularity of the type considered may be viewed as

* Received by the editors October 4, 1979, and in revised form September 5, 1980.

† Department of Mathematical Sciences, Oakland University, Rochester, Michigan 48063. This work was supported in part by Oakland University Research Funds.

introducing an impulse or a point "explosion," and the problem is to determine its effect on the hyperbolic solution. Such a study could serve to assist in formulating methods for treating more general types of singular hyperbolic problems. A number of examples of closed form solutions of (1.1) are given to illustrate some of the typical types of solution behavior that could be expected.

A study similar to this was carried out by the author in connection with the initial value problem for the radial heat problem (1.3) [3]. In that case, a solution exists in the classical sense if $\int_{0+}^{a} \mathscr{S}(r)r^{-\mu}\,dr$ exists for any finite, positive a. We shall call upon the results for this heat problem to treat (1.1), (1.2) by the use of the methods of related partial differential equations [5], [6], [7]. By these means, we can transform results for (1.3) into corresponding results for (1.1) and (1.2). Differentiability properties of the solutions of (1.1) and (1.2) will then follow from general results associated with the Laplace transform. As would be expected, more restrictions are needed for the existence and smoothness of solutions of (1.1) and (1.2) than are needed for (1.3).

The basic mathematical notions, techniques, and background needed for this study will be developed in § 2. Included in this will be a uniqueness theorem that will make clear the types of solution functions permitted. A somewhat detailed treatment of a simple pole will be given in § 3. Results for a more general type of pole will be obtained in § 4. Finally, § 5 will treat logarithmic and mixed singularities.

The reader is referred to R. W. Carroll [17] for a general background in related partial differential equations and transmutations. In this, the author treats the Euler-Poisson-Darboux problem in a function space setting. Also, see [18] for a treatment of singular equations.

2. Preliminaries. We now summarize the basic ideas, tools, and results that will be needed in the ensuing development. A solution notation similar to that used in [3] will be introduced that will permit us to easily distinguish between the different solutions of (1.1) and (1.2). We also include (i) a uniqueness result that will make clear the types of solutions of (1.1) and (1.2) under consideration and (ii) some typical types of calculations involving inverse Laplace transforms that will be called upon later.

As was mentioned, we will need to make use of results for the radial heat problem (1.3). Throughout this paper, we will denote the "appropriate" solution of this problem by $u^{\mu}(r, t, \phi)$ in order to clearly indicate the precise value of the parameter μ and the underlying data function $\phi(r)$. Similarly, we denote a solution of (1.1) by $W^{\mu}(r, t, \phi)$ and a solution of (1.2) by $E_k^{\mu}(r, t, \phi)$.

DEFINITION 2.1. An entire function $\phi(z) = \sum_{j=0}^{\infty} a_j z^j$ is of growth (ρ, τ) if and only if

$$\limsup_{j\to\infty} \left(\frac{j}{e\rho}\right)|a_j|^{\rho/j} \leqq \tau \quad \text{(see [1])}.$$

THEOREM 2.1. *Let $\psi(r) = \sum_{j=0}^{\infty} a_j r^{2j}$ be an entire function in r^2 of growth $(1, \sigma)$. Then there exists a solution $u^{\mu}(r, t, \psi)$ of (1.3) of the form $\sum_{j=0}^{\infty} a_j R_j^{\mu}(r, t)$ in the time strip $|t| < 1/4\sigma$ in which the $R_j^{\mu}(r, t)$ denote the radial heat polynomials* [2].

THEOREM 2.2. *Let $\psi(r)$ be an entire function of growth $(1, \sigma)$ in r^2. Then a solution of (1.1) with $\phi(r)$ replaced by $\psi(r)$ is given by*

$$W^{\mu}(r, t, \psi) = \Gamma\left(\frac{3}{2}\right)\mathscr{L}_s^{-1}\left\{s^{-3/2}u^{\mu}\left(r, \frac{1}{4s}, \psi\right)\right\}_{s\to t^2}, \tag{2.1}$$

and a solution of (1.2) with $\phi(r)$ replaced by $\psi(r)$ is given by

$$E_k^{\mu}(r, t, \psi) = t^{1-k}\Gamma\left(\frac{k+1}{2}\right)\mathscr{L}_s^{-1}\left\{s^{-(k+1)/2}u^{\mu}\left(r, \frac{1}{4s}, \psi\right)\right\}_{s\to t^2}, \tag{2.1'}$$

in which $\mathscr{L}_s^{-1}\{\cdot\}_{s\to t^2}$ *denotes the inverse Laplace transform with* s *the variable of the transform and* t^2 *the variable of inversion (see* [6], [7]). *The transformations* (2.1) *and* (2.1′) *provide the basic means for obtaining solutions of* (1.1) *and* (1.2) *from the corresponding solution of* (1.3). *Since* $\psi(r)$ *is entire in* r^2, *the functions* $W^\mu(r, t, \psi)$ *and* $E_k^\mu(r, t, \psi)$ *that correspond to* $\psi(r)$ *exist and are entire for* $r, t \geqq 0$. *The results in* [3] *show that* $\psi(r)$ *can be replaced by* $\phi(r) = \mathscr{S}(r) \cdot \psi(r)$ *and that solutions of* (1.3) *can be obtained for a variety of singularities in* $\mathscr{S}(r)$. *The transformations* (2.1) *and* (2.1′) *apply in these cases as well if* $\mathscr{S}(r)$ *is somewhat restricted (the "smoothing" associated with the heat problem permits more badly behaved singularities in the data).*

We now give a brief discussion of the uniqueness question for solutions of (1.1) and (1.2) and consider some of its implications. Recall that

DEFINITION 2.2. A solution $W^\mu(r, t)$ of (1.1) or $E_k^\mu(r, t)$ of (1.2) is dissipative if, for $t > 0$, $\lim_{r\to 0} W^\mu(r, t)$ or $\lim_{r\to 0} E_k^\mu(r, t)$ exists and is finite [3].

The importance of this definition can be illustrated by the following examples. In (1.1), select $\mu = 4$, $\mathscr{S}(r) = r^{-2}$ and $\psi(r) = 1$. Then

$$W^4(r, t, \phi) = \begin{cases} \dfrac{t}{r^2} & \text{for } t < r, \\ r^{-2}[t - C(t^2 - r^2)^{1/2}] & \text{for } t \geqq r, \end{cases}$$

in which C is an arbitrary real constant. Among these solutions, the only dissipative one results from the choice $C = 1$. Without a condition of this type, uniqueness fails because $r = 0$ has been excluded from the data and solution regions and it is possible to construct nontrivial solutions of (1.3) that have support at $r = 0$ but which vanish for $r > 0$, $t = 0$. (The Tychonov and Widder type uniqueness theorems for solutions of the heat problem require that $r = 0$ be included in the data and solution regions [11], [14], [15].) Similarly, we note that the function $u(r, t) = \ln r + 2t/r^2$ is a solution of (1.3) for $\mu = 4$ and $r > 0$ that corresponds to $\phi(r) = \ln r$. It fails, however, to have the dissipative property. So also do its transforms (2.1) and (2.1′). The solution function, in this case, has a more badly behaved singularity at $r = 0$ than do the initial data. By a slight modification of the proof of the uniqueness theorem 6.1 of [4], we can establish the following result for $W^\mu(r, t)$ (and $E_k^\mu(r, t)$ for $\mu + 1 \geqq k \geqq 0$).

THEOREM 2.3. *Let* $\mu \geqq 1$ *and let* $W^\mu(r, t) \in C^2$ *and satisfy* (1.1) *in each of the regions* $0 < r < t$ *and* $0 < t < r$. *Suppose that*

(a) $W^\mu(r, 0+) = W_t(r, 0+) = 0$ *for* $r > 0$,

(b) $W^\mu(r, t)$ *is dissipative, and*

(c) $W^\mu(r, t)$ *is continuous across the line* $t = r$.

Then $W^\mu(r, t) \equiv 0$ *for* $r \geqq 0$, $t > 0$.

The condition (b) replaces the inclusion of the line $r = 0$ in the earlier theorem. As a consequence of this result, we transform only dissipative solutions of (1.3) and restrict the choice of singularities so that condition (c) of theorem 2.3 is satisfied.

From (2.1) and (2.1′), we note that

$$W^\mu(r, t, \phi) = tE_2^\mu(r, t, \phi). \tag{2.2}$$

More typically, we will need to carry out inversions of the following type:

$$\mathscr{L}_s^{-1}\left\{\frac{e^{-r^2\xi s}u^\lambda(r\sqrt{1-\xi}, 1/4s, \psi)}{s^p}\right\}_{s\to\tau},$$

in which $r>0$, $\xi>0$, $\lambda \geqq 0$. From (2.1'),

$$\mathscr{L}_s^{-1}\left\{s^{-1/2}u^{\lambda}\left(r\sqrt{1-\xi}, \frac{1}{4s}, \psi\right)\right\}_{s\to\tau} = \frac{1}{\Gamma(1/2)}\tau^{-1/2}E_0^{\lambda}(r\sqrt{1-\xi}, \tau^{1/2}, \psi),$$

and, for $p>\frac{1}{2}$,

$$\mathscr{L}_s^{-1}\{s^{-(p-1/2)}\, e^{-r^2\xi s}\}_{s\to\tau} = \frac{(\tau-r^2\xi)_+^{p-3/2}}{\Gamma(p-1/2)}, \tag{2.3}$$

in which [16]

$$(a-b)_+^{\upsilon} = \begin{cases} 0 & \text{if } a<b, \\ (a-b)^{\upsilon} & \text{if } a>b. \end{cases} \tag{2.4}$$

Using the convolution property, we have

$$\begin{aligned} &\mathscr{L}_s^{-1}\left\{\frac{e^{-r^2\xi s}}{s^p}u^{\lambda}\left(r\sqrt{1-\xi}, \frac{1}{4s}, \psi\right)\right\}_{s\to\tau} \\ &\quad = \frac{1}{\sqrt{\pi}\Gamma(p-1/2)}\int_0^{\tau-r^2\xi}(\tau-r^2\xi-\sigma)_+^{p-3/2}\sigma^{-1/2}E_0^{\lambda}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)\, d\sigma, \end{aligned} \tag{2.5}$$

for $p>\frac{1}{2}$. In this, τ is usually replaced by t^2. Since the inversion (2.5) occurs in connection with various types of poles, it is clear from (2.4) and (2.5) that special attention needs to be given to the solutions of (1.1) and (1.2) at $t=r$. If $p \leqq \frac{1}{2}$ in (2.3), the inverse transform will exist in (2.5) in the sense of distributions but will usually fail to lead to solutions of (1.1) and (1.2) that are continuous at $t=r$. Standard results from Laplace transform theory show that a condition that is sufficient (but not necessary) to ensure that solutions of (1.1) and (1.2) (that involve the inversion (2.5)) $\in C^l$ at $t=r$ is

$$l<p-\tfrac{1}{2}, \qquad l=0, 1, 2, \cdots. \tag{2.6}$$

In applying (2.1) and (2.1') to solutions of (1.3), one must frequently interchange the order of integrating a function of several variables and computing its inverse Laplace transform. For the types of data treated here, this can be readily validated using standard results on the interchange of orders on integration in connection with uniform convergence. See, for example, [10, p. 497].

3. An elementary pole. In this section, we discuss (1.1) and (1.2) when $\phi(r)$ has the form $\phi(r)=r^{-1}\psi(r)$ with $\psi(r)$ entire of growth $(1, \sigma)$ in r^2. This case will provide us with some interesting special cases of wave propagation (for various values of k and μ) and will furnish a model for treating other types of singularities through use of (2.1) and (2.1').

By [3, Theorem 4.3], an elementary change of variables and the definition of a, it follows that, for $\mu>1$,

$$u^{\mu}(r, t, \phi) = (4\pi t)^{-1/2}\int_0^1 \xi^{-1/2}(1-\xi)^{(\mu-3)/2}\, e^{-r^2\xi/4t}u^{\mu-1}(r\sqrt{1-\xi}, t, \psi)\, d\xi. \tag{3.1}$$

An application of (2.1) to (3.1) and references to the remarks in the final paragraph of § 2 shows that

$$\begin{aligned} &W^{\mu}(r, t, \phi) \\ &\quad = \frac{1}{2}\int_0^1 \xi^{-1/2}(1-\xi)^{(\mu-3)/2}\left\{\mathscr{L}_s^{-1}\left[s^{-1}\, e^{-r^2\xi s}u^{\mu-1}\left(r\sqrt{1-\xi}, \frac{1}{4s}, \psi\right)\right]_{s\to\tau}\right\} d\xi, \end{aligned} \tag{3.2}$$

in which $\tau = t^2$. But, by (2.5) with $p = 1$, the bracketed term in the right member of (3.2) has the evaluation

$$\frac{1}{\pi}\int_0^{\tau - r^2\xi} (\tau - r\xi - \sigma)_+^{-1/2}\sigma^{-1/2}E_0^{\mu-1}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)\, d\sigma.$$

Upon replacing τ by t^2 and inserting this back into (3.2), we obtain

$$\begin{aligned} &W^{\mu}(r, t, \phi) \\ &\quad = \frac{1}{2\pi}\int_E \xi^{-1/2}(1-\xi)^{(\mu-3)/2} \\ &\qquad \left\{\int_0^{t^2 - r^2\xi} (t^2 - r^2\xi - \sigma)^{-1/2}\sigma^{-1/2}E_0^{\mu-1}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)\, d\sigma\right\} d\xi, \end{aligned} \tag{3.3}$$

in which the region of integration E is taken to be the interval $[0, 1]$ if $t^2 \geqq r^2$ and the interval $[0, t^2/r^2]$ if $t < r$. This shows that the form of the solution $W^{\mu}(r, t, \phi)$ is dependent upon the position of the point (r, t) in relation to the characteristic line $t = r$. Since $E_0^{\mu-1}(r, t, \psi)$ has the dissipative property, it is readily seen that $W^{\mu}(r, t, \phi)$ also has the dissipative property. Summarizing, we have

THEOREM 3.1. *Let $\mu > 1$ and let $\psi(r)$ be an entire function of r^2. Then the dissipative solution of* (1.1) *corresponding to $\phi(r) = r^{-1}\psi(r)$ is given by* (3.3).

Similarly, an application of (2.1′) to (3.1) for $\mu > 1$ gives

$$\begin{aligned} E_k^{\mu}(r, t, \phi) = &\frac{\Gamma\left(\frac{k+1}{2}\right)t^{1-k}}{2\sqrt{\pi}}\int_0^1 \xi^{-1/2}(1-\xi)^{(\mu-3)/2}\mathcal{L}_s^{-1} \\ &\cdot \left\{s^{-k/2}\, e^{-r^2\xi s}u^{\mu-1}\left(r\sqrt{1-\xi}, \frac{1}{4s}, \psi\right)\right\}_{s \to t^2} d\xi. \end{aligned} \tag{3.4}$$

Again, by (2.5) with $p = k/2$, the bracketed term in this becomes, for $k > 1$,

$$\frac{\pi^{-1/2}}{\Gamma\left(\frac{k-1}{2}\right)}\int_0^{\tau - t^2\xi} (\tau - r^2\xi - \sigma)_+^{(k-3)/2}\sigma^{-1/2}E_0^{\mu-1}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)\, d\sigma.$$

Upon replacing τ by t^2 and replacing this back into (3.4), we get

$$\begin{aligned} E_k^{\mu}(r, t, \phi) = &\frac{(k-1)t^{1-k}}{4\pi}\int_0^1 \xi^{-1/2}(1-\xi)^{(\mu-3)/2} \\ &\cdot \left\{\int_0^{t^2 - r^2\xi} (t^2 - r^2\xi - \sigma)^{(k-3)/2}\sigma^{-1/2} \right. \\ &\left. \cdot E_0^{\mu-1}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)\, d\sigma\right\} d\xi. \end{aligned} \tag{3.5}$$

When we evaluate this, the bracketed integral vanishes if $t^2 < r^2\xi$. Using the condition (2.6) and its implications, we see that $E_k^{\mu}(r, t, \phi) \in C^l$ at $t = r$ if $2l + 1 < k$ for $l = 0, 1, 2, \cdots$. Summarizing, we have:

THEOREM 3.2. *Let $\mu > 1$ and let $\psi(r)$ be an entire function of r^2. Then the dissipative solution of* (1.2) *corresponding to $\phi(r) = r^{-1}\psi(r)$ is given by* (3.5) *for $k > 1$. If $2l + 1 < k$, then $E_k^{\mu}(r, t, \phi) \in C^l$ at $t = r$.*

Whether the function $W^{\mu}(r, t, \phi)$ of Theorem 3.1 or $E_k^{\mu}(r, t, \phi)$ has more than the minimum number of derivatives at $t = r$ will depend upon the properties of $\psi(r)$, i.e., whether $\psi(0) = 0$. The following examples will illustrate the possibilities.

Example 1. Select $\psi(r)=1$. Then for any $\mu>1$, $E_0^{\mu-1}(r\sqrt{1-\xi}, \sigma^{1/2}, \psi)=1$, and it follows that the inside integral in the right-hand side of (3.3) becomes

$$\int_0^{t^2-r^2\xi}(t^2-r^2\xi-\sigma)^{-1/2}\sigma^{-1/2}\,d\sigma,$$

which has the value π. Hence,

$$W^{\mu}\left(r, t, \frac{1}{r}\right)=\frac{1}{2}\int_E \xi^{-1/2}(1-\xi)^{(\mu-3)/2}\,d\xi, \qquad \mu>1. \tag{3.6}$$

If $t\geqq r$, the E in (3.6) is the interval $[0, 1]$ and the right side of (3.6) reduces to $\frac{1}{2}B((\mu-1)/2, \frac{1}{2})$. However, if $t<r$, $E=[0, t^2/r^2]$ and the integral in (3.6) defines the incomplete beta function $\frac{1}{2}B(\frac{1}{2}, (\mu-1)/2, t^2/r^2)$ (see [13, p. 356]). Using the hypergeometric series for this last function, we obtain

$$W^{\mu}\left(r, t, \frac{1}{r}\right)=\begin{cases}\dfrac{1}{2}B\left(\dfrac{\mu-1}{2}, \dfrac{1}{2}\right) & \text{for } t\geqq r,\\[2ex] \dfrac{t}{r}{}_2F_1\left(\dfrac{3-\mu}{2}, \dfrac{1}{2}; \dfrac{3}{2}; \dfrac{t^2}{r^2}\right) & \text{for } t<r.\end{cases} \tag{3.7}$$

If μ is an odd integer with $\mu\geqq 3$, the series for ${}_2F_1$ terminates (the choice $\mu=3$ yields the example given in the introduction). Along the characteristic line $t=r$, the incomplete beta function reduces to $B((\mu-1)/2, \frac{1}{2})$ so that W^{μ} is continuous there. Since $\partial W^{\mu}/\partial t=0$ for $t>r$ and $\partial W^{\mu}/\partial t\neq 0$ for $t<r$, $\partial W^{\mu}/\partial t$ is discontinuous there.

Remark 1. The function

$$E_1^{\mu}(r, t, r^{-1})=\frac{1}{2\pi}\int_0^1 \xi^{-1/2}(1-\xi)^{(\mu-3)/2}(t^2-r^2\xi)_+^{-1/2}\,d\xi,$$

is a "solution" of (1.2) corresponding to $k=1$. If $\mu>2$, E_1^{μ} is continuous at $t=r$. This shows that condition (2.6) is sufficient but not necessary. E_1^{μ} is not continuous at $t=r$ if $\mu\leqq 2$.

4. Other types of poles. We can use the techniques of § 3 to obtain representations of dissipative solutions of (1.1) and (1.2) when the function $\mathcal{S}(r)$ contains a pole of a more complicated form that can depend upon certain parameters. We treat solution representations of two different types corresponding to $\phi(r)=r^{2-\mu-2\alpha}\psi(r)$, α real. In view of the relation (2.2), our general discussion will pertain to solutions of (1.2).

Case 1. $\phi(r)=r^{2-\mu-2\alpha}\psi(r)$, $0\leqq\alpha<\frac{1}{2}$, $\mu>2$. With elementary changes in the variables of integration and the definitions of $S_{\mu}(r, t)$ and a, it follows from [3, Theorem 4.2] that

$$u^{\mu}(r, t, \phi)=\frac{(4t)^{-(\mu/2+\alpha-1)}}{\Gamma(\mu/2+\alpha-1)}\int_0^1 (1-\xi)^{-\alpha}\xi^{\mu/2+\alpha-2}\,e^{-r^2\xi/4t}u^{2-2\alpha}(r\sqrt{1-\xi}, t, \psi)\,d\xi.$$

An application of (2.1′) to this function u^{μ} gives

$$E_k^{\mu}(r, t, \phi)=\frac{\Gamma\left(\dfrac{k+1}{2}\right)t^{1-k}}{\Gamma(\mu/2+\alpha-1)}\int_E (1-\xi)^{-\alpha}\xi^{\mu/2+\alpha-2}K(r, t, \xi)\,d\xi, \tag{4.1}$$

in which

$$(4.2)\quad \text{(a)}\quad K(r,t,\xi)=\mathscr{L}_s^{-1}\left\{\frac{e^{-r^2\xi s}u^{2-2\alpha}(r\sqrt{1-\xi},\,1/4s,\,\psi)}{s^{(k+3)/2-\mu/2-\alpha}}\right\}_{s\to t^2},$$

$$\text{(b)}\quad E=\begin{cases}[0,1] & \text{if } t\geqq r,\\ [0,t^2/r^2] & \text{if } t<r.\end{cases}$$

From (2.4) and (2.5) we find that the function K in (4.2a) is given by

$$(4.3)\quad K(r,t,\xi)=\begin{cases}0 & \text{if } t^2<r^2\xi,\\ \dfrac{1}{\sqrt{\pi}\Gamma\left(\dfrac{k+1-\mu}{2}-\alpha\right)}\displaystyle\int_0^{t^2-r^2\xi}(t^2-r^2\xi-\sigma)^{(k-\mu)/2-\alpha} & \\ \qquad\cdot\sigma^{-1/2}E_0^{2-2\alpha}(r\sqrt{1-\xi},\,\sigma^{1/2},\,\psi)\,d\sigma & \text{if } t^2>r^2\xi.\end{cases}$$

We have obtained (4.1) and (4.3) on the assumption that $\frac{1}{2}(k+3-\mu)-\alpha>\frac{1}{2}$. Suppose $\psi(0)\neq 0$. Using the condition (2.6) here shows that the function in $(4.1)\in C^l$ if $l<\frac{1}{2}(k+2-\mu)-\alpha$ or $2l-2+\mu+2\alpha<k$. From this and the relationship (2.5), we have:

THEOREM 4.1. *Let $E_k^\mu(r,t,\phi)$ be a dissipative solution of* (1.2) *corresponding to $\phi(r)=r^{2-\mu-2\alpha}\cdot\psi(r)$, $0\leqq\alpha<\frac{1}{2}$, $\mu>2$. Then $E_k^\mu(r,t,\phi)$ is defined by* (4.1) *with $K(r,t,\xi)$ given by* (4.3). *If $\psi(0)\neq 0$, then the function $E_k^\mu(r,t,\phi)\in C^l$ $(l=0,1,2,\cdots)$ provided that $2l-2+\mu+2\alpha<k$.*

THEOREM 4.2. *Let $W^\mu(r,t,\phi)$ be a dissipative solution of* (1.1) *corresponding to $\phi(r)=r^{2-\mu-2\alpha}\cdot\psi(r)$, $0\leqq\alpha<\frac{1}{2}$, $\mu>2$. Then $W^\mu(r,t,\phi)=tE_2^\mu(r,t,\phi)$ with $E_2^\mu(r,t,\phi)$ defined by* (4.1) *and $K(r,t,\xi)$ given by* (4.3). *If $\psi(0)\neq 0$, then $W^\mu(r,t,\phi)$ is continuous across $t=r$ but is not differentiable across this line if $\mu\leqq 4$ and $\alpha\leqq 2-\frac{1}{2}\mu$. If $\mu>4$ and $\psi(0)\neq 0$, $W^\mu(r,t,\phi)$ fails to be continuous at $t=r$.*

Example 2. Suppose that $\psi(r)=1$ and $\mu/2+\alpha<2$. Then $E_0^{2-2\alpha}(r\sqrt{1-\xi},\,\sigma^{1/2},\,\psi)=1$ and an elementary change of variables gives

$$K(r,t,\xi)=\begin{cases}B\left(\dfrac{1}{2},\,2-\dfrac{\mu}{2}-\alpha\right)(t^2-r^2\xi)^{(3-\mu)/2-\alpha}, & t^2>r^2\xi,\\ 0, & t^2<r^2\xi.\end{cases}$$

From (2.4) and (2.5), we obtain

$$W^\mu(r,t,\phi)=\frac{B\left(\dfrac{1}{2},\,2-\dfrac{\mu}{2}-\alpha\right)}{2\Gamma(\mu/2+\alpha-1)\Gamma(2-\mu/2-\alpha)}\int_E\xi^{\mu/2+\alpha-2}(1-\xi)^{-\alpha}(t^2-r^2\xi)^{(3-\mu)/2-\alpha}\,d\xi,$$

in which $E=[0,1]$ if $t\geqq r$ and $[0,t^2/r^2]$ if $t<r$. From the integral defining the hypergeometric function ${}_2F_1(a,b;c;z)$ (see [13, p.54]), we obtain the evaluations

$$W^\mu(r,t,\phi)=\begin{cases}\dfrac{t}{r^{\mu-2+2\alpha}}\,{}_2F_1\left(\alpha,\dfrac{\mu}{2}+\alpha-1;\dfrac{3}{2};\dfrac{t^2}{r^2}\right), & t<r,\\ \dfrac{\Gamma(3/2)\Gamma(1-\alpha)t^{3-\mu-2\alpha}}{\Gamma\left(\dfrac{5-\mu}{2}-\alpha\right)\Gamma\left(\dfrac{\mu}{2}\right)}\,{}_2F_1\left(\dfrac{\mu-3}{2},\dfrac{\mu}{2}+\alpha-1;\dfrac{\mu}{2};\dfrac{r^2}{t^2}\right), & t\geqq r.\end{cases}$$

For the special case $\mu = 3$, this simplifies to

$$W^3(r, t, \phi) = \begin{cases} \dfrac{1}{2(1-2\alpha)r} \cdot [(r+t)^{1-2\alpha} - (r-t)^{1-2\alpha}], & t < r, \\ \dfrac{1}{2(1-2\alpha)r} \cdot [(t+r)^{1-2\alpha} - (t-r)^{1-2\alpha}], & t \geqq r. \end{cases} \tag{4.4}$$

Let $h(r, \alpha)$ denote the value of $W^3(r, r, \phi)$ for a fixed $r > 0$. Then $h(r, \alpha) = (1-2\alpha)^{-1}(2r)^{-2\alpha}$. It is immediate that $\lim_{\alpha \to (1/2)^-} h(r, \alpha) = \infty$. If $r \leqq e/2$, $h(r, \alpha)$ is monotone increasing in α. If $r > e/2$, it is easy to show, by elementary calculus, that $h(r, \alpha)$ decreases for $0 \leqq \alpha \leqq \frac{1}{2}(1 - 1/\ln 2r)$ and then increases. The graph of W^3 versus t for a fixed value of r ($r = 3$) and various choices of α is given in Fig. 1. The choice $\alpha = .22$ yields the minimum value of $h(3, \alpha)$. If $\alpha > 0$, there is a vertical cusp in the solution curve at $t = 3$.

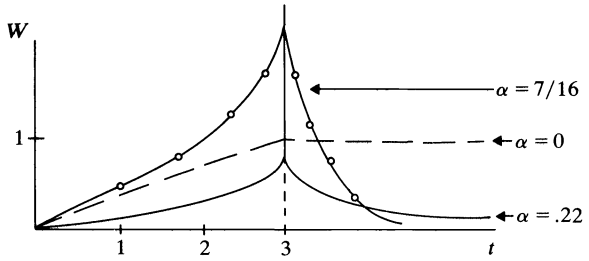


FIG. 1

Example 3. The example given in § 2 to illustrate the dissipative property for the choice $C = 1$ corresponds to the choices $\mu = 4$ and $\alpha = 0$ in Theorem 4.2. The graph of this function $W^4(r, t, \phi)$ ($= t/r^2$ for $t < r$ and $r^{-2}[t - (t^2 - r^2)^{1/2}]$ for $t \geqq r$) versus t for various choices of r is given in Fig. 2. We again see that, for each r, W is continuous at $t = r$ but is not differentiable there.

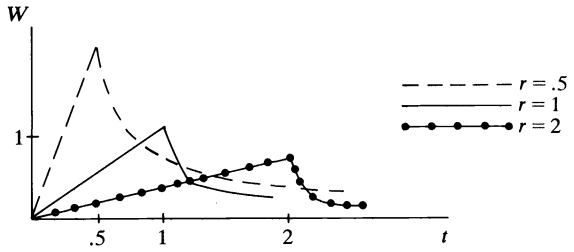


FIG. 2

Case 2. $\phi(r) = r^{2-\mu-2\alpha}\{\beta + r^2\psi(r)\}$, β real, $\alpha < 1$. It follows by [3, Theorem 4.1] that the dissipative solution of (1.3) is given by

$$\begin{aligned} u^\mu(r, t, \phi) = {} & \frac{\beta(4t)^{1-\mu/2-\alpha}}{\Gamma(\mu/2+\alpha-1)} \int_0^1 \xi^{\mu/2+\alpha-2}(1-\xi)^{-\alpha} e^{-r^2\xi/4t}\, d\xi \\ & + \frac{(4t)^{2-\mu/2-\alpha}}{\Gamma(\mu/2+\alpha-2)} \int_0^1 \xi^{\mu/2+\alpha-3}(1-\xi)^{1-\alpha} e^{-r^2\xi/4t} u^{4-2\alpha} \\ & \cdot (r\sqrt{1-\xi}, t, \psi)\, d\xi. \end{aligned} \tag{4.5}$$

Applying (2.1'), (2.4) and (2.5) with $p=(k+5-\mu)/2-\alpha$ in the second integral in (4.5), we find that

$$
\begin{aligned}
E_k^{\mu}(r,t,\phi)=&\frac{\beta t^{1-k}\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(\mu/2+\alpha-1)}\int_0^1 \xi^{\mu/2+\alpha-2}(1-\xi)^{-\alpha}H^*(r,t,\xi)\,d\xi\\
&+\frac{t^{1-k}\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(\mu/2+\alpha-2)}\int_0^1 \xi^{\mu/2+\alpha-3}(1-\xi)^{1-\alpha}K^*(r,t,\xi)\,d\xi,
\end{aligned}
\tag{4.6}
$$

in which $H^*(r,t,\xi)=K^*(r,t,\xi)=0$ for $t^2<r^2\xi$ and

$$
\text{(a)}\quad H^*(r,t,\xi)=\frac{(t^2-r^2\xi)^{(k+1-\mu)/2-\alpha}}{\Gamma\left(\frac{k+3-\mu}{2}-\alpha\right)}\quad\text{for } t^2>r^2\xi,
\tag{4.7}
$$

$$
\begin{aligned}
\text{(b)}\quad K^*(r,t,\xi)=&\frac{\pi^{-1/2}}{\Gamma\left(\frac{k+2-\mu}{2}-\alpha\right)}\\
&\cdot\int_0^{t^2-r^2\xi}(t^2-r^2\xi-\alpha)^{(k+2-\mu)/2-\alpha}\sigma^{-1/2}E_0^{4-2\alpha}(r\sqrt{1-\xi},\sigma^{1/2},\psi)\,d\sigma.
\end{aligned}
$$

We note that the second integral in the right member of (4.6) $\in C^l$ for $l=0,1,2,\cdots$, if k is selected so that $2l-4+\mu+2\alpha<k\leqq 2l-2+\mu+2\alpha$ and $\mu/2+\alpha>2$. On the other hand, the first member on the right side of (4.6) defines a hypergeometric function. By applying the usual differentiation formulas for this, we find that this first term $\in C^l$, for $l=0,1,2,\cdots$, if k is selected so that $2l-3+\mu+4\alpha<k$ and $\mu/2+\alpha>1$. If $k=2$, as in the case of the wave problem (1.1), this last condition on l becomes $2l+\mu+4\alpha<5$. From this, we obtain:

THEOREM 4.3. *Let $E_k^{\mu}(r,t,\phi)$ denote a solution of* (1.2) *corresponding to $\phi(r)=r^{2-\mu-2\alpha}(\beta+r^2\psi(r))$, β real and $\alpha<1$ as given by* (4.6). *Depending upon the choice of β and the value of $\psi(0)$, we have*

(a) *If $\beta\neq 0$, $\psi(0)\neq 0$ and $\mu/2+\alpha>2$, then $E_k^{\mu}(r,t,\phi)\in C^l$, $l=0,1,2,\cdots$, if $k>\max(2l-3+\mu+4\alpha, 2l-4+\mu+2\alpha)$.*

(b) *If $\beta=0$, $\psi(0)\neq 0$ and $\mu/2+\alpha>2$, then $E_k^{\mu}(r,t,\phi)\in C^l$, $l=0,1,2,\cdots$, if $k>2l-4+\mu+2\alpha$.*

(c) *If $\beta\neq 0$, $\psi(0)=0$ and $\mu/2+\alpha>1$, then $E_k^{\mu}(r,t,\phi)\in C^l$, $l=0,1,2,\cdots$, if $k>2l-3+\mu+4\alpha$.*

This theorem applies, of course, to the problem (1.1) when $k=2$. In particular, if $\beta=0$ and $\psi(0)\neq 0$, (b) shows that $W^{\mu}(r,t,\phi)\in C^0$ at $t=r$ if $2<\mu/2+\alpha<3$ but fails to be differentiable there. If $\beta\neq 0$ and $\psi(0)=0$, the conditions in (c) becomes $2l+\mu+4\alpha<5$ and $\mu+2\alpha>2$. From these we see that $W^{\mu}(r,t,\phi)\in C^0$ at $t=r$ if $\mu+2\alpha>2$ and $\mu+4\alpha<5$ and $W^{\mu}(r,t,\phi)\in C^1$ at $t=r$ if $2<\mu+2\alpha$ and $\mu+4\alpha<3$. In the first case, we can select μ so that $2<\mu<5$ and then choose α positive but sufficiently small so that the required inequalities are satisfied. To get differentiability at $t=r$, we must restrict μ to the interval $2<\mu<3$. This discussion with some typical calculations for the hypergeometric function justify:

THEOREM 4.4. *Suppose that $2<\mu<5$ in* (1.1). *Then the most badly behaved pole in $\phi(r)$ that will lead to a dissipative solution of* (1.1) *that is continuous at $t=r$ has the form*

$\phi(r) = r^{-(\mu+1)/2+\varepsilon}$, $\varepsilon > 0$. *The solution corresponding to this* $\phi(r)$ *is*

$$W^{\mu}(r, t, \phi) = \begin{cases} \dfrac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{\mu-1}{4}+\frac{\varepsilon}{2}\right)t^{(1-\mu)/2+\varepsilon}}{\Gamma\left(\frac{\mu}{2}\right)\Gamma\left(\frac{5-\mu}{4}+\frac{\varepsilon}{2}\right)} {}_2F_1\left(\frac{\mu-1}{4}-\frac{\varepsilon}{2}, \frac{\mu+1}{4}-\frac{\varepsilon}{2}; \frac{\mu}{2}; \frac{r^2}{t^2}\right), & t > r, \\ \dfrac{t}{r^{(\mu+1)/2-\varepsilon}} {}_2F_1\left(\frac{5-\mu}{4}-\frac{\varepsilon}{2}, \frac{\mu+1}{4}-\frac{\varepsilon}{2}; \frac{3}{2}; \frac{t^2}{r^2}\right), & t \leqq r. \end{cases}$$

At $t = r$, the two parts of this have the common value

$$\frac{\Gamma(3/2)\Gamma(\varepsilon)}{\Gamma\left(\frac{\mu+1}{4}+\frac{\varepsilon}{2}\right)\Gamma\left(\frac{5-\mu}{4}+\frac{\varepsilon}{2}\right)} \cdot r^{(1-\mu)/2+\varepsilon}$$

Finally, $\lim_{t\to\infty} W^{\mu}(r, t, \phi) = 0$.

Remark 2. In the case of the radial heat problem (1.3), the data can have a pole of the form $r^{-\mu+\varepsilon}$, $\varepsilon > 0$, for a given μ and give rise to a differentiable solution. This clearly points out the difference between the smoothing in the heat solution and the wave solution.

5. Logarithmic singularities. We finally treat (1.1) and (1.2) when $\phi(r) = \Phi(r) \ln r$, in which $\Phi(r)$ is entire in r^2 or else contains a pole. As in §§ 3 and 4, one can transform representations of solutions of (1.3) corresponding to such data by means of (2.1) and (2.1′). While we shall do this when $\Phi(r) = r^{2-\mu}$, it is useful to examine some general results for logarithmic singularities. We will have occasion to use the following inversion for Laplace transforms,

$$\mathcal{L}_s^{-1}\left\{\frac{\ln s}{s^l}\right\}_{s\to\tau} = \frac{\Gamma^{(1)}(l)\tau^{l-1}}{(\Gamma(l))^2} - \frac{1}{\Gamma(l)}\tau^{l-1}\ln\tau, \tag{5.1}$$

for $l > 0$.

For $\mu > 2$, a dissipative solution of (1.3) with the above choice for $\phi(r)$ has the form

$$u^{\mu}(r, t, \phi) = u^{\mu}(r, t, \Phi)\ln r + U^{\mu}(r, t), \tag{5.2}$$

in which $U(r, t)$ is a dissipative solution of the nonhomogeneous heat problem

$$\begin{aligned} &U_t(r, t) = \Delta_{\mu}U(r, t) + f(r, t), \qquad r, t > 0, \\ &U(r, 0+) = 0, \end{aligned} \tag{5.3}$$

where

$$f(r, t) = \frac{2u_r^{\mu}(r, t, \Phi)}{r} + \frac{\mu-2}{r^2}u^{\mu}(r, t, \Phi). \tag{5.4}$$

The solution of (5.3) can be expressed in the symbolic form

$$U(r, t) = \int_0^t e^{(t-\eta)\Delta_{\mu}} f(r, \eta)\, d\eta, \tag{5.5}$$

with this formula to be interpreted as follows. Treating η as a parameter, first construct a dissipative solution $u^{\mu}(r, t, f(r, \eta))$ of (1.3). Then replace t by $t - \eta$ and integrate this resulting function from 0 to t. From (5.4), it is evident that $U(r, t)$ is constructed from a function involving a pole and this imposes the restriction $\mu > 2$. The dissipative solution of (1.2) (and, hence, (1.1)) can be obtained from (5.2) and (5.5) by applying the transform (2.1′). Thus we have:

THEOREM 5.1. *For $\mu > 2$, let $E_k^\mu(r, t, \Phi)$ be the dissipative solution of* (1.2) *corresponding to $\phi(r) = \Phi(r) \ln r$. Then this solution has the form*

$$E_k^\mu(r, t, \phi) = E_k^\mu(r, t, \Phi) \ln r + G(r, t), \tag{5.6}$$

in which $G(r, t)$ satisfies the nonhomogeneous problem

$$\begin{aligned} &G_{tt}(r, t) + \frac{k}{t} G_t(r, t) = \Delta_\mu G(r, t) + f(r, t), \qquad r, t > 0, \\ &G(r, 0+) = G_t(r, 0+) = 0, \qquad r > 0. \end{aligned} \tag{5.7}$$

Moreover,

$$G(r, t) = t^{1-k} \Gamma\left(\frac{k+1}{2}\right) \mathscr{L}_s^{-1}\left\{ s^{-(k+1)/2} U^\mu\left(r, \frac{1}{4s}\right)\right\}_{s \to t^2}. \tag{5.8}$$

(A) *$\Phi(r)$ entire in r^2.* If $\Phi(r) = \sum_{j=0}^\infty a_j r^{2j}$ is entire in r^2, then $u^\mu(r, t, \Phi)$ is entire in r^2 and the function $f(r, t)$ in (5.4) involves, at worst, a pole of the form r^{-2}. One can then use (5.5) to obtain the corresponding $U^\mu(r, t)$ in (5.2). Since the term $a_0 \ln r$ clearly leads to the most badly behaved portion of the solution (5.2), we construct the solution of (1.2) (and hence (1.1)) corresponding to the choice of $\Phi(r) = 1$.

If $\Phi(r) = 1$, then $u^\mu(r, t, 1) = 1$ so that $f(r, t) = (\mu - 2) r^{-2}$. By (5.5),

$$U^\mu(r, t) = (\mu - 2) \int_0^t e^{(t-\eta)\Delta_\mu}(r^{-2})\, d\eta.$$

Using [3, Theorem 2.1] to compute $e^{(t-\eta)\Delta_\mu}(r^{-2})$, we obtain after some changes of variables of integration that

$$U^\mu(r, t) = \frac{\mu - 2}{4} \int_0^1 \xi^{-1} \left\{ \int_0^1 \sigma^{\mu/2-2}\, e^{-r^2(1-\sigma)/4t\xi}\, d\sigma \right\} d\xi.$$

If we add this to the term $u^\mu(r, t, 1) \ln r = \ln r$ in (5.2) and apply (2.1′), we obtain

$$\begin{aligned} &E_K^\mu(r, t, \ln r) \\ &\quad = \ln r + \frac{(\mu - 2) t^{1-k}}{4} \int_0^1 \int_0^1 \xi^{-1} \sigma^{\mu/2-2} \left(t^2 - \frac{r^2(1-\sigma)}{\xi} \right)_+^{(k-1)/2} d\sigma\, d\xi. \end{aligned} \tag{5.9}$$

By an application of (2.4), one can obtain the precise regions of integration according as $t \geqq r$ or $t < r$ namely

$$(\xi, \sigma) \in [0, 1] \times [0, 1] \cap \left\{ (\xi, \sigma) : \sigma \geqq \max\left(0, 1 - \frac{t^2 \xi}{r^2}\right) \right\}.$$

Example 4. Consider (1.1) with $\varphi(r) = \ln r$ and $\mu = 4$. Using (2.2) with (5.9) we obtain, by a partial integration, that

$$W^4(r, t, \ln r) = \begin{cases} t \ln r + \dfrac{t^3}{3r^2}, & t < r, \\ t \ln r + \dfrac{t^3}{3r^2} - \dfrac{1}{3r^2} \displaystyle\int_{r^2/t^2}^1 \left(t^2 - \frac{r^2}{\xi} \right)^{3/2} d\xi, & t \geqq r. \end{cases}$$

It is easy to see that W^4 and $\partial W^4 / \partial t$ are continuous at $t = r$.

(B) $\varphi(r) = r^{2-\mu} \ln r^2$. We can apply the transform (2.1′) directly to the solution of (1.3) corresponding to this data as given by [2, formula (7.2)]. Using (5.1) and (2.5) we

have

$$
\begin{aligned}
&E_k^{\mu}(r, t, r^{2-\mu}\ln r^2)\\
&= \frac{-\Gamma\left(\frac{k+1}{2}\right)\Gamma^{(1)}\left(\frac{k+3-\mu}{2}\right)t^{1-k}}{\Gamma\left(\frac{\mu}{2}-1\right)\left(\Gamma\left(\frac{k+3-\mu}{2}\right)\right)^2}\int_0^1 \sigma^{\mu/2-2}(t^2-r^2\sigma)_+^{(k+1-\mu)/2}\,d\sigma\\
&\quad+\frac{\Gamma\left(\frac{k+1}{2}\right)t^{1-k}}{\Gamma\left(\frac{\mu}{2}-1\right)\Gamma\left(\frac{k+3-\mu}{2}\right)}\int_0^1 \sigma^{\mu/2-2}(t^2-r^2\sigma)_+^{(k+1-\mu)/2}\ln(t^2-r^2\sigma)\,d\sigma\\
&\quad+\frac{\Gamma\left(\frac{k+1}{2}\right)t^{1-k}}{\Gamma\left(\frac{\mu}{2}-1\right)\Gamma\left(\frac{k+3-\mu}{2}\right)}\int_0^1 (1-\sigma)^{\mu/2-2}\ln\sigma\{t^2-r^2(1-\sigma)\}_+^{(k+1-\mu)/2}\,d\sigma\\
&\quad-\frac{\Gamma\left(\frac{k+1}{2}\right)t^{1-k}}{\Gamma\left(\frac{\mu}{2}-1\right)\Gamma\left(\frac{k+3-\mu}{2}\right)}\int_0^1 \sigma^{\mu/2-2}\ln\sigma(t^2-r^2\sigma)_+^{(k+1-\mu)/2}\,d\sigma\\
&\quad+\frac{\Gamma^{(1)}\left(\frac{\mu}{2}-1\right)\Gamma\left(\frac{k+1}{2}\right)t^{1-k}}{\left(\Gamma\left(\frac{\mu}{2}-1\right)\right)^2\Gamma\left(\frac{k+3-\mu}{2}\right)}\int_0^1 \sigma^{\mu/2-2}(t^2-r^2\sigma)_+^{(k+1-\mu)/2}\,d\sigma.
\end{aligned}
\tag{5.10}
$$

Example 5. Consider (1.1) with $\mu=3$ and $\varphi(r)=r^{-1}\ln r^2$. After some involved calculations, we can show that

$$
W^3(r, t, r^{-1}\ln r^2)=\begin{cases}\dfrac{t+r}{r}\ln(r+t)+\dfrac{t-r}{r}\ln(r-t)-\dfrac{2t}{r}, & t<r,\\[2ex] \left(1+\dfrac{t}{r}\right)\ln(t+r)+\left(1-\dfrac{t}{r}\right)\ln(t-r)-2, & t>r,\end{cases}
$$

and $\lim_{t\to r} W^3(r, t, r^{-1}\ln r^2)=2\ln 2r-2$. The graph of this W^3 function in the W-t "plane" for various values of r is given in Fig. 3. In Fig. 3 the dashed curve is the projection onto the W-t plane of the values $W(r, r, r^{-1}\ln r^2)$.

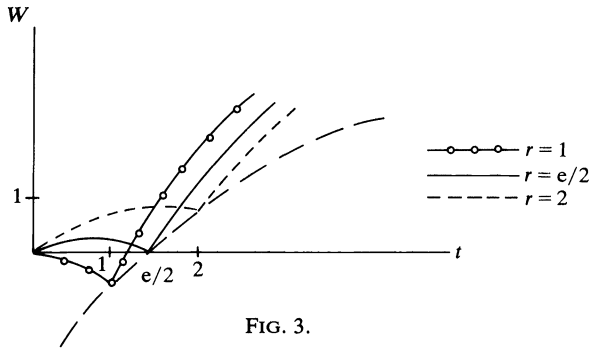


FIG. 3.

REFERENCES

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

[2] L. R. BRAGG, *The radial heat polynomials and related functions*, Trans. Amer. Math. Soc., 119 (1965), pp. 270–290.

[3] ———, *On the solution of radial heat problems with singular data*, SIAM J. Appl. Math., 15 (1967), pp. 1258–1271.

[4] ———, *Fundamental solutions and properties of solutions of the initial value radial Euler-Poisson-Darboux problem*, J. Math. Mech., 18 (1969), pp. 607–616.

[5] ———, *Hypergeometric operator series and related partial differential equations*, Trans. Amer. Math. Soc., 143 (1969), pp. 319–336.

[6] L. R. BRAGG AND J. W. DETTMAN, *Related partial differential equations and their applications*, SIAM J. Appl. Math., 16 (1968), pp. 459–467.

[7] ———, *An operator calculus for related partial differential equations*, J. Math. Anal., 22 (1968), pp. 261–271.

[8] ———, *Expansions of solutions of certain hyperbolic and elliptic problems in terms of Jacobi polynomials*, Duke Math. J., 36 (1969), pp. 129–144.

[9] ———, *Multinomial representation of solutions of a class of singular initial value problems*, Proc. Amer. Math. Soc., 21 (1969), pp. 629–634.

[10] W. FULKS, *Advanced Calculus*, John Wiley, New York, 1969.

[11] I. HIRSCHMAN AND D. WIDDER, *The Convolution Transform*, Princeton Univ. Press, Princeton, NJ, 1955.

[12] H. KAUFMANN AND G. ROBERTS, *Tables of Laplace Transforms*, W. B. Saunders, Philadelphia, 1966.

[13] W. MAGNUS, F. OBERHETTINGER AND R. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[14] A. TYCHONOFF, *Théorèmes d'unicité pour l'équation de la chaleur*, Mat. Sb., 42 (1935), pp. 119–215.

[15] D. WIDDER, *Positive temperatures on an infinite rod*, Trans. Amer. Math. Soc., 55 (1944), pp. 85–95.

[16] A. ZEMANIAN, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

[17] R. W. CARROLL, *Transmutation and Operator Differential Equations*, Notas de Mathematica 37, North-Holland, New York, 1979.

[18] R. W. CARROLL AND R. W. SHOWALTER, *Singular and Degenerate Cauchy problems*, Academic Press, New York, 1976

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0002 $01.00/0

SERIES OF ORTHOGONAL POLYNOMIALS AS BOUNDARY VALUES*

GILBERT G. WALTER† AND PAUL G. NEVAI‡

Abstract. The relations between expansions in orthogonal polynomials of generalized functions on the real axis and certain holomorphic functions in the upper and lower half planes are studied. The holomorphic functions are given by series of functions of the second kind which satisfy the same recurrence formula as the polynomials. A space of generalized functions associated with the polynomials is first introduced. Each element in this space has an analytic representation given by such series whose jump across the real axis is given by the element. Under certain conditions the singularities of the analytic representation may be related to singularities of the associated power series.

1. Introduction. Each function or generalized function f (in D') defined on the real line has an analytic representation consisting of a pair of holomorphic functions defined in the upper and lower half plane respectively [1]. The symmetric difference of these two functions tends to the given f as the real axis is approached.

For many of the classical orthogonal polynomials, such an analytic representation may be obtained from the second solution q_n to the differential equation satisfied by the polynomials p_n [8], [10], [11]. Thus if a function or generalized function f has an expansion in such orthogonal polynomials, say $\sum a_n p_n$, then the series $\sum a_n q_n$, if it converges, should supply the analytic representation. This series does converge for appropriate generalized functions when the p_n are Legendre polynomials [9], Hermite polynomials [10] or standard Laguerre polynomials [12].

This series can then be used to obtain results analogous to some well known properties of trigonometric series. Indeed a (π)-periodic generalized function (distribution) on the real axis has a cosine series $\sum a_n \cos nx$ with $a_n = O(n^p)$, and its analytic representation may be given by $\frac{1}{2}\sum a_n e^{inz}$ and $-\frac{1}{2}\sum a_n e^{-inz}$ which converge off the real axis. Their difference is a real harmonic function and their sum its conjugate. As the Im $z \to 0$ the difference approaches the original series and the sum the conjugate series which also converges to a generalized function.

In this work we shall extend the results for classical orthogonal polynomials to "arbitrary" ones. Since in this case no differential equation is available, we use the difference equation satisfied by the polynomials to obtain the functions of the second kind. We then introduce a space of generalized functions which are appropriate to these particular polynomials and which is contained in the space of distributions. Elements of this space have convergent expansions and have an analytic representation given by series of these functions of the second kind. These series in turn have many of the properties of trigonometric series.

2. Some basic properties. Let $\{p_n\}$ be the set of orthogonal polynomials on the finite or infinite interval (a, b) with respect to the probability measure given by the monotone function $\alpha(x)$. That is, $p_n(x)$ is a polynomial of degree n and

$$\int_a^b p_n(x)p_m(x)\,d\alpha(x) = \delta_{mn}, \qquad n, m = 0, 1, 2, \cdots, \tag{2.1}$$

where $p_0(x) \equiv 1$.

* Received by the editors August 26, 1980.

† Department of Mathematics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. This material is based upon work supported by the National Science Foundation under grant MCS78-01868.

The properties of such polynomials are developed in Szegö [8], in particular in Chapters 2, 3, 7, 12 and 13, as well as in a number of other sources. One such property is that the difference equation

$$p_n(x) = (A_n x + B_n)p_{n-1}(x) - C_n p_{n-2}(x), \qquad n = 1, 2, \cdots, \tag{2.2}$$
$$p_{-1}(x) = 0, \qquad p_0(x) = 1$$

is satisfied [8, p. 42].

Associated with p_n we have the analytic function (function of the second kind) $f_n(z)$ given for $z \in \mathbb{C} - [a, b]$. It is

$$f_n(z) = \frac{1}{2\pi i}\int_a^b \frac{p_n(x)}{x - z}\, d\alpha(x), \qquad n = 0, 1, \cdots, \tag{2.3}$$

and satisfies the same recurrence formula (2.2) as p_n except that the first two terms are

$$f_{-1}(z) = \frac{1}{2\pi i} \quad \text{and} \quad f_0(z) = \frac{1}{2\pi i}\int_a^b \frac{1}{x - z}\, d\alpha. \tag{2.4}$$

(See Geronimus [3, p. 57].) These functions of the second kind are related to the second solution q_n of the differential equation satisfied by p_n in the classical case. That is, if the $\{p_n\}$ are Jacobi, Laguerre or Hermite polynomials with weight function $w(x)$, then

$$q_n(z)w(z) = f_n(z), \qquad n = 0, 1, 2, \cdots, \quad \operatorname{Im} z \neq 0. \tag{2.5}$$

See respectively [8, p. 74], [3, p. 62] and [11].

We next observe that by (2.3) the $f_n(z)$ are the expansion coefficients of $1/2\pi i(x - z)$, and hence form a series which converges in $L^2(d\alpha; (a, b))$; i.e.,

$$(x - z)^{-1} = 2\pi i \sum_{n=0}^{\infty} f_n(z)p_n(x), \qquad x \in (a, b), \quad z \in \mathbb{C} - [a, b]. \tag{2.6}$$

For Jacobi polynomials this series converges pointwise when x lies inside and z outside an ellipse with foci at ± 1. The same is true for Laguerre and Hermite polynomials except that a parabola and a horizontal strip replace the ellipse. Of course we cannot expect to obtain such results in general, but we do have more restrictive ones.

PROPOSITION 2.1. *Let $d\alpha$ be an arbitrary measure on the finite interval (a, b) for which the infinite system $\{p_n\}_0^\infty$ exists. Then*

$$\sum f_n(z)p_n(x)$$

converges uniformly for $x \in [a, b]$ and $|\operatorname{Im} z| \geqq 1 + b - a$.

The proof will involve certain functions given by:

DEFINITION 2.1. The *Christoffel function* $\lambda_n(d\alpha, z)$ is defined by

$$\lambda_n(d\alpha, z) = \min_{\Pi(z)=1} \int_a^b |\Pi(t)|^2\, d\alpha(t), \tag{2.7}$$

where Π is an arbitrary polynomial of degree $n - 1$. It also satisfies

$$\lambda_n^{-1}(d\alpha, z) = \sum_{k=0}^{n-1} p_k^2(d\alpha, z); \tag{2.8}$$

see [2, p. 25].

LEMMA 2.2. *Let* $z \in \mathbb{C}$ *with* $\operatorname{Im} z \neq 0$. *Then*

$$|f_n(z)| \leqq \frac{1}{2\pi |p_n(z)||\operatorname{Im} z|}. \tag{2.9}$$

Proof. We first show that

$$4\pi^2 |f_n(z)|^2 \leqq \frac{\lambda_{n+1}(d\alpha, z)}{|\operatorname{Im} z|^2} \tag{2.10}$$

by calculating that

$$\begin{aligned} 2\pi i f_n(z)\Pi(z) &= \int_a^b \frac{\Pi(z)p_n(t)}{t-z}\, d\alpha(t) \\ &= \int_a^b \frac{\Pi(z)-\Pi(t)}{t-z}\, p_n(t)\, d\alpha(t) + \int_a^b \frac{\Pi(t)}{t-z}\, p_n(t)\, d\alpha(t). \end{aligned}$$

The first integral is zero since p_n is orthogonal to all polynomials of degree $<n$. Hence by Schwarz's inequality

$$\begin{aligned} 4\pi^2 |f_n(z)\Pi(z)|^2 &\leqq \int_a^b |\Pi(t)|^2 \frac{d\alpha(t)}{|t-z|^2} \int_a^b p_n^2(t)\, d\alpha(t) \\ &\leqq \int_a^b |\Pi(t)|^2 \frac{d\alpha(t)}{|\operatorname{Im} z|^2}. \end{aligned}$$

By taking the minimum over $\Pi(t)$ such that $\Pi(z) = 1$, we obtain (2.10).

We now apply (2.8) to (2.10) to obtain

$$4\pi^2 |f_n(z)|^2 \leqq \left[\sum_{k=0}^{n} |p_k^2(z)||\operatorname{Im} z|^2 \right]^{-1}, \tag{2.11}$$

from which the conclusion follows. □

Proof of Proposition 2.1. By the lemma it suffices to establish the convergence of

$$\sum \frac{|p_n(x)|}{|p_n(z)|} = \sum \prod_{k=1}^{n} \frac{|x - x_{kn}|}{|z - x_{kn}|},$$

where the x_{kn} denote the zeros of p_n (which are real and lie in (a, b)).

Clearly $|x - x_{nk}| < b - a$ and $|z - x_{nk}| \geqq |\operatorname{Im} z|$; hence

$$\frac{|p_n(x)|}{|p_n(z)|} \leqq \frac{(b-a)^n}{|\operatorname{Im} z|^n} \leqq \frac{(b-a)^n}{(1+b-a)^n}.$$

Thus the series converges uniformly and the conclusion follows. □

In the case of an infinite interval no such simple result holds. Another condition involving classical weight functions is needed. We specify them more precisely by

DEFINITION 2.2. By a *classical weight function* we denote the function $w(x) = (x-a)^\alpha (b-x)^\beta$ for a finite interval (a, b), or $w(x) = (x-a)^\alpha e^{-x}$ for a semi-infinite interval (a, ∞), or $w(x) = e^{-x^2/2}$ for $(-\infty, \infty)$, $\alpha \geqq 0$, $\beta \geqq 0$.

PROPOSITION 2.3. *Let* $\alpha'(x) \geqq \mu w(x)$ *a.e. on* (a, b) *where* $\mu > 0$ *and* w *is a classical weight function; let* $\{p_n\}$ *be the associated orthogonal polynomials of* $d\alpha$. *Then for each* $x_0 > 0$ *there exists a* $y_0 > 0$ *such that*

$$\sum f_n(z)p_n(t)w^{1/2}(t)$$

converges uniformly for $t \in (a, b)$ *and* $z \in S = \{x + iy : |y| \geqq y_0, |x| \leqq x_0\}$.

Proof. We use the fact that for such $w(x)$, $p_n(x)w^{1/2}(x) = O(n^k)$ for some integer k and $f_n(z) = O((y_0/z)^n)$ for z in such a strip. The norm of p_n with respect to w satisfies

$$\int_a^b p_n^2(x)w(x)\,dx \leqq \mu^{-1}\int_a^b p_n^2(x)\,d\alpha(x) = \mu^{-1}.$$

Hence by [8, p. 181]

$$|p_n^2(x)| \leqq \mu^{-1}\sum_{\nu=0}^{n}|\tilde{p}_\nu(x)|^2,$$

where $\tilde{p}_n$ is the classical orthogonal polynomial associated with $\tilde{w}$. Since each of the three classical types satisfies $\tilde{p}_n(x)w^{1/2}(x) = O(n^k)$, so does $p_n(x)w^{1/2}(x)$. To prove the other part we use the following.

LEMMA 2.4. *Let $\{p_n\}$ be orthogonal polynomials with respect to $d\alpha$ on (a, b). Then for each $x_0 > 0$, there exists $y_0 > 0$ such that $f_n(z) = O((y_0/z)^n)$ uniformly for z in the strip*

$$S_{x_0,y_0} = \{x + iy \mid |y| \geqq y_0, |x| \leqq x_0\}.$$

The proof is based on (2.3), which may be rewritten as

$$\begin{aligned} f_n(z)z^n &= \frac{1}{2\pi i}\int_a^b \frac{p_n(x)z^n}{x-z}\,d\alpha(x) \\ &= \frac{1}{2\pi i}\int_a^b \frac{z^n - x^n}{x-z}p_n(x)\,d\alpha(x) + \frac{1}{2\pi i}\int_a^b \frac{x^n}{x-z}p_n(x)\,d\alpha(x) \\ &= 0 + r_n(z), \end{aligned} \tag{2.12}$$

since p_n is orthogonal to x^k for $k < n$. The last integral $r_n(z)$ satisfies $r_n(0) = 0$ and $r_n(z) \to 0$ as $|\operatorname{Im} z| \to \infty$.

Hence the maximum of $|r_n(z)|$ in the strip $S_{x_0,1}$ occurs at some boundary point z_1. By taking $y_0 = |z_1|$ we observe that

$$|r_n(z)| \leqq \frac{|z_1|^n}{2|\operatorname{Im} z_1|}\int_a^b |p_n(x)|\,d\alpha(x), \qquad z \in S_{x_0,y_0},$$

from which the inequality follows. □

3. A space of generalized functions. In this section we study a space of generalized functions appropriate to a general class of orthogonal polynomials. We assume initially only that $d\alpha$ has a complete system $\{p_n\}_{n=0}^{\infty}$ of polynomials.

DEFINITION 3.1. Let f be a function on (a, b) such that $\sum n^2|a_n^2| < \infty$, where $a_n = \int f p_n\,d\alpha$. Then L is the operator given by

$$L(f) = L(\sum a_n p_n) = \sum n a_n p_n.$$

DEFINITION 3.2. Let A consist of all complex-valued functions $\phi(x)$ on (a, b) such that for each $k = 0, 1, 2, \cdots, L^k\phi \in L^2_{d\alpha}$; i.e.,

$$\rho_k^2(\phi) = \int_a^b |L^k\phi(x)|^2\,d\alpha(x) < \infty \qquad \left(= \sum n^{2k}|<\phi, p_n>|^2\right).$$

Remark 3.1. (a) L is self-adjoint with respect to A; i.e.,

$$\langle L\phi, \psi\rangle = \langle \phi, L\psi\rangle \quad \text{for } \phi, \psi \in A, \quad \text{where } \langle \phi, \psi\rangle = \int \phi\bar{\psi}\,d\alpha.$$

(b) A is a linear subspace of $L^2_{d\alpha}$.

(c) ρ_k is a seminorm on A.

(d) All polynomials belong to A.

(e) The topology generated by the family of seminorms $\{\rho_k\}$ is stronger than that induced by $L^2_{d\alpha}$. (See [13, p. 9] for the definition of this topology.)

LEMMA 3.1. *A with the topology generated by $\{\rho_k\}$ is complete.*

Proof. Let $\{\phi_n\}$ be a Cauchy sequence in A; then $\{L^k\phi_n\}$ is Cauchy and hence convergent in $L^2_{d\alpha}$ for $k = 0, 1, 2, \cdots$. Let the limits be ψ_k and denote by

$$\phi_n = \sum a_{nj}p_j, \qquad \psi_k = \sum b_{kj}p_j.$$

Then $L^k\phi_n \to \psi_k$, in $L^2_{d\alpha}$, becomes

$$\sum (j^k a_{nj} - b_{kj})^2 \to 0 \quad \text{as } n \to \infty.$$

But $j^k a_{nj} \to b_{kj}$ and $j^k a_{nj} \to j^k b_{0j}$; hence $j^k b_{0j} = b_{kj}$ or $\psi_k = L^k\psi_0$. Hence $\psi_0 \in A$ and is the limit in A of ϕ_n. □

LEMMA 3.2. *Let the $d\alpha$ satisfy $\alpha'(x) \geqq \mu w(x)$ a.e. in (a, b) for some classical w, $\mu > 0$. Then for each integer $k \geqq 0$, there is a $p > 0$ such that the kth derivative of p_n satisfies*

$$p_n^{(k)}(x)w^{1/2}(x) = O(n^p)$$

uniformly for $x \in (a, b)$, $p = p(k)$.

Proof. Let $\tilde{p}_n$ denote the orthogonal polynomials associated with w. Then

$$p_n(x) = \sum_{i=0}^{n} \lambda_i \tilde{p}_i(x), \tag{3.1}$$

and hence by Cauchy's inequality

$$|p_n^{(k)}(x)|^2 \leqq \sum_{i=0}^{n} |\lambda_i|^2 \sum_{i=0}^{n} |\tilde{p}_i^{(k)}(x)|^2. \tag{3.2}$$

But we have

$$1 = \int_a^b p_n^2(x)\, d\alpha(x) \geqq \mu \int_a^b p_n^2(x)w(x)\, dx, \tag{3.3}$$

and therefore by Bessel's inequality

$$\sum |\lambda_i|^2 \leqq \mu^{-1}. \tag{3.4}$$

Since the $\tilde{p}_n$ are Jacobi, Laguerre or Hermite polynomials, the inequality

$$\tilde{p}_n^{(k)}(x)w^{1/2}(x) = O(n^{\tilde{p}}) \tag{3.5}$$

holds uniformly on (a, b), (e.g., $\tilde{p} = 2k + \max(\alpha, \beta)$ in the Jacobi case [8, p. 170]). By substituting (3.5) into (3.2) we obtain the conclusion. □

If we are interested in the behavior of p_n on bounded intervals only, the hypothesis may be weakened considerably.

LEMMA 3.3. *Let $[c, d] \subseteq \operatorname{supp} d\alpha$ and let $(\alpha')^{-\varepsilon} \in L[c, d]$ for some $\varepsilon > 0$. Then there is a p such that*

$$p_n^{(k)}(x) = O(n^{p+2k}) \quad \textit{uniformly on } [c, d].$$

Proof. For such α it follows that for some $A > 1$

$$\Pi_n^2(x) \leqq n^A \int_a^b \Pi_n^2(t)\, d\alpha(t), \qquad x \in [c, d], \quad n \geqq 1,$$

for every Π_n and hence for p_n (see [5, p. 158]). Thus, for $k = 0$, p may be taken to be

$A/2$. By Markov's inequality,

$$p_n'(x) = O(n^{p+2}) \quad \text{on } [c, d],$$

and by iterating this expression the conclusion follows. □

DEFINITION 3.3. By $(f\,d\alpha)^{(-p)}$ we denote the pth order antiderivative of $f\,d\alpha$ given by

$$(f\,d\alpha)^{(-p)}(x) = \frac{1}{\Gamma(p)} \int_a^x f(t)(x-t)^{p-1}\,d\alpha(t)$$

for $f \in L^2_{d\alpha}$.

LEMMA 3.4. *Let $d\alpha$ satisfy*

$$\int w^{-1}\,d\alpha < \infty$$

for some classical weight w. Then for each $q > 0$, there is an integer k such that

$$(p_n\,d\alpha)^{(-k)}(x) = O(n^{-q})$$

uniformly on (a, b).

Proof. We first observe that the conclusion is true for the classical polynomials which correspond to the three weight functions given in Definition 2.2. This follows from Rodrigues' formulae [8, pp. 67, 101, 106]. Again let $\tilde{p}_n$ denote the classical polynomial. We obtain

$$\int_a^b p_n\,d\alpha\,\tilde{p}_m = \int_a^b \frac{p_n\,d\alpha}{w}\,\tilde{p}_m w = 0, \qquad m < n, \tag{3.6}$$

since p_n is orthogonal to any polynomial of degree $< n$. Hence the expansion of the measure $p_n\,d\alpha/w$ with respect to the $\tilde{p}_m$ is of the form

$$\frac{p_n\,d\alpha}{w} \sim \sum_{i=n}^{\infty} \lambda_{i,n}\tilde{p}_i. \tag{3.7}$$

But by Schwarz's inequality

$$\begin{aligned} |\lambda_{i,n}|^2 &\leqq \int p_n^2\,d\alpha \int \tilde{p}_i^2\,w\,\frac{d\alpha}{w} \\ &\leqq 1 \cdot \int w^{-1}\,d\alpha \sup_{a<x<b} \tilde{p}_i^2(x)w(x) = O(i^{2r}) \end{aligned} \tag{3.8}$$

for each of the classical cases for some $r > 0$.

We then multiply (3.7) by w and integrate k times to obtain

$$(p_n\,d\alpha)^{(-k)}(x) = \sum_{i=n}^{\infty} \lambda_{i,n}(\tilde{p}_i w)^{(-k)}(x). \tag{3.9}$$

For k sufficiently large the right side will converge uniformly on (a, b) and its terms will be dominated by terms of the form $O(i^{r-q})$ independent of n. Hence the conclusion follows. □

THEOREM 3.5. *Let $d\alpha$ be a measure such that, respectively,* (i) *for some $\varepsilon > 0$ $(\alpha')^{-\varepsilon} \in L_{\text{loc}}$ on (a, b),* (ii) $\int_a^b w^{-1}\,d\alpha < \infty$ *for some classical weight function w. Then, respectively:*

(i) *A is a subspace of $E(a, b)$ and convergence in A implies convergence in E.*

(ii) *$D(a, b)$ is a subspace of A and convergence in D implies convergence in A.*

Here $E(a, b)$ is the space of all C^∞ functions on (a, b), while $D(a, b)$ consists of those with compact support in (a, b). Convergence in E is uniform convergence of the derivatives $\phi_m^{(k)}$ to $\phi^{(k)}$ on compact subsets of (a, b). Convergence in D is the same with the added condition that all the ϕ_m have a common support.

Proof. Conclusion (i) is clear since, by Lemma 3.3, the series expansion of $\phi \in A$ may be differentiated k times to obtain

$$\phi^{(k)} = \sum_{n=k}^{\infty} a_n p_n^{(k)},$$

which will converge uniformly on compact subsets of (a, b). The same is true for

$$\phi_m = \sum a_{nm} p_n.$$

Since ϕ_m converges to ϕ in the sense of A, the sequence $\{a_{nm} n^{p+1}\}_{n=0}^{\infty}$ converges to $\{a_n n^{p+1}\}_{n=0}^{\infty}$ in the sense of l^2 for each integer $p > 0$. Hence on compact subsets of (a, b) we have

$$\left| \sum_{n=k}^{\infty} a_{nm} p_n^{(k)} - \sum_{n=k}^{\infty} a_n p_n^{(k)} \right| \leqq \sum_n |a_{nm} - a_n| \, |p_n^{(k)}|$$

$$\leqq C \sum_n |a_{nm} - a_n| n^p \leqq C \left[\sum_n |a_{nm} - a_n|^2 n^{2p+2} \right]^{1/2} [\sum n^{-2}]^{1/2}$$

for some constant C.

Thus $\phi_m^{(k)} \to \phi^{(k)}$ uniformly on compact subsets of (a, b).

Conclusion (ii) follows from the fact that for $\phi \in D$ the expansion coefficients satisfy

$$|a_n| = \left| \int_a^b \phi p_n \, d\alpha \right| = \left| \int_a^b \phi^{(k)} (p_n \, d\alpha)^{(-k)} \right|$$

$$\leqq \int_a^b |\phi^{(k)}| \| (p_n \, d\alpha)^{(-k)} \|_\infty.$$

By Lemma 3.4 $a_n = O(n^{-q})$ and hence $\phi \in A$. The convergence in D which consists of having each ϕ_m have support in a common closed subinterval of (a, b) and having its derivatives converge uniformly implies convergence in A by the same considerations. □

We are interested in extending the set of functions which have expansions in terms of $\{p_n\}$ to a space sufficiently large to include all functions which are potentially interesting. To do this we consider the dual space A' of A consisting of all continuous linear functionals on A. Ordinary functions may be embedded in A' by using the convention

$$\langle f, \phi \rangle = \int f \phi \, d\alpha.$$

That is, the functional whose values are $\langle f, \phi \rangle$ is associated with the function f. Since A is a complete countably normed space, its dual is a space of generalized functions [13, p. 39] on (a, b).

THEOREM 3.6. *Let $d\alpha$ be a measure such that for some $\varepsilon > 0$, $(\alpha')^{-\varepsilon} \in L_{\text{loc}}$ and $\int w^{-1} \, d\alpha < \infty$ for some classical weight function w on (a, b). Then the space of continuous linear functionals A' of A satisfies the following*:

(i) $L^2_{d\alpha} \subset A'$.

(ii) $E' \subset \alpha' A' \subset D'$.

(iii) $f \in A'$ *if and only if* $\sum \langle f, p_n \rangle p_n$ *converges to* f *in the sense of* A'.

(iv) $\sum b_n p_n$ *converges in* A' *if and only if* $b_n = O(n^p)$ *for some* $p > 0$.

If furthermore $\alpha'(x) > \mu w(x)$ *for some* $\mu > 0$, *then*:

(v) *Differentiation is a continuous operation in* A *and the operation* Δ *given by* $\langle \Delta f, \phi \rangle = -\langle f, \phi' \rangle$ *is continuous in* A'.

Proof of (v). Let $\phi_m \to 0$ in the sense of A and let

$$\phi_m = \sum_{j=0}^{\infty} a_{mj} p_j.$$

Then we have

$$\phi'_m = \sum_{j=1}^{\infty} a_{mj} p'_j = \sum_{j=1}^{\infty} a_{mj} \sum_{k=0}^{j-1} \lambda_{jk} p_k = \sum_{k=0}^{\infty} \Big(\sum_{j=k+1}^{\infty} a_{mj} \lambda_{jk} \Big) p_k, \tag{3.10}$$

where

$$|\lambda_{jk}| \leqq \Big(\int (p'_j)^2 \, d\alpha \Big)^{1/2} = O(j^{r/2})$$

for some $r > 0$ by Lemma 3.2. Hence the series $\sum \lambda_{jk}^2 j^{-r-2}$ converges and the coefficients in (3.10) satisfy

$$\Big(\sum_{j=k+1}^{\infty} a_{mj} \lambda_{jk} \Big)^2 k^q \leqq \sum_{j=k+1}^{\infty} a_{mj}^2 j^{r+2+q} \sum_{j=1}^{\infty} \lambda_{jk}^2 j^{-r-2}, \qquad q > 0. \tag{3.11}$$

Since the first factor on the right converges to 0 as $m \to \infty$ and the second is bounded, it follows that $\phi'_m \to 0$ in the sense of A. The proof of the other parts follows from standard techniques for dual spaces and is similar to classical cases [13, p. 257]. □

4. Analytic representation of elements A'. In this section we develop some properties of the analytic representation $\hat{f}$ of the generalized function $f \in A'$ given in terms of series of functions of the second kind f_n.

DEFINITION 4.1. Let $f \in A'$, $f = \sum_{n=0}^{\infty} a_n p_n$; then the *analytic representation* $\hat{f}$ of f is the continuation to the upper and lower open half plane of the function given by

$$\hat{f}(z) = \sum a_n f_n(z).$$

Remark 4.1. The series $\sum a_n f_n$ converges for $z \in S_{x_0, y_0}$ since by Lemma 2.2 the $f_n(z) = O((y_0/z)^n)$ and by Theorem 3.5 $a_n = O(n^p)$.

PROPOSITION 4.1. *Let* $d\alpha$ *be as in Theorem* 3.5, *and let* $f \in A'$. *Then there exists an integer* $p > 0$ *and a bounded continuous function* F *on* (a, b) *such that*:

(i) $$\alpha' f = D^p F \quad \text{in } D'(a, b).$$

(ii) $$\hat{f}(z) = \left\{ \frac{p!}{2\pi i} \int_a^b \frac{F(x)}{(x - z)^{p+1}} \, dx + \rho(z) \right\}$$

for $\operatorname{Im} z \neq 0$, $\rho(z)$ *a polynomial in* $(b - z)^{-1}$ *of degree* $\leqq p$.

(iii) $$\hat{f}(z) = \frac{1}{2\pi i} \langle f, (\cdot - z)^{-1} \rangle, \qquad \operatorname{Im} z \neq 0.$$

Proof. Let $f \in A'$ and be given by $f = \sum a_n p_n$. Then by Theorem 3.5 there is a q such that $a_n = O(n^q)$. By Lemma 3.3, to this q there corresponds an integer $p \geqq 0$ such that $(p_n \, d\alpha)^{(-p)}(x) = O(n^{-q-2})$ uniformly for $x \in (a, b)$. Hence $F(x) = \sum a_n (p_n \, d\alpha)^{(-p)}(x)$ converges uniformly on (a, b) to the bounded continuous function $F(x)$.

Now F, because it is bounded, is also in D'. Hence both F and the series may be differentiated, and since D is a continuous operator on D',

$$D^pF = \sum a_n p_n \alpha'.$$

But $f \in A'$ corresponds to $\alpha' f \in D'$, which is given by the same series; hence conclusion (i) follows.

For conclusion (ii) we write

$$\begin{aligned}\frac{p!}{2\pi i}\int_a^b \frac{F(x)}{(x-z)^{p+1}}\,dx &= \frac{p!}{2\pi i}\int_a^b \frac{\sum a_n (p_n\,d\alpha)^{(-p)}(x)}{(x-z)^{p+1}}\,dx \\ &= \sum_{n=0}^{\infty} a_n \frac{p!}{2\pi i}\int_a^b \frac{(p_n\,d\alpha)^{(-p)}(x)}{(x-z)^{p+1}}\,dx \\ &= \sum_{n=0}^{\infty} a_n \frac{1}{2\pi i}\Big\{ -\frac{(p_n\,d\alpha)^{(-p)}(b)(p-1)!}{(b-z)^p} - \frac{(p_n\,d\alpha)^{(-p+1)}(b)(p-2)!}{(b-z)^{p-1}} - \cdots \\ &\qquad - \frac{(p_n\,d\alpha)^{(-2)}(b)1!}{(b-z)^2} - \frac{(p_n\,d\alpha)^{(-1)}(b)}{(b-z)} + \int_a^b \frac{p_n(x)}{x-z}\,d\alpha(x)\Big\},\end{aligned} \tag{4.1}$$

by integration by parts. For $n \geqq p$, $(p_n\,d\alpha)^{(-p)}(b) = (p_n\,d\alpha)^{(-p+1)}(b) = (p_n\,d\alpha)^{(-1)}(b) = 0$ since p_n is orthogonal to polynomials of degree $<n$. In the case $b = \infty$ the integrated terms must be interpreted as

$$\frac{(p_n\,d\alpha)^{(-p)}(b)}{(b-z)^p} = \lim_{x\to\infty} \int_a^x \frac{\dfrac{(x-t)^{p-1}}{(p-1)!}\,p_n(t)\,d\alpha(t)}{(x-z)^p},$$

which is zero for all values of n. Hence (4.1) becomes

$$\begin{aligned}\frac{p!}{2\pi i}\int_a^b \frac{F(x)}{(x-z)^{p+1}}\,dx &= \sum_{n=0}^{p-1} a_n \sum_{i=0}^{p} \lambda_{ni}(b-z)^{-i} + \sum_{n=0}^{\infty} a_n \frac{1}{2\pi i}\int_a^b \frac{p_n(x)\,d\alpha(x)}{(x-z)} \\ &= -\rho(z) + \sum_{n=0}^{\infty} a_n f_n(z).\end{aligned} \tag{4.2}$$

For z in the strip S_{x_0,y_0}, this latter series converges to $\hat{f}(z)$.

We next observe that $(x-z)^{-1} \in A$ for fixed z such that $\operatorname{Im} z \neq 0$. This follows from the fact that, for large n,

$$\int_a^b (x-z)^{-1} p_n(x)\,d\alpha(x) = \frac{p!}{2\pi i}\int_a^b \frac{(p_n\,d\alpha)^{(-p)}(x)}{(x-z)^{p+1}}\,dx = O(n^{-k}), \tag{4.3}$$

as in (4.1) for each integer $k \geqq 0$. This gives the other expression for $\hat{f}(z)$, (iii). □

LEMMA 4.2. *Let $\psi \in D$. Then for each integer $p \geqq 0$*

$$D_x^p\{\hat{\psi}(x+iy) - \hat{\psi}(x-iy)\} \to \psi^{(p)}(x) \quad \text{as } y \to 0^+$$

uniformly on compact subsets of (a, B), where

$$\hat{\psi}(z) = \frac{1}{2\pi i}\int_a^b \frac{\psi(x)}{x-z}\,dx. \tag{4.4}$$

Proof. This result is merely a form of Abel's theorem, since $\psi^{(p)}$ has compact support in (a, b) and the harmonic function $D^p\{\hat{\psi}(x+iy) - \hat{\psi}(x-iy)\}$ has $\psi^{(p)}$ as its boundary value. Such harmonic functions converge uniformly on compact subsets to their continuous boundary values [7, p. 65]. □

THEOREM 4.3. *Let $d\alpha$ be as in Proposition* 4.1, *and let $f \in A'$. Then*:

(i) $\hat{f}(\cdot + iy) - \hat{f}(\cdot - iy) \to f$ *in the sense of* D' *as* $y \to 0^+$.

(ii) *If f is a Peano derivative of some order at $x_0 \in (a, b)$, then the convergence is pointwise at x_0.*

Proof. Let $f_y(x) = \hat{f}(x + iy) - \hat{f}(x - iy)$; clearly $f_y \in D'$. Let $\phi \in D$; then

$$\begin{aligned}
\langle f_y, \phi \rangle &= \int_a^b f_y(t) \cdot \phi(t)\, dt \\
&= \int_a^b \frac{p!}{2\pi i} \left\{ \int_a^b \frac{F(x)}{(x - t - iy)^{p+1}}\, dx + \rho(t + iy) \right. \\
&\qquad \left. - \int_a^b \frac{F(x)}{(x - t + iy)^{p+1}}\, dx - \rho(t - iy) \right\} \phi(t)\, dt \qquad (4.5) \\
&= (-1)^{p+1} \int_a^b F(x) D_x^p \{ \hat{\phi}(x - iy) - \hat{\phi}(x + iy) \}\, dx \\
&\quad + \int_a^b \frac{p!}{2\pi i} (\rho(t + iy) - \rho(t - iy)) \phi(t)\, dt \\
&\to (-1)^p \int_a^b F(x) \phi(x)^{(p)}\, dx + 0
\end{aligned}$$

by Lemma 4.2 and the fact that ρ is analytic inside (a, b). Hence we have

$$\langle f_y, \phi \rangle \to \langle D^p F, \varphi \rangle \quad \text{for } \phi \in D. \qquad (4.6)$$

In order to prove the statement about Peano derivatives, we first observe that Proposition 4.1 holds for all p greater than the given value. Hence we may assume that the Peano derivative $f(x_0)$ (see [14, II, p. 59] for the definition) has the same order p and is given by

$$f(x_0) = \lim_{x \to x_0} \frac{(F(x) + P(x))p!}{(x - x_0)^p}, \qquad (4.7)$$

where $P(x)$ is a polynomial of degree $\leqq p - 1$. We define $G(x)$ to be the continuous bounded function in $L^2(a, b)$,

$$G(x) = \begin{cases} \dfrac{(F(x) + P(x))p!}{(x - x_0)^p}, & x \neq x_0, \\ f(x_0), & x = x_0. \end{cases} \qquad (4.8)$$

Then we have, for $\operatorname{Im} z \neq 0$,

$$\hat{f}(z) = \frac{1}{2\pi i} \int_a^b \frac{G(t)}{(t - z)^{p+1}} (t - x_0)^p\, dt + \text{holomorphic function}. \qquad (4.9)$$

We then need to show that $\hat{f}_y(x_0)$ converges to $f(x_0) = G(x_0)$. Since

$$\begin{aligned}
\hat{f}_y(x_0) &= \frac{1}{2\pi i} \int_a^b G(t)(t - x_0)^p \left\{ \frac{1}{(t - x_0 - iy)^{p+1}} - \frac{1}{(t - x_0 + iy)^{p+1}} \right\} dt \\
&= \int_a^b G(t) K_y(t - x_0)\, dt,
\end{aligned} \qquad (4.10)$$

we need merely show therefore that $K_y(t-x_0)$ is in a quasi-positive kernel (see [14, I, p. 86]), i.e., that it satisfies

(4.11)

(a) $$\int_{-\infty}^{\infty} K_y \to 1 \quad \text{as } y \to 0^+,$$

(b) $$\int_{-\infty}^{\infty} |K_y| \leqq A, \qquad y > 0,$$

(c) for each $\alpha > 0$, $K_y(t) \to 0$ uniformly for $|t| > \alpha$ as $y \to 0^+$.

We integrate by parts p times to deduce that

$$\int K_y = \int P_y = 1,$$

where $P_y(t) = (1/\pi)(y/(t^2+y^2))$, the Poisson kernel. Since $K_y(t) = (1/y)K_1(t/y)$, we need merely to show that $\int |K_1| < \infty$ to deduce that (b) is true. But this is obvious, as is (c), and hence conclusion (b) follows. □

COROLLARY 4.4. *Let $f \in A'$; then the function given by*

$$u(x, y) = \sum a_n \{f_n(x+iy) - f_n(x-iy)\}$$

is harmonic in the upper half plane and converges to f in the sense of $D'(a, b)$ as $y \to 0^+$.

We may also define a conjugate harmonic function

(4.12) $$v(x, y) = \sum i a_n \{f_n(x+iy) + f_n(x-iy)\}.$$

Now, however, the limit of $v(x, y)$ as $y \to 0^+$ is not necessarily in A'. In fact, for the Hermite polynomials, for example, the space A' consists of distributions f in D' for which $fe^{-x^2/2}$ is a tempered distribution. But even $[f_0(x+iy)+f_0(x-iy)]e^{-x^2/2}$ is $O(e^{x^2/2})$, and hence is not in A'. However, we do have:

COROLLARY 4.5. *Let $f \in A'$; then the harmonic conjugate $v(x, y) \to \tilde{f} \in D'$ in the sense of D' as $y \to 0^+$.*

5. Singular points. In the case of classical orthogonal polynomials, the location of singular points of the analytic representation of an expansion can sometimes be related to singular points of an associated power series (see [9], [4]). However, this seems to work only in the case of a finite interval (a, b). Hence we shall restrict ourselves to such an interval, which we take to be $(-1, 1)$.

We shall also place a further restriction on $d\alpha$, or rather on the recurrence formula (2.1). Since for $d\alpha$ considered in § 4, $A_n \to 2$, $B_n \to 0$, and $C_n \to 1$ as $n \to \infty$, [8, p. 310] we shall assume that each differs from the limit by a polynomial in n^{-1}. That is, we assume

(5.1) $$A_n = 2 + \sum_{i=1}^{k} \frac{a_i}{n^i}, \quad B_n = \sum_{i=1}^{k} \frac{b_i}{n^i}, \quad C_n = 1 + \sum_{i=1}^{k} \frac{c_i}{n^i}, \qquad n = 1, 2, \cdots.$$

This gives us the well-known Pollaczek polynomials. Pollaczek [6] shows that the singularities of the "kernel"

(5.2) $$K(r, x) = \sum_{n=1}^{\infty} r^n p_n(x), \quad |r| < 1, \qquad x \in (-1, 1)$$

are located at $r = 0$ and at $r = x \pm \sqrt{x^2-1}$ [6, p. 13]. The kernel may be extended to complex values of x with the same conclusion about singularities. In fact, the series

defining $K(r, x)$ is convergent for x in the interior of an ellipse passing through this point [8, p. 312].

The same conclusions follow for the kernel

$$\tilde{K}(r, z) = \sum_{n=1}^{\infty} r^n f_n(z), \qquad |r| < 1, \quad z \in S \tag{5.3}$$

except that it now converges for z outside of the ellipse. This can be used to locate the singular points of the analytic representation of an element of A'. This was done for Legendre series in [9] and Gegenbauer series in [12], and makes use of Hadamard's "multiplication of singularities" procedure.

THEOREM 5.1. *Let $d\alpha$ be a measure on $[-1, 1]$ satisfying the conditions of Proposition* 4.1 *such that $\{p_n\}$ satisfies the recurrence formula* (2.2) *with coefficients satisfying* (5.1); *let $f \in A'$ with series expansion*

$$f \sim \sum a_n p_n$$

and analytic representation

$$\hat{f}(z) = \sum a_n f_n(z).$$

Then $\hat{f}(z)$ is singular at the point x_0 in $(-1, 1)$ if and only if the function

$$\phi(w) = \sum_{n=0}^{\infty} a_n w^n$$

has singular points on the unit circle at $e^{\pm i\theta_0}$ such that $\cos \theta_0 = x_0$.

If we make the observation that $a_n = O(n^p)$, the proof is exactly that given in [9] for Legendre polynomials and will be omitted.

REFERENCES

[1] H. BREMERMANN, *Distributions, Complex Variables and Fourier Transforms*, Addison-Wesley, Reading, MA, 1965.

[2] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, Oxford, 1971.

[3] JA. L. GERONIMUS, *Orthogonal polynomials*, Amer. Math. Soc. Transl., Series 2, 108 (1977), pp. 37–130.

[4] R. P. GILBERT, *Function Theoretic Methods in Partial Differential Equations*, Academic Press, New York, 1969.

[5] P. G. NEVAI, *Orthogonal polynomials*, Mem. Amer. Math. Soc., 18 (1979).

[6] F. POLLACZEK, *Sur une généralisation des polynomes de Jacobi*, Mémorial des Sci. Math., XXI, Paris, 1956.

[7] E. M. STEIN, *Singular Integrals and Differentiability Property of Functions*, Princeton Univ. Press, Princeton, NJ, 1970.

[8] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1959.

[9] G. WALTER, *On real singularities of Legendre expansions*, Proc. Amer. Math. Soc., 19 (1968), pp. 1407–1412.

[10] ———, *Hermite series as boundary values*, Trans. Amer. Math. Soc., 218 (1976), pp. 155–171.

[11] ———, *Hermite series solutions of differential equations*, J. Differential Equations, 10 (1971), pp. 1–16.

[12] A. ZAYED, *Generalized functions and boundary value problems*, Ph.D. thesis, Univ. of Wisconsin, Milwaukee, WI, 1979.

[13] A. H. ZEMANIAN, *Generalized Integral Transformations*, Interscience, New York, 1968.

[14] A. ZYGMUND, *Trigonometric Series* I, II, Cambridge, University Press, London, 1959.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0003 $01.00/0

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF NONLINEAR VOLTERRA EQUATIONS WITH COMPLETELY POSITIVE KERNELS*

PH. CLÉMENT† AND J. A. NOHEL‡

Abstract. We consider the nonlinear Volterra equation

(V) $$u(t)+(b*Au)(t)\ni f(t), \qquad 0\leqq t<\infty$$

in the general setting $b:[0,\infty)\to R$ a given kernel, A a nonlinear m-accretive operator on a real Banach space X, $f:[0,\infty)\to X$ a given function and $*$ the convolution. We study the existence of positive solutions of (V) and their asymptotic behavior as $t\to\infty$, together with estimates of their rates of decay, under physically reasonable assumptions on b, A, f motivated by the problem of heat flow in materials with memory. The concept of complete positivity of the kernel b and its characterization play a crucial role in the analysis.

1. Introduction. In this paper we discuss the positivity of solutions, and their asymptotic behavior as $t\to\infty$, of the nonlinear Volterra equation

(V) $$u(t)+(b*Au)(t)\ni f(t), \qquad 0\leqq t<\infty$$

in the general setting: $b:[0,\infty)\to R$ is a given kernel, A is a nonlinear (possibly multivalued) m-accretive operator defined on a real Banach space X, $f:[0,\infty)\to X$ is a given function and $*$ denotes the convolution on $[0,t]$: $(b*z)(t)=\int_0^t b(t-\tau)z(\tau)\,d\tau$; the integral in (V) is understood in the sense of Bochner. The assumptions which are imposed on b, A, f are motivated by the problem of nonlinear heat flow in a material with memory discussed in § 4, in which the general positivity and asymptotic theory developed in §§ 2, 3 is applied to this physical problem. A different application of the general theory is given in Example 3.4 of § 3 to a nonlinear conservation law with memory.

The present study generalizes and complements earlier work of Clément and Nohel [3] on positivity and of Clément [2] on limiting behavior of positive solutions of (V). The generalization enables us to discuss the physical problem described in § 4. General existence, uniqueness and continuous dependence results of solutions of (V) which need not be positive have been established by Crandall and Nohel [5] and by Gripenberg [6]; these will be referred to as needed.

We will motivate the assumptions on the kernel b which will be needed throughout the analysis by means of a simple linear problem at the end of this section. These considerations suggest the concept of complete positivity of the kernel b (Definition 1.1 below) which plays an important role in the analysis. Some properties and a useful characterization of completely positive kernels are obtained in § 2.

We shall consider equation (V) in the slightly less general form

(V_g) $$u(t)+(b*Au)(t)\ni u_0+(b*g)(t), \qquad 0\leqq t<\infty\ .$$

We assume throughout the following minimal assumptions:

* Received by the editors May 8, 1980, and in revised form November 19, 1980.

† University of Technology, Delft, the Netherlands and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. The research of the author was sponsored by the U.S. Army under contracts DAAG29-75-C-0024 and DAAG29-80-C-0041.

‡ Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. The research of this author was sponsored by the U.S. Army under contracts DAAG29-75-C-0024 and DAAG29-80-C-0041 and grant DAAG29-77-G0004.

$$b \in L^1_{\mathrm{loc}}(0, \infty);$$

(H_1) A m-accretive in a real Banach space X;

$$u_0 \in \overline{D(A)} \quad \text{and} \quad g \in L^1_{\mathrm{loc}}(0, \infty; X).$$

The motivation for taking $f = u_0 + b * g$ in (V) is given in § 3 (see the argument at the beginning of § 3 following (V_g)). The main results of this paper give a rather complete description of the asymptotic behavior of the positive solutions of the abstract equation (V) as $t \to \infty$, including a priori estimates for their rates of decay. The results are then applied to the physical problem described above.

The additional assumption we shall make on the kernel b in order to insure positivity of solutions was first introduced in [3]; it is motivated by the following remark. If $b \equiv 1$ then (V_g) reduces to the evolution equation

(DE) $$\frac{du}{dt} + Au \ni g, \qquad u(0) = u_0.$$

It is well known [1] that if the resolvent $J_\lambda = (I + \lambda A)^{-1}$ of A maps a closed convex cone P of X into itself for every $\lambda > 0$, then $u(t) \in P$ for all $t > 0$, provided that $u_0 \in P$ and $g(t) \in P$ a.e. on $[0, \infty)$. Let us take, for instance,

$$X = \{u \in C[a, b] | u(a) = u(b) = 0\}$$

equipped with the supremum norm, $D(A) = \{u \in X | u \in C^2[a, b]$ and $u_{xx} \in X\}$ and $Au(x) = -u_{xx}(x)$ for $u \in D(A)$. It is standard that A is m-accretive in X. Moreover, if $P = \{u \in X | u(x) \geqq 0 x \in [a, b]\}$, then $J_\lambda P \subset P$ for every $\lambda > 0$; thus, as is classical, the solution of the heat equation is nonnegative provided that the initial value u_0 and the forcing form g are nonnegative.

We want to consider a class of kernels b under which the solution of (V) (resp. (V_g)) preserves this positivity property. This requirement is useful and natural in the application to the model of heat flow in a material with memory discussed in § 4, and in Example 3.4 of § 3.

Consider (V_g) with $Au = -u_{xx}$ with $D(A)$ as in the above example. It is easy to give necessary conditions to be imposed on b in order that positivity be preserved by (V_g) whenever u_0 and g are positive. Let $\bar{\lambda}$ denote the principal eigenvalue and $\bar{u}$ the corresponding principal eigenfunction of A, normalized by $\max_{x \in [a, b]} \bar{u}(x) = 1$. Clearly $\bar{\lambda} = (\pi/(b-a))^2$ and $\bar{u}(x) = \sin(\pi/(b-a))(x-a)$. If $u_0 = \alpha \bar{u}$, $g(t) = \bar{\lambda}\beta(t)\bar{u}$ with $\alpha \geqq 0$ and $\beta(t) \geqq 0$ where $\beta \in L^1_{\mathrm{loc}}(0, \infty)$, then, as can be verified directly, the strong solution of (V_g) is

(1.1) $$u(t) = [\alpha s(\bar{\lambda} b)(t) + (\beta * r(\bar{\lambda} b))(t)]\bar{u}, \qquad 0 \leqq t < \infty,$$

where the functions $s(b)$ and $r(b) : [0, \infty) \to R$ are respectively solutions of the linear Volterra equations

$(s(b))$ $$s(b)(t) + (b * s(b))(t) = 1, \qquad 0 \leqq t < \infty,$$

$(r(b))$ $$r(b)(t) + (b * r(b))(t) = b(t), \qquad 0 \leqq t < \infty.$$

Recall the standard fact (see, e.g., R. K. Miller [11]) that if $b \in L^1_{\mathrm{loc}}(0, \infty)$ the functions $s(b)$, $r(b)$ are uniquely defined and $s(b)$, $r(b) \in L^1_{\mathrm{loc}}(0, \infty)$. Moreover, if $F \in L^1_{\mathrm{loc}}(0, \infty)$ the unique solution of the linear Volterra equation

(1.2) $$u(t) + (b * u)(t) = F(t), \qquad 0 \leqq t < \infty$$

is given by

$$u(t) = F(t) - (r(b) * F)(t), \qquad 0 \leqq t < \infty. \tag{1.3}$$

In particular, taking $F \equiv 1$ in (1.2), we have

$$s(b)(t) = 1 - \int_0^t r(b)(\tau)\, d\tau, \qquad 0 \leqq t < \infty, \tag{1.4}$$

so that $s(b)$ is absolutely continuous on $[0, \infty)$ whenever $b \in L^1_{\text{loc}}(0, \infty)$. The function $s(b)$ is called the fundamental solution of (1.2), while the function $r(b)$ is called the resolvent kernel associated with b. The reader should also recall (see [5]) that *a strong solution of the abstract Volterra equation* (V_g) *on* $[0, T]$ is a function $u:[0, T] \to X$ such that $u \in L^1(0, T; X)$, $u(t) \in D(A)$ a.e. on $[0, T]$ and there exist

$$v \in L^1(0, T; X), \qquad v(t) \in Au(t)$$

on $[0, T]$ such that $u(t) + (b * v)(t) = u_0 + (b * g)(t)$ a.e. on $[0, T]$.

Returning to the solution (1.1) of (V_g) with $Au = -u_{xx}$, $\bar{\lambda} > 0$, $\bar{u}$, u_0, g defined above, we note that $\bar{u}(x) > 0$ for $x \in (a, b)$. Thus the solution $u(t)$ will be nonnegative for every $\alpha \geqq 0$ and for every $\beta \in L^1_{\text{loc}}(0, \infty)$, $\beta \geqq 0$, only if the functions $r(b)$, $s(b)$ are nonnegative on $[0, \infty)$. Moreover, if one imposes the requirement that the solution (1.1) of (V_g) should be nonnegative and independent of the length of the interval (a, b), it is clear that both of the functions $r(\lambda b)$ and $s(\lambda b)$ must be nonnegative for every $\lambda > 0$. We remark that these latter necessary conditions imposed on the kernel b have been shown to be sufficient to guarantee the preservation of positivity by the solution operator of the nonlinear equation (V) in the general case of A m-accretive on X (see [3, Thm. 4.5]).

The above considerations suggest the following concept of complete positivity of the kernel b:

DEFINITION 1.1. We shall say that the kernel b is *completely positive on* $[0, T]$ if $b \in L^1(0, T)$ and if the functions $r(\lambda b)$ and $s(\lambda b) = 1 - 1 * r(\lambda b)$ are nonnegative on $[0, T]$ for every $\lambda > 0$.

Some known sufficient conditions which insure the complete positivity of the kernel b on $[0, T]$ are:

(i) $b \in L^1(0, T)$ is nonnegative, nonincreasing and $\log b$ is convex (see Miller [10], Levin [8], Clément and Nohel [3]).

(ii) (special case of (i)). $b \in L^1(0, T)$ and b is completely monotonic on $(0, T)$ (see Miller [10]).

2. Completely positive kernels. In this section we give an alternate and useful characterization of completely positive kernels (Theorem 2.2) which will be needed for the development of the asymptotic properties of positive solutions of the abstract Volterra equation (V_g). For this purpose we consider the linear scalar Volterra equation (1.2) in the form

$$u + b * u = u_0 + b * g, \tag{2.1}$$

where $b \in L^1(0, T)$, $u_0 \in R$, $g \in L^1(0, T)$ and $T > 0$. Its unique solution (see (1.3), (1.4)) is given by

$$u(t) = u_0 s(b)(t) + (r(b) * g)(t), \qquad 0 \leqq t \leqq T. \tag{2.2}$$

In the following proposition we list some elementary properties of completely positive kernels which are needed in the sequel.

PROPOSITION 2.1. *Assume that b is completely positive on $[0, T]$ for some $T>0$. Then:*

1) *b is nonnegative on $[0, T]$ and, for every $\mu>0$, $s(\mu b)$ is nonnegative and nonincreasing on $[0, T]$.*

2) *For every $\mu>0$, $r(\mu b)$ is itself completely positive on $[0, T]$.*

Next, assume b is completely positive on $[0, T]$ for every $T>0$. Then:

3) *If $b \in L^1[0, \infty)$, then for every $\mu>0$*

$$\lim_{t\to\infty} s(\mu b)(t) = \left(1+\mu\int_0^\infty b(\tau)\,d\tau\right)^{-1},$$

$$\|r(\mu b)\|_{L^1(0,\infty)} = \left(\mu\int_0^\infty b(\tau)\,d\tau\right)\left(1+\mu\int_0^\infty b(\tau)\,d\tau\right)^{-1}.$$

4) *If $b \notin L^1(0, \infty)$, then for every $\mu>0$*

$$\lim_{t\to\infty} s(\mu b)(t) = 0 \quad \text{and} \quad \|r(\mu b)\|_{L^1(0,\infty)} = 1.$$

5) *If $b \notin L^1(0, \infty)$ and $b \in AC[0, \infty)$, then for every $\mu>0$, $r(\mu b) \in C[0, \infty]$ and* $\lim_{t\to\infty} r(\mu b)(t) = 0$.

Proof. 1) For every $\lambda>0$, $v_\lambda := \lambda^{-1} r(\lambda b)$ satisfies

$$v_\lambda + \lambda b * v_\lambda = b. \tag{2.3}$$

From the convolution theorem we have

$$\|v_\lambda\|_{L^1(0,T)} \leqq \|b\|_{L^1[0,T]}(1+\lambda\|v_\lambda\|_{L^1[0,T]}),$$

which implies that $\|v_\lambda\|_{L^1[0,T]}$ is bounded as a function of λ for $\lambda \in [0, 1/2\|b\|]$. Consequently $\lim_{\lambda\to 0}\|v_\lambda - b\|_{L^1(0,T)} = 0$, and b is nonnegative since v_λ are nonnegative. The last assertion of 1) is an immediate consequence of (1.4) and the definition of complete positivity.

2) Let $\mu>0$. We have to prove

$$r(\lambda r(\mu b))(t) \geqq 0, \qquad t \in [0, T] \quad \text{a.e.}, \quad \lambda>0, \tag{2.4}$$

$$s(\lambda r(\mu b))(t) \geqq 0, \qquad t \in [0, T] \quad \text{a.e.}, \quad \lambda>0. \tag{2.5}$$

Inequality (2.4) is a consequence of the easily verified identity

$$r(\lambda r(\mu b)) = \frac{\lambda}{1+\lambda} r((\mu+\lambda\mu)b) \tag{2.6}$$

and the fact that $r(\lambda b)$ is nonnegative. Inequality (2.5) is a consequence of the identity

$$s(\lambda r(\mu b)) = s((\mu+\lambda\mu)b) + \mu b * s(\mu+\lambda\mu)b) \tag{2.7}$$

and the fact that b and $s(\mu b)$ are nonnegative.

3) and 4). From (1.4) and the fact that b is completely positive on $[0, \infty)$, it follows that $r(\mu b) \in L^1(0, \infty)$ for every $\mu>0$. Hence again for (1.4) $\lim_{t\to\infty} s(\mu b)(t)$ exists. The first assertion of 3) now follows from the definition of s and the fact that $b \in L^1(0, \infty)$ (see for instance [8]). From (1.4) again we have

$$\|r(\mu b)\|_{L^1(0,\infty)} = \int_0^\infty r(\mu b)(\tau)\,d\tau = 1 - s(\mu b)(\infty), \tag{2.8}$$

which proves the last assertion of 3). In order to establish 4) we note that

$$s(\mu b)(t)\int_0^t b(\tau)\,d\tau \leqq \int_0^t b(t-\tau)s(\mu b)(\tau)\,d\tau$$

holds since $s(\mu b)$ is nonnegative and nonincreasing and b is nonnegative. Thus from the definition of s we obtain

$$s(\mu b)(t)\Big[1+\mu\int_0^t b(\tau)\,d\tau\Big]\leqq 1, \qquad t\geqq 0. \tag{2.9}$$

If $b\notin L^1(0,\infty)$, it follows that $\lim_{t\to\infty} s(\mu b)(t)=0$, and from (1.4) $\|r(\mu b)\|_{L^1(0,\infty)}=1$.

5) Since $s(\mu b)\in AC_{\mathrm{loc}}[0,\infty)$ we can differentiate equation (s). By noting that $s'=-r$ (see (1.4)) we obtain

$$r(\mu b)(t)=-\frac{d}{dt}s(\mu b)(t)=\frac{d}{dt}(\mu b * s(\mu b))(t)=\mu b(0)s(\mu b)(t)+\mu\Big(\frac{db}{dt}*s(\mu b)\Big)(t).$$

From part 4) we know that $s(\mu b)\in C[0,\infty]$ and $\lim_{t\to\infty} s(\mu b)(t)=0$. Since $b\in AC[0,\infty)$, $db/dt\in L^1(0,\infty)$, and thus from the above identity we obtain $r(\mu b)\in C[0,\infty]$ and $\lim_{t\to\infty} r(\mu b)(t)=0$. This completes the proof of Proposition 2.1.

In the next result we give an alternate and useful characterization of completely positive kernels b. Some arguments used are similar to those of [6].

THEOREM 2.2. *Let $T>0$, $b\in L^1(0,T)$, $b\not\equiv 0$. Then b is completely positive on $[0,T]$ if and only if there exist $\alpha\geqq 0$ and $k\in L^1(0,T)$ nonnegative and nonincreasing satisfying*:

$$\alpha b(t)+k*b(t)=1, \qquad t\in[0,T]. \tag{2.10}$$

Remarks.

(i) It follows from (2.10) that $\alpha>0$ if and only if $b\in L^\infty(0,T)$. If this is the case, $b=\alpha^{-1}s(\alpha^{-1}k)$ and thus $b\in AC[0,T]$. Conversely, if $b\in AC[0,T]$ then $\alpha=b(0)^{-1}>0$. Moreover, observe that if $\alpha>0$ then $k\in BV[0,T]$ (equivalently, $k(0^+)<\infty$) if and only if $b'\in BV[0,T]$.

The importance of the remark $\alpha>0$, ($k\in BV[0,T]$) is that for kernels b satisfying the type of regularity

$$b\in AC[0,T], \quad b(0)>0, \quad b'\in BV[0,T], \tag{H}$$

the existence and uniqueness of a generalized solution $u\in C([0,T];\overline{D(A)})$ of the abstract Volterra equation (V) has been established by Crandall and Nohel [5, Thm. 4], whenever the operator A is m-accretive and $f(0)\in\overline{D(A)}$, $f\in W^{1,1}(0,T;X)$. For the special case $X=H$ a real Hilbert space and $A=\partial\varphi$ we refer to [5, Remarks in § 4]. Recently Gripenberg [6, Thm. 2] has extended this result to the case of kernels $b=b_1+b_2$, where b_1 satisfies the above regularity assumption and where $b_2\in L^1(0,T)$, b_2 is positive, nonincreasing and $\log b_2$ is convex on $(0,T)$ with A and f as above. This result with $b_1\equiv 0$ and A linear was established by Clément and Nohel [3]. These more general completely positive kernels b correspond to the case $\alpha=0$. The problem of existence of generalized solutions of (V) with only the assumption that b is completely positive is under study and will be treated elsewhere.

For clarity of exposition we recall some basic facts about strong and generalized solutions of the abstract Volterra equation

$$u+b*Au\ni f, \qquad t\in[0,T], \tag{V}$$

where $T>0$ is arbitrary, $b\in L^1(0,T)$, $f\in L^1(0,T;X)$ and A is m-accretive on X (for details see Crandall and Nohel [5], Clément and Nohel [3]). The Yosida approximation A_λ of A is defined by

$$A_\lambda=\frac{1}{\lambda}(I-J_\lambda), \qquad J_\lambda=(I+\lambda A)^{-1}, \qquad \lambda>0.$$

Evidently, A_λ is m-accretive, single-valued and Lipschitz continuous for every $\lambda > 0$. The approximating equation

$$(V_\lambda) \qquad u + b * A_\lambda u = f$$

has a unique strong solution u_λ on $[0, T]$ (use a standard contraction argument). A function $u \in L^1(0, T; X)$ is called a *generalized solution* of (V) on $[0, T]$ if the "sequence" $\{u_\lambda\}$ of strong solutions of (V_λ) converges to $u \in L^1(0, T; X)$ as $\lambda \downarrow 0$. It is important to note that in the existence results of Crandall and Nohel [5, Thm. 4] and of Gripenberg [6, Thm. 2] referred to above, the generalized solution $u \in C([0, T]; X)$ by construction.

(ii) It follows from Theorem 2.2 and Remark (i) that if b is completely positive, then b need not be nonincreasing; it also need not be convex and a fortiori log convex. Choose $\alpha = 1$ and $k(t) = 1$ for $t \in [0, 1]$ and $k(t) = 0$ for $t > 1$; then $b = s(k)$ is completely positive. But as shown in Levin [8], $b' = -r(k)$ is negative on some interval $[0, \alpha)$ with $\alpha \in (1, 2)$ and positive in $(\alpha, 2]$. Thus b is not nonincreasing on $[0, 2]$. Moreover, assume b to be convex on $[0, \infty)$. Then b is strictly increasing for $t > \alpha$, and moreover, $\lim_{t\to\infty} b(t) = \infty$. But this is impossible, since $b(t) \leqq 1$ as seen from (2.10) and the fact that k, b are nonnegative and $\alpha = 1$. Thus b is not convex.

(iii) If b is completely positive and absolutely continuous on $[0, T]$, then it follows from (2.10) that $b(t) \leqq b(0)$ for $t \in [0, T]$.

(iv) It follows from Theorem 2.2 that if $b \in L^1_{\text{loc}}(0, \infty)$, b is positive, decreasing, $\log b$ is convex and $b(0^+) = \infty$, then the linear Volterra equation of the first kind

$$(2.11) \qquad k * b(t) = 1, \qquad t > 0$$

possesses a unique solution $k \in L^1_{\text{loc}}(0, \infty)$ which is nonnegative and nonincreasing. However, given $k \in L^1[0, T]$, k nonnegative and nonincreasing, (2.11) may not have a solution in $L^1(0, T)$. (Take $k(t) \equiv 1$.) Thus when $\alpha = 0$, (2.10) does not provide a way to generate completely positive kernels which are not absolutely continuous on $[0, T]$.

Before giving the proof of Theorem 2.2 we recall a result due to Levin [8] which will be used repeatedly. If u satisfies $u + b * u = f$ with $b \in L^1(0, T)$, b nonnegative and nonincreasing, $f \in L^1(0, T)$, nonnegative, nondecreasing, then u is nonnegative on $[0, T]$.

Proof of Theorem 2.2. a) Let b be completely positive on $[0, T]$, $b \not\equiv 0$. We wish to show that there exist $\alpha \geqq 0$, $k \in L^1(0, T)$ nonnegative and nonincreasing such that (2.10) holds. First, we need the following estimate:

$$(2.12) \qquad \sup_{\lambda>0} \int_0^T \lambda s(\lambda b)(\tau)\, d\tau \leqq \frac{2T}{\int_0^T b(\tau)\, d\tau}.$$

Indeed, define $\tilde{s}(\lambda b)(t) = s(\lambda b)(t)$ for $t \in [0, T]$ and $\tilde{s}(\lambda b)(t) = 0$ for $t > T$ and $\lambda > 0$. Similarly we define $\tilde{b}(t) = b(t)$ for $t \in [0, T]$ and 0 otherwise. If $z_\lambda(t) = \tilde{s}(\lambda b)(t) + (\lambda \tilde{b} * \tilde{s}(\lambda b))(t)$ for $t \geqq 0$ and $\lambda = 0$, then it is easily verified that $z_\lambda(t) = 1$ for $t \in [0, T]$ and $z_\lambda(t) = 0$ for $t \geqq 2T$, $\lambda > 0$. For $t \in [T, 2T]$, using the fact that $\tilde{s}(\lambda b)$ is nonnegative and nonincreasing and $\tilde{b}$ is nonnegative, one gets

$$\begin{aligned}(\tilde{s}(\lambda b) * \lambda \tilde{b})(t) &= \int_0^T \tilde{s}(\lambda b)(t-\tau)\lambda \tilde{b}(\tau)\, d\tau \\ &\leqq \int_0^T \tilde{s}(\lambda b)(T-\tau)\lambda b(\tau)\, d\tau \leqq 1.\end{aligned}$$

Thus $z_\lambda(t) \in [0, 1]$ for $t \in [T, 2T]$ and $\lambda > 0$, and hence from the convolution theorem we obtain

$$\int_0^\infty \tilde{s}(\lambda b)(\tau)\, d\tau \left(1 + \int_0^\infty \lambda \tilde{b}(\tau)\, d\tau\right) = \int_0^\infty z_\lambda(\tau)\, d\tau$$

and

$$\left(\lambda \int_0^T s(\lambda b)(\tau)\, d\tau\right)\left(\int_0^T b(\tau)\, d\tau\right) \leqq 2T,$$

which establishes (2.12).

Next, we define

$$v_\lambda(t) = \int_t^T \lambda s(\lambda b)(\tau)\, d\tau, \qquad \lambda > 0, \quad t \in [0, T]. \tag{2.13}$$

Since $\lambda s(\lambda b)$ is nonnegative,

$$\operatorname{Var}[v_\lambda; [0, T]] = \int_0^T \lambda s(\lambda b)(\tau)\, d\tau,$$

and from (2.12) $\sup \operatorname{Var}[v_\lambda; [0, T]] < \infty$. Note that $v_\lambda(T) = 0$. Hence, from Helly's theorem [13], there is a $v : [0, T] \to R$, nonnegative and nonincreasing, satisfying

$$\operatorname{Var}[v; [0, T]] \leqq \frac{2T}{\int_0^T b(\tau)\, d\tau},$$

and there exists a sequence $\lambda_n \uparrow \infty$ as $n \to \infty$ such that

$$\lim_{n\to\infty} \int_0^T g(t)\, dv_{\lambda_n}(t) = \int_0^T g(t)\, dv(t) \tag{2.14}$$

holds for every $g \in C[0, T]$. From (2.9) we get the estimate

$$\lambda s(\lambda b)(t) \leqq \frac{1}{\int_0^t b(\tau)\, d\tau}, \qquad t > 0, \quad \lambda > 0. \tag{2.15}$$

This implies that, for every $\varepsilon > 0$ ($\varepsilon < T$), $v \in \operatorname{lip}[\varepsilon, T]$ and a fortiori $v \in AC[\varepsilon, T]$. Thus there exist $\beta \leqq 0$ and $h \in AC[0, T]$, nonnegative and nonincreasing, such that $v(t) = \beta e(t) + h(t)$ for $t \in [0, T]$ holds, with $e(0) = 0$ and $e(t) = 1$ for $t > 0$. From (2.14) we obtain

$$\lim_{n\to\infty} \int_0^T g(t)\, dv_{\lambda_n}(t) = \beta g(0) + \int_0^T g(t) h'(t)\, dt \tag{2.16}$$

for every $g \in C[0, T]$ or, from the definition of v_λ,

$$\lim_{n\to\infty} \int_0^T g(t) \lambda_n s(\lambda_n b)(t)\, dt = -\beta g(0) - \int_0^T g(t) h'(t)\, dt. \tag{2.17}$$

If we take $g \in C[0, T]$ such that $g(0) = 0$ in (2.17), then

$$\lim_{n\to\infty} \int_0^T g(t) \lambda_n s(\lambda_n b)(t)\, dt = -\int_0^T g(t) h'(t)\, dt. \tag{2.18}$$

Since $\lambda_n s(\lambda_n b)$ are nonincreasing, it follows that $-h'$ is also nonincreasing. Moreover,

for every $w \in C[0, T]$, it follows from (2.12) that

$$\lim_{n\to\infty} \int_0^T s(\lambda_n b)(t-\tau)w(\tau)\,d\tau = 0 \tag{2.19}$$

uniformly on $[0, T]$. Thus for every $w \in C[0, T]$ we have (from the definition of $s(\lambda b)$)

$$\int_0^T w(\tau)\,d\tau = (s(\lambda_n b) * w)(t) + (\lambda_n s(\lambda_n b) * b * w)(t). \tag{2.20}$$

Using (2.19), (2.17), (2.20) and letting $n \to \infty$, we obtain

$$\int_0^T w(\tau)\,d\tau = -\beta(b * w)(t) - (b * w * h')(t), \qquad t \in [0, T]. \tag{2.21}$$

Finally, since $w \in C[0, T]$ is arbitrary we have

$$-\beta b - h' * b = 1, \qquad t \in [0, T], \tag{2.22}$$

and thus (2.10) holds with $\alpha = -\beta$, $k = -h'$, $\alpha \geqq 0$, $k \in L^1(0, T)$, nonnegative, nonincreasing.

b) Let $\alpha \geqq 0$, $k \in L^1(0, T)$ nonnegative and nonincreasing, and $b \in L^1(0, T)$ such that (2.10) holds. We have to prove that $r(\lambda b) \geqq 0$ and $s(\lambda b) \geqq 0$ for every $\lambda > 0$ and $t \in [0, T]$.

First, we assume $\alpha > 0$. Then (Remark (i)) $b \in AC[0, T]$; from the definition of $s(\lambda b)$, $\lambda > 0$ and (1,4) we obtain $\lambda b(0)s(\lambda b) + \lambda db/dt * s(\lambda b) = r(\lambda b)$. Dividing by $\lambda b(0)$ and solving for $s(\lambda b)$ (using (1.3) with $f = r(\lambda b)/\lambda b(0)$) we obtain

$$s(\lambda b) = \lambda^{-1} b(0)^{-1} r(\lambda b) - r\left(b(0)^{-1}\frac{db}{dt}\right) * \lambda^{-1} b(0)^{-1} r(\lambda b). \tag{2.23}$$

Differentiating (2.10) with $\alpha = b^{-1}(0) > 0$ we deduce that $-r(b(0)^{-1}\,db/dt) = b(0)k$. Thus from (2.23) we get

$$s(\lambda b) = \lambda^{-1} b(0)^{-1} r(\lambda b) + k * \lambda^{-1} r(\lambda b). \tag{2.24}$$

Using (1.4) we get

$$r(\lambda b) + b(0)(k + \lambda) * r(\lambda b) = \lambda b(0). \tag{2.25}$$

Since $k + \lambda$ is positive, nonincreasing and $\lambda b(0)$ is positive, nondecreasing, we conclude by Levin's result mentioned above that $r(\lambda b) \geqq 0$ on $[0, T]$. From (2.24) and the fact that k is nonnegative, we have $s(\lambda b) \geqq 0$ on $[0, T]$, which establishes the complete positivity of b in the case $\alpha > 0$.

Next we assume $\alpha = 0$. For $\varepsilon > 0$, define b_ε by

$$\varepsilon b_\varepsilon + k * b_\varepsilon = 1 \quad \text{on } [0, T]. \tag{2.26}$$

Thus by the proof for the case $\alpha > 0$ one has $r(\lambda b_\varepsilon) \geqq 0$ and $s(\lambda b_\varepsilon) \geqq 0$ for every $\lambda > 0$ on $[0, T]$. From (2.25) we have

$$\varepsilon \lambda^{-1} r(\lambda b_\varepsilon) + (k + \lambda) * \lambda^{-1} r(\lambda b_\varepsilon) = 1. \tag{2.27}$$

From the definition of $r(\lambda b)$ we have

$$\lambda^{-1} r(\lambda b) + \lambda b * \lambda^{-1} r(\lambda b) = b.$$

Since $k * b = 1$ holds for $t \in [0, T]$, we obtain

$$(k + \lambda) * \lambda^{-1} r(\lambda b) = 1, \qquad t \in [0, T]. \tag{2.28}$$

Next, for every $z \in AC[0, T]$ such that $z(0) = 0$ and $z \geqq 0$ we define $u = \lambda^{-1} r(\lambda b) * z$ and $u_\varepsilon = \lambda^{-1} r(\lambda b_\varepsilon) * z$. We know that $u_\varepsilon \geqq 0$ and we want to prove that $u \geqq 0$, which implies $r(\lambda b) \geqq 0$. From (2.28) we obtain the equivalent identity

$$\varepsilon u + (k + \lambda) * u = 1 * z + \varepsilon * \frac{du}{dt}, \qquad t \in [0, T], \tag{2.29}$$

and from (2.27) we obtain

$$\varepsilon u_\varepsilon + (k + \lambda) * u_\varepsilon = 1 * z, \qquad t \in [0, T]. \tag{2.30}$$

Hence $\varepsilon(u - u_\varepsilon) + (k + \lambda) * (u - u_\varepsilon) = \varepsilon * du/dt$. Let w_ε satisfy

$$\varepsilon w_\varepsilon + (k + \lambda) * w_\varepsilon = \varepsilon * \left|\frac{du}{dt}\right|. \tag{2.31}$$

Then $w_\varepsilon - (u - u_\varepsilon)$ is the unique solution of

$$\varepsilon v + (k + \lambda) * v = \varepsilon * \left(\left|\frac{du}{dt}\right| - \frac{du}{dt}\right). \tag{2.32}$$

From Levin's result mentioned above we have $w_\varepsilon - (u - u_\varepsilon) \geqq 0$. Similarly, one shows that $w_\varepsilon + (u - u_\varepsilon) \geqq 0$ by considering the equation

$$\varepsilon v + (k + \lambda) * v = \varepsilon * \left(\left|\frac{du}{dt}\right| + \frac{du}{dt}\right). \tag{2.33}$$

Thus $|u - u_\varepsilon| \leqq w_\varepsilon$ holds, and from (2.31), the fact that $k + \lambda$ is positive, nonincreasing, and w_ε is nonnegative we obtain

$$(k(T) + \lambda) \int_0^T w_\varepsilon(t)\, dt \leqq (k + \lambda) * w_\varepsilon \leqq \varepsilon \int_0^T \left|\frac{du}{dt}\right| dt. \tag{2.34}$$

Therefore, $\lim_{\varepsilon \downarrow 0} \int_0^T w_\varepsilon(t)\, dt = 0$ and $\lim_{\varepsilon \downarrow 0} u_\varepsilon = u$ in $L^1(0, T)$. Consequently, $u = \lambda^{-1} r(\lambda b) * z \geqq 0$ on $[0, T]$ for every $z \in AC[0, T]$, $z \geqq 0$ and $z(0) = 0$. This implies $r(\lambda b) \geqq 0$ on $[0, T]$ for every $\lambda > 0$.

Finally, we observe that

$$s(\lambda b) = k * \lambda^{-1} r(\lambda b), \qquad t \in [0, T]. \tag{2.35}$$

Indeed, $k * \lambda^{-1} r(\lambda b)$ satisfies (see (2.28))

$$k * \lambda^{-1} r(\lambda b) + \lambda b * k * \lambda^{-1} r(\lambda b) = (k + \lambda) * \lambda^{-1} r(\lambda b) = 1, \qquad t \in [0, T].$$

Since k and $r(\lambda b)$ are nonnegative, we obtain $s(\lambda b) \geqq 0$. This concludes the proof of Theorem 2.

3. Qualitative properties of abstract Volterra equations with completely positive kernels. In this section we study some properties of generalized solutions, including positivity and the asymptotic behaviour of positive solutions as $t \to \infty$, of the nonlinear abstract Volterra equation

$$u + b * Au \ni u_0 + b * g, \qquad t \geqq 0. \tag{V_g}$$

Although our results are stated for generalized solutions, it is obvious that the results hold for strong solutions whenever strong solutions are shown to be generalized solutions (see Remark (i)).

The justification for taking $f = u_0 + b * g$ in (V) is as follows. If b satisfies assumption (H) (§ 2, Remark (i) following Theorem 2.2), if $f \in W^{1,1}(0, T; X)$ and if $f(0) \in \overline{D(A)}$, then there exists a unique $u_0 \in \overline{D(A)}$ and a unique $g \in L^1(0, T; X)$ such that

$$f(t) = u_0 + (b * g)(t), \qquad 0 \leqq t \leqq T. \tag{3.1}$$

Indeed, $u_0 = f(0) \in \overline{D(A)}$ and g is the unique solution of the linear equation

$$b(0)g(t) + (b' * g)(t) = f'(t), \qquad 0 \leqq t \leqq T. \tag{3.2}$$

Conversely, if b satisfies assumption (H) and $u_0 \in \overline{D(A)}$, $g \in L^1(0, T; X)$, then f given by (3.1) satisfies $f(0) \in \overline{D(A)}$, $f \in W^{1,1}(0, T; x)$. We shall make the following general assumptions:

$$(\tilde{\mathrm{H}}) \quad \begin{aligned} &A \text{ is } m\text{-accretive in } X; \\ &u_0 \in \overline{D(A)}; \\ &g \in L^1_{\mathrm{loc}}(0, \infty; X); \\ &b \text{ is completely positive on } [0, \infty). \end{aligned}$$

The basic preliminary result assuming the global existence of solutions of (V_g) under assumption $(\tilde{\mathrm{H}})$ is known:

THEOREM 3.1. *If A, u_0, g and b satisfy assumption $(\tilde{\mathrm{H}})$ then:*

1. *If u_1 and u_2 are the generalized solutions of (V_g) corresponding to the data $u_{0,i}$, g_i, $i = 1, 2$, then the following estimate holds:*

$$\|u_1(t) - u_2(t)\| \leqq \|u_{0,1} - u_{0,2}\| + (b * \|g_1 - g_2\|)(t), \qquad t \geqq 0 \quad a.e. \tag{3.3}$$

2. *If P is a closed convex cone in X, if $J_\lambda(P) \subseteq P$ for every $\lambda > 0$, and if $u_0 \in P$ and $g(t) \in P$ a.e. on $[0, \infty]$, then $u(t) \in P$ a.e. in $[0, \infty)$; moreover, if $v - u \in P$ implies $J_\lambda v - J_\lambda u \in P$ for every $\lambda > 0$, $u, v \in X$, and if $u_{0,2} - u_{0,1} \in P$, $g_2(t) - g_1(t) \in P$ a.e. on $[0, \infty)$, then u_i, $i = 1, 2$, the corresponding generalized solutions of (V_g), satisfy $u_2(t) - u_1(t) \in P$ a.e. on $[0, \infty)$.*

Remarks.

(i) The existence of a generalized solution in the linear case under the assumption b completely positive was proved in [3]. In the nonlinear case, when $b \in AC[0, T]$, $b(0) > 0$ and $b \in BV[0, T]$, or when $b \in L^1(0, T)$, b is positive, non-increasing and $\log b$ is convex on $(0, T)$, the existence of generalized solutions of (V_g) follows from results Crandall and Nohel [5] and Gripenberg [6], already discussed in Remark (i) following Theorem 2.2. Moreover, if more regularity is assumed on b and f, then (see [5], [6]) the generalized solution is also a strong solution of (V_g).

(ii) Estimate (3.3) was proved in [2].

(iii) The positivity result in part 2 was proved in [3]. The last assertion of part 2 can be established in a similar way.

We next obtain some results concerning the asymptotic behavior of solutions of (V_g) as $t \to \infty$. We first consider the case $b \in L^1(0, \infty)$.

THEOREM 3.2. *Let A, u_0, g, b, satisfy the general assumptions $(\tilde{\mathrm{H}})$ with $b \not\equiv 0$ and $b \in L^1(0, \infty)$.*

1. *Let $g^\infty \in L^\infty(0, \infty; X)$ and assume there exists $g^\infty \in X$ such that $\lim_{t\to\infty} \|g(t) - g^\infty\| = 0$. Let u be the generalized solution of (V_g) and define $u^\infty = J_{\bar{b}}(u_0 + \bar{b}g^\infty)$, where $\bar{b} = \int_0^\infty b(t)\, dt > 0$. Then the following estimate, which implies strong convergence of $u(t)$ to u^∞ as $t \to \infty$, holds:*

$$\|u(t) - u^\infty\| \leqq \frac{\int_t^\infty b(\tau)\, d\tau}{\bar{b}} \|u_0 - u^\infty\| + (b * \|g - g^\infty\|)(t), \qquad 0 \leqq t < \infty. \tag{3.4}$$

2. *In addition, let* $b \in L^\infty(0,\infty)$ *and* $\lim_{t\to\infty} b(t) = 0$. *Let* $g = g_1 + g_2$ *where* g_1 *satisfies the assumptions of* g *in part* 1, *and where* $g_2 \in L^1(0,\infty;X) + L^p(0,\infty;X)$, $p \in (1,\infty)$. *Let* u *be the generalized solution of* (V_g) *and let* $u^\infty = J_{\bar b}(u_0 - \bar b g_1^\infty)$. *Then the following estimate, which implies strong convergence of* $u(t)$ *to* u^∞ *as* $t \to \infty$, *holds:*

$$\|u(t)-u^\infty\| \leqq \frac{\int_t^\infty b(s)\,ds}{\bar b}\|u_0-u^\infty\| + (b * \|g_1-g_1^\infty\|)(t) + (b * \|g_2\|)(t), \qquad 0 \leqq t < \infty, \tag{3.5}$$

where $g_1^\infty = \lim_{t\to\infty} g_1(t)$.

Remark. Part 1 of Theorem 3.2 was proved in Clément [2] and the proof will be omitted. Part 2 is proved below.

Next we consider the case where $b \notin L^1(0,\infty)$, which is needed for the application in § 4. In order to establish the strong convergence of u to u^∞ as $t \to \infty$ we shall require that the nonlinear operator A in (V) satisfy a rather strong coercivity condition.

THEOREM 3.3. *Let* A, u_0, g *and* b *satisfy the general assumptions* $(\tilde H)$ *with* $b \notin L^1(0,\infty)$ *and* A *is coercive in the sense that there exists* $\omega > 0$ *for which* $A - \omega I$ *is accretive in* X.

1. *Let* g *be in* $L^\infty(0,\infty;X)$ *and let* $g^\infty \in X$ *such that* $\lim_{t\to\infty}\|g(t)-g^\infty\| = 0$. *Let* u *be the generalized solution of* (V_g) *and let* u^∞ *be the unique element in* X *satisfying* $Au^\infty \ni g^\infty$. *Then the following estimate, which implies strong convergence of* $u(t)$ *to* u^∞ *as* $t \to \infty$, *holds*:

$$\|u(t)-u^\infty\| \leqq \int_t^\infty r(\omega b)(\tau)\,d\tau\|u_0-u^\infty\| + \omega^{-1}(r(\omega b) * \|g-g^\infty\|)(t), \qquad 0 \leqq t \leqq \infty. \tag{3.6}$$

2. *In addition, let* b *be* $AC[0,\infty]$ *and* $g = g_1 + g_2$, *where* g_1, g_2 *satisfy the assumptions of Theorem* 3.2, *part* 2, *with* $g_1^\infty = \lim_{t\to\infty} g_1(t)$. *Let* u *be the generalized solution of* (V_g), *and let* u^∞ *be the unique element in* X *satisfying* $Au^\infty \ni g_1^\infty$. *Then the following estimate, which implies strong convergence of* $u(t)$ *to* u^∞ *as* $t \to \infty$, *holds*:

$$\|u(t)-u^\infty\| \leqq \int_t^\infty r(\omega b)(\tau)\,d\tau\|u_0-u^\infty\| + \omega^{-1}(r(\omega b) * \|g_1-g_1^\infty\|)(t) + \omega^{-1}(r(\omega b) * \|g_2\|)(t). \tag{3.7}$$

Remarks.

(i) Since b is completely positive and $b \notin L^1(0,\infty)$, it follows (see Proposition 2.1) that $r(\omega b) \in L^1(0,\infty)$, and therefore, if the assumptions of part 1 hold, (3.6) implies $\lim_{t\to\infty}\|u(t)-u^\infty\| = 0$.

When b also satisfies $b \in AC[0,\infty]$, it follows (see Proposition 2.1) that $r(\omega b) \in L^1(0,\infty) \cap C[0,\infty)$ and $\lim_{t\to\infty} r(\omega b)(t) = 0$. Therefore (3.7) implies $\lim_{t\to\infty}\|u(t)-u^\infty\| = 0$, if the assumptions of part 2 hold.

(ii) As is clear from the proofs, the assumption $g \in L^\infty(0,\infty;X)$ and there exists g^∞ such that $\lim_{t\to\infty}\|g(t)-g^\infty\| = 0$ in part 1 of Theorem 3.2 can be weakened to $g \in L^1_{\mathrm{loc}}(0,\infty;X)$ and there exists $g^\infty \in X$ such that $\lim_{t\to\infty}(b * \|g-g^\infty\|)(t) = 0$. Similar generalizations can be made in Theorem 3.2, part 2 and in Theorem 3.3.

Proof of Theorem 3.2, *part* 2. As in the proof of Theorem 3.2, part 1 in [2], we first prove the result with A replaced by A_λ, $\lambda > 0$, and then we pass to the limit as $\lambda \downarrow 0$. For $\lambda > 0$, let u_λ be the strong solution of the approximating equation

$$u_\lambda + b * A_\lambda u_\lambda = u_0 + b * g, \qquad t \in [0,\infty).$$

Using the definition of A_λ and applying (2.2) we see that u_λ satisfies the equation

$$u_\lambda = r(\lambda^{-1}b) * J_\lambda u_\lambda + s(\lambda^{-1}b)u_0 + \lambda r(\lambda^{-1}b) * g \tag{3.8}$$

for $t \in [0, \infty)$. Since A_λ is also m-accretive, there is a unique u_λ^∞ satisfying the limiting equation

$$u_\lambda^\infty + \bar{b}A_\lambda u_\lambda^\infty = u_0 + \bar{b}g_1^\infty. \tag{3.9}$$

Using the fact that $b \in L^1(0, \infty)$ and $g = g_1 + g_2$ we can rewrite (3.9) in the equivalent form

$$u_\lambda^\infty + b * A_\lambda u_\lambda^\infty = u_0 + b * g + b * (g_1^\infty - g_1) - b * g_2 - \xi w_\lambda^\infty, \tag{3.10}$$

where

$$\xi(t) = \int_t^\infty b(s)\, ds \quad \text{and} \quad w_\lambda^\infty = A_\lambda u_\lambda^\infty - g_1^\infty.$$

Let $\eta : [0, \infty) \to R$ be the unique solution of the linear equation

$$\eta + \lambda^{-1} b * \eta = \xi;$$

then obviously

$$\eta w_\lambda^\infty + \lambda^{-1} b * \eta w_\lambda^\infty = \xi w_\lambda^\infty, \qquad 0 \leqq t \leqq \infty. \tag{3.11}$$

Using (3.10), (3.11), (2.2) and the definition of A_λ we obtain

$$\begin{aligned} u_\lambda^\infty = {} & r(\lambda^{-1}b) * J_\lambda u_\lambda^\infty + s(\lambda^{-1}b)u_0 + \lambda r(\lambda^{-1}b) * g + \lambda r(\lambda^{-1}b) * (g_1^\infty - g_1) \\ & + \lambda r(\lambda^{-1}b) * g_2 - \eta w_\lambda^\infty. \end{aligned} \tag{3.12}$$

Subtracting (3.12) from (3.8) we obtain

$$\begin{aligned} \|u_\lambda(t) - u_\lambda^\infty\| \leqq {} & (r(\lambda^{-1}b) * \|u_\lambda^\infty - u_\lambda\|)(t) + \lambda (r(\lambda^{-1}b) * \|g_1^\infty - g_1\|)(t) \\ & + \lambda (r(\lambda^{-1}b) * \|g_2\|)(t) + |\eta|(t)\|w_\lambda^\infty\|. \end{aligned} \tag{3.13}$$

It is shown in Clément [2, see argument following (3.18)] that $\eta \geqq 0$. Thus by using the same argument as in [2] one gets (take the convolution of (3.13) with $\lambda^{-1}b$)

$$\|u_\lambda(t) - u_\lambda^\infty\| \leqq \xi(t)\|w_\lambda^\infty\| + (b * \|g_1 - g_1^\infty\|)(t) + (b * \|g_2\|)(t), \qquad 0 \leqq t < \infty. \tag{3.14}$$

The conclusion (3.5) follows from using (3.9) and rewriting

$$\xi(t)\|w_\lambda^\infty\| = \frac{\int_t^\infty b(s)\, ds}{\int_0^\infty b(s)\, ds} \|u_0 - u_\lambda^\infty\|,$$

and then letting $\lambda \downarrow 0$. Note that

$$u^\infty = (I + \bar{b}A)^{-1}(u_0 + \bar{b}g_1^\infty) = \lim_{\lambda \downarrow 0} (I + \bar{b}A_\lambda)^{-1}(u_0 + \bar{b}g_1^\infty) = \lim_{\lambda \downarrow 0} u_\lambda^\infty.$$

Proof of Theorem 3.3. We first establish the results with A replaced by $A_{(\lambda)} = \omega I + B_\lambda$, $\lambda > 0$, where B_λ is the Yosida approximation of B, defined by $B = A - \omega I$. Note that B is m-accretive in X. Let u_λ be the strong solution of the approximation equation to (V_g) written in the form

$$u_\lambda + \omega b * u_\lambda + \omega b * \omega^{-1} B_\lambda u_\lambda = u_0 + \omega b * \omega^{-1} g. \tag{3.15}$$

Since the kernel $b \notin L^1(0, \infty)$, we transform this equation into a form which has the property that its new kernel will be in $L^1(0, \infty)$ and completely positive. Indeed, if we

take the convolution of (3.15) by $r(\omega b)$, subtract the result from (3.15) and use the definition of $r(\omega b)$ we get the approximating equation equivalent to (3.15):

$$u_\lambda + r(\omega b) * \omega^{-1} B_\lambda u_\lambda = u_0 + r(\omega b) * (\omega^{-1} g - u_0). \tag{3.16}$$

From Proposition 2.1, $r(\omega b)$ is completely positive and $r(\omega b) \in L^1(0, \infty)$, with $\int_0^\infty r(\omega b)(\delta)\, d\delta = 1$.

To prove Theorem 3.3, part 1, we wish to apply Theorem 3.2, part 1, to (3.16). If g satisfies the assumptions of Theorem 3.3, part 1, so does $\omega^{-1} g - u_0$. Thus all assumptions of Theorem 3.2, part 1 are satisfied with b replaced by $r(\omega b)$, A replaced by $\omega^{-1} B_\lambda$, g replaced by $\omega^{-1} g - u_0$, and u by u_λ. We obtain (by (3.4))

$$\|u_\lambda(t) - u_\lambda^\infty\| \leqq \int_t^\infty r(\omega b)(\tau)\, d\tau \|u_0 - u_\lambda^\infty\| + \omega^{-1}(r(\omega b) * \|g - g^\infty\|)(t), \qquad 0 \leqq t < \infty, \tag{3.17}$$

where u_λ^∞ is the unique solution of the limiting equation

$$u_\lambda^\infty + \omega^{-1} B_\lambda u_\lambda^\infty = \omega^{-1} g_\infty, \tag{3.18}$$

which exists because B_λ is m-accretive and $\omega > 0$. Note that (3.17) is the estimate (3.6) with u replaced by u_λ.

If g satisfies the assumptions of Theorem 3.3, part 2, and $b \in AC[0, \infty]$, it follows from Proposition 2.1 that $r(\omega b) \in L^1(0, \infty) \cap C[0, \infty]$. Thus with $g = g_1 + g_2$ we can apply Theorem 3.2, part 2, to (3.16), and we obtain (from (3.5)) the estimate

$$\|u_\lambda(t) - u_\lambda^\infty\| \leqq \int_t^\infty r(\omega b)(\tau)\, d\tau \|u_0 - u_\lambda^\infty\| + (\omega^{-1} r(\omega b) * \|g_1 - g_1^\infty\|)(t) + (\omega^{-1} r(\omega b) * \|g_2\|)(t), \qquad 0 \leqq t < \infty, \tag{3.19}$$

where u_λ^∞ is the unique solution of the limiting equation

$$u_\lambda^\infty + \omega^{-1} B_\lambda u_\lambda^\infty = \omega^{-1} g_1^\infty. \tag{3.20}$$

Note that (3.19) is the estimate (3.7) with u replaced by u_λ.

Since B is m-accretive, $\lim_{\lambda \downarrow 0} u_\lambda^\infty = u^\infty$, where in the case of (3.18) u^∞ satisfies the limiting equation

$$u^\infty + \omega^{-1} B u^\infty \ni \omega^{-1} g^\infty, \tag{3.21}$$

or equivalently u^∞ satisfies the limiting equation

$$A u^\infty \ni g^\infty. \tag{3.22}$$

Similarly, in the case of (3.20) we find that u^∞ satisfies the limiting equation

$$A u^\infty \ni g_1^\infty. \tag{3.23}$$

It remains to prove that $\lim_{\lambda \downarrow 0} u_\lambda = u$ in $L^1(0, T; X)$ for every $T > 0$, where u is the generalized solution of (V_g). Having done so we see immediately that the estimates (3.17), (3.19) hold with u_λ replaced by u, thus obtaining (3.6) and (3.7). We know that $\lim_{\lambda \downarrow 0} \int_0^T \|\tilde{u}_\lambda(t) - u(t)\|\, dt = 0$, where $\tilde{u}_\lambda \in L^1_{\mathrm{loc}}[0, \infty; X]$ satisfies

$$\tilde{u}_\lambda + b * A_\lambda \tilde{u}_\lambda = u_0 + b * g. \tag{3.24}$$

Introduce the notation $\delta = \lambda(1 + \lambda\omega)^{-1}$. Since $A = \omega I + B$, one easily checks that

$$A_\lambda = \omega(1 + \omega\lambda)^{-1} I + B_\delta.$$

Thus the solution $\tilde{u}_\lambda$ of (3.24) satisfies the equation

$$\tilde{u}_\lambda + b * B_\delta \tilde{u}_\lambda = u_0 + b * (g - \omega(1+\omega\lambda)^{-1}\tilde{u}_\lambda), \tag{3.25}$$

or, by (3.15),

$$u_\delta + b * B_\delta u_\delta = u_0 + b * (g - \omega u_\delta). \tag{3.26}$$

To compare the solutions of (3.25) and (3.26) we apply the inequality (3.3) of Theorem 3.1, and we obtain (note that $u_{01} = u_{02} = u_0$)

$$\|\tilde{u}_\lambda - u_\delta\| \leqq \omega b * (1+\lambda_\omega)^{-1}\|\tilde{u}_\lambda - u_\delta\| + \lambda\omega^2(1+\lambda\omega)^{-1} b * \|\tilde{u}_\lambda\|.$$

and hence also the estimate

$$\|\tilde{u}_\lambda - u_\delta\| \leqq \omega b * \|\tilde{u}_\lambda - u_\delta\| + \delta\omega^2 b * \|\tilde{u}_\lambda\|. \tag{3.27}$$

It follows from (3.27) that for every $T > 0$

$$\int_0^T \|\tilde{u}_\lambda - u_\delta\|(t)\, dt \leqq \delta\omega^2 \|b\|_{L^1(0,T)} \cdot (1 + \|r(-\omega b)\|_{L^1(0,T)}) \int_0^T \|\tilde{u}_\lambda(t)\|\, dt. \tag{3.28}$$

Since $\tilde{u}_\lambda$ converge to u in $L^1(0, T; X)$, we finally obtain $\lim_{\lambda \downarrow 0} \int_0^T \|u_\delta - u\|(t)\, dt = 0$, which proves that u_λ converge in $L^1[0, T; X]$ to the generalized solution u of (V_g) for every $T > 0$. This completes the proof of Theorem 3.3.

Example 3.4. *A conservation law with memory.* As an illustration of Theorems 3.1 and 3.2 we consider the existence and qualitative properties of positive solutions of the problem

$$u(t, x) + \int_0^t b(t-s)\phi(u(s, x))_x\, ds = u_0(x), \qquad t \geqq 0, \quad x \in R. \tag{c}$$

We assume that $\phi \in C^1(R)$ is a given function. If $b \equiv 1$, problem (c) is equivalent to the nonlinear conservation law in one space dimension,

$$u_t + \phi(u)_x = 0, \qquad u(0, x) = u_0(x), \qquad x \in R.$$

Although no particular physical significance is claimed for (c), it evidently contains the usual conservation law as a special case. The latter has been studied extensively from special points of view. Crandall [4] has shown that if $\phi : R \to R$ is a given smooth strictly increasing function (actually ϕ continuous is sufficient) such that $\phi(0) = 0$, then the operator A defined by $Au = \phi(u)_x$ on the Banach space $X = L^1(R)$, with $D(A) = \{u \in L^1(R) : \phi(u_x) \in L^1(R)\}$ (see [4, Def. 1.1 and Thm. 1.1]), is m-accretive on X, and $\overline{D(A)} = X$. Moreover, one has $J_\lambda(0) = 0$ and $J_\lambda u \leqq J_\lambda v (\lambda > 0)$, whenever $u \leqq v$, $u, v \in L^1(R)$.

In (c) assume that $b \in L^1_{\text{loc}}(0, \infty)$, $b \not\equiv 0$ completely positive on $[0, \infty)$, and $u_0 \in \overline{D(A)}$; to be specific take b nonnegative, nonincreasing and $\log b$ convex on $(0, \infty)$. Then by Gripenberg's result [6] (see Remark (i) following Theorem 2.2) and by Theorem 3.1 problem (c) has a unique generalized solution u; u is nonnegative whenever u_0 is nonnegative, and $u_1 \geqq u_2$ whenever $u_{01} \geqq u_{02}$. If, in addition, $b \in L^1(0, \infty)$, then this generalized solution u converges strongly in $L^1(R)$ as $t \to \infty$ to the element $u_0 \in D(A)$ which is the unique solution of the limit equation

$$u_\infty(x) + \left(\int_0^\infty b(t)\, dt \right) \phi(u_\infty(x))_x = u_0(x), \qquad x \in R;$$

u_∞ exists and is uniquely defined since $\int_0^\infty b(t)\, dt > 0$ and A is m-accretive.

4. Nonlinear heat flow in a material with memory. Consider nonlinear heat flow in a homogeneous bar of unit length of a material with memory. Let $u = u(t, x)$ denote the temperature at time t and position x and let the ends of the rod at $x = 0$ and $x = 1$ be maintained at zero temperature. For simplicity and without loss of generality let the history of u be prescribed as zero when $t < 0$ and when $0 \leqq x \leqq 1$ (if not this introduces an additional known forcing term in (4.3) below). The equation satisfied by u is derived from the assumptions that in such materials the internal energy ε and the heat flux q are functionals of u and of the gradient of u respectively (rather than functions of u and u_x). Specifically, we assume following the theory developed by Coleman, Gurtin, MacCamy and Nunziato for heat flow in materials for fading memory type (see, e.g., MacCamy [9], Nunziato [12]) that ε and q are taken respectively as the functionals

$$\varepsilon(t, x) = b_0 u(t, x) + \int_0^t \beta(t-s) u(s, x)\, ds, \qquad t \geqq 0, \quad 0 < x < 1, \tag{4.1}$$

$$q(t, x) = -c_0 \sigma(u_x(t, x)) + \int_0^t \gamma(t-s) \sigma(u_x(s, x))\, ds, \qquad t \geqq 0, \quad 0 < x < 1; \tag{4.2}$$

here $b_0 > 0, c_0 > 0$ are positive constants and the functions $\beta, \gamma: [0, \infty) \to R$ are given sufficiently smooth functions called the internal energy and heat flux relaxation functions, respectively. The given function $\sigma: R \to R$ satisfies the assumption

$$(\sigma) \qquad \sigma \in C^1(R), \quad \sigma(0) = 0, \quad \sigma'(\xi) \geqq p_0 > 0, \qquad \xi \in R, \quad \text{for some } p_0 > 0.$$

In the physical literature the relaxation functions β, γ are usually taken as finite linear combinations of decaying exponentials with positive coefficients. The theory developed in § 3 will be shown to be applicable under the considerably more general and physically reasonable assumptions of Theorem 4.1 below. We remark that for physical reasons one should require at least that $\beta, \gamma \in L^1(0, \infty)$ and

$$(\beta) \qquad b_0 + \int_0^\infty \beta(t)\, dt > 0 \quad \text{and} \quad (\gamma) \quad c_0 - \int_0^\infty \gamma(t)\, dt > 0.$$

If $h = h(t, x)$ represents the external heat supply applied to the rod for $t \geqq 0$ and $0 \leqq x < 1$, and if $u(0, x) = u_0(x), 0 < x < 1$ represents the initial temperature distribution in the rod, we apply the law of balance of heat ($\varepsilon_t = -\operatorname{div} q + h$) to obtain the following initial-boundary value problem to be satisfied by the temperature u:

$$\begin{aligned} &\frac{\partial}{\partial t}[b_0 u(t, x) + (\beta * u)(t, x)] = c_0 \sigma(u_x(t, x))_x - (\gamma * \sigma(u_x)_x)(t, x) + h(t, x), \\ &\qquad\qquad 0 < t < \infty, \quad 0 < x < 1, \\ &u(t, 0) = u(t, 1) \equiv 0, \qquad t > 0, \\ &u(0, x) = u_0(x), \qquad 0 < x < 1. \end{aligned} \tag{4.3}$$

We remark that if the history of u for $t < 0$ is not zero, the integrals in (4.1) and (4.2) range over the interval $(-\infty, t)$ (rather than $(0, t)$), and the resulting equation corresponding to (4.3) would have additional known forcing terms stemming from the integrals over $(-\infty, 0)$ in (4.1) and (4.2). We also remark that with proper interpretation of the differential operator $-\sigma(u_x)_x$ and suitable boundary conditions, the problem (4.3) and the theory for it developed below apply equally well in more than one space dimension.

We next transform the problem (4.3) to a more convenient and recognizable form. Define

$$C(t) = c_0 - \int_0^t \gamma(\tau)\, dt, \qquad 0 \leqq t < \infty, \tag{4.4}$$

$$G(t, x) = b_0 u_0(x) + \int_0^t h(\tau, x)\, dt, \qquad 0 \leqq t < \infty, \quad 0 < x < 1. \tag{4.5}$$

Noting that

$$c_0 \sigma(u_x(t, x))_x - (\gamma * \sigma(u_x)_x)(t, x) = \frac{\partial}{\partial t}(C * \sigma(u_x)_x)(t, x),$$

integrating the equation in (4.3) and using the initial condition, we obtain the Volterra equation

$$u + \beta * u + C * Au = G, \qquad 0 \leqq t < \infty, \quad 0 < x < 1; \tag{V_1}$$

here we have taken the constant b_0 as 1 without loss of generality. The nonlinear operator A is defined to be $Au = -(\partial/\partial x)\sigma(u_x)$, together with the boundary conditions $u(t, 0) = u(t, 1) \equiv 0$. Thus, if X is the Hilbert space $L^2(0, 1)$ and

$$D(A) = \left\{ u \in H_0^1(0, 1): -\frac{\partial}{\partial x}\sigma(u_x) \in L^2(0, 1) \right\},$$

it is well known that if assumptions (σ) are satisfied, then $A = \partial\varphi$, where $\varphi: L^2(0, 1) \to (-\infty, +\infty]$,

$$\varphi(y) = \begin{cases} \displaystyle\int_0^1 W\left(\frac{dy}{dx}\right)(x)\, dx & \text{if } y \in H_0^1\ (0, 1), \\ +\infty & \text{otherwise,} \end{cases} \tag{4.6}$$

where $W(z) = \int_0^z \sigma(\xi)\, d\xi$. Thus φ is convex, l.s.c. and proper on $L^2(0, 1)$ (in fact, $\varphi(y) \geqq 0$), and A is maximal monotone and hence m-accretive on $X = L^2(0, 1)$. Moreover, by an integration by parts and the Pioncaré inequality, A satisfies the coercivity condition

$$(Au, u) \geqq p_0 \pi^2 \int_0^1 |u|^2\, dx. \tag{4.7}$$

The Volterra equation (V_1) may be written in the standard form (V_g) by defining the resolvent kernel $r(\beta)$ of β to be the unique solution of the linear equation

$$r(\beta) + \beta * r(\beta) = \beta, \qquad 0 \leqq t < \infty; \tag{$r(\beta)$}$$

clearly, if $\beta \in L^1(0, \infty)$, then $r(\beta) \in L_{\text{loc}}(0, \infty)$ (at least). Next, define $b: [0, \infty) \to R$ by

$$b = C - r(\beta) * C, \tag{4.8}$$

where C is the function defined in (4.4). Then the variation of constants formula shows that (V_1) is equivalent to the Volterra equation

$$u + b * Au = G - r(\beta) * G, \qquad 0 \leqq t < \infty;$$

taking $b_0 = 1$ in (4.5) one sees that (V_1) is equivalent to the equation

$$u + b * Au = u_0 + 1 * (h - r(\beta) * h - u_0 r(\beta)), \qquad 0 \leqq t < \infty. \tag{4.9}$$

The heat flow problem (4.3) under study is completely described by (4.9).

Our objective is to apply the theory developed in §§ 2 and 3 to discuss the global existence, uniqueness, positivity and decay of solutions of the nonlinear Volterra equation (4.9) (equivalent to (V_1) and to (4.3)), under physically reasonable assumptions on the relaxation functions β, γ, the external heat supply h and the initial temperature distribution u_0. Our main result is:

THEOREM 4.1. *Let β be bounded, nonnegative, nonincreasing and convex on $[0,\infty)$. Let γ be nonnegative, nonincreasing, log convex and bounded on $[0,\infty)$. Let $C(\infty)=c_0-\int_0^\infty \gamma(t)\,dt>0$, and let*

$$\beta'(t)+\frac{\gamma(0)}{c_0}\beta(t)\leqq 0 \quad a.e. \quad for\ t\in[0,\infty). \tag{4.10}$$

Let the assumption (σ) be satisfied, and let $A=\partial\varphi$, where φ is defined in (4.6).

1. *If $u_0\in L^2(0,1)$ and if the forcing function $h\in L^2_{loc}([0,\infty)\times(0,1))$, then the nonlinear Volterra equation* (4.9) (*equivalent to the heat flow problem* (4.3)) *possesses a unique strong solution u on $[0,\infty)$, such that $\sqrt{t}\,u'\in L^2_{loc}(0,\infty;L^2(0,1))$; if $u_0\in H^1_0(0,1)$, then $u'\in L^2_{loc}(0,\infty;L^2(0,1))$.*

2. *If the data u_0 and h satisfy $u_{0,1}(x)\leqq u_{0,2}(x)$ a.e. on $[0,1]$ and $h^1(t,x)\leqq h^2(t,x)$ a.e. on $[0,\infty)\times[0,1]$, then the corresponding strong solutions u_i $(i=1,2)$ satisfy $u_1(t,x)\leqq u_2(t,x)$ a.e. $[0,\infty)\times[0,1]$; in particular, if $u_0(x)\geqq 0$ and $h(t,x)\geqq 0$ a.e. on $[0,1]$ and $[0,\infty)\times[0,1)$ respectively, then $u(t,x)\geqq 0$ a.e. on $[0,\infty)\times[0,1]$.*

3. *If, in addition, $\beta\in L^1(0,\infty)$, and if $h=h_1+h_2$ (where $h_1\in L^\infty(0,\infty;L^2(0,1))$ and there exists $h_1^\infty\in L^2(0,1)$ such that $\lim_{t\to\infty}\|h_1(t)-h_1^\infty\|_{L^2(0,1)}=0$, and where $h_2\in L^p(0,\infty;L^2(0,1))$ for some $p\geqq 1$) then the strong solution u of* (4.9) *converges strongly in $L^2(0,1)$ as $t\to\infty$ to the element $u^\infty\in L^2(0,1)$; u^∞ is the unique solution of the limit equation $Au^\infty=g_1^\infty$, where*

$$g_1^\infty=\frac{h_1^\infty}{c_0}\left(1+\frac{\bar\gamma}{C(\infty)}\right),\qquad \bar\gamma=\int_0^\infty\gamma(t)\,dt. \tag{4.11}$$

In particular, if $h_1^\infty=0$, then $u^\infty=0$.

We pause to comment about the assumptions concerning the relaxation functions β and γ. Since in the physical literature these functions are taken as linear combinations of decaying exponentials with positive coefficients (or even only a single such exponential), it is reasonable to assume that both functions are nonnegative bounded, nonincreasing and convex on $[0,\infty)$; while we only require that $\log\gamma$ is convex, both functions are log convex in this physical case. The assumption $C(\infty)=c_0-\int_0^\infty\gamma(t)\,dt>0$ is motivated as follows. Suppose that the temperature $u(t,x)\to\bar u(x)$ as $t\to\infty$ (where $\bar u(x)$ is the equilibrium temperature), and for definiteness suppose that $d\bar u/dx>0$ at $x\in(0,1)$ (i.e., the equilibrium gradient of temperature is positive). Then assumptions (σ) and $\gamma\in L^1(0,\infty)$ applied to (4.2) yield

$$\lim_{t\to\infty}q(t,x)=-C(\infty)\sigma\left(\frac{d\bar u}{dx}\right),\qquad \sigma\left(\frac{d\bar u}{dx}(x)\right)>0.$$

Thus the assumption $C(\infty)>0$ insures that the equilibrium flux is negative, and this fact is essential in order to guarantee "forward" heat flow at equilibrium. Since γ is nonnegative, nonincreasing and bounded on $[0,\infty)$, the assumption $C(\infty)>0$ also implies that $C(t)=c_0-\int_0^t\gamma(\tau)\,d\tau$ with $c_0>0$ is strictly positive (as well as nonincreasing and bounded) on $0\leqq t<\infty$; this is essential for "forward" heat flow for all $t\geqq 0$. The reader should recall that even in the linear case the "backward" heat equation does not, in general, lead to well posed problems. The assumption (4.10), which, together with

the logarithmic convexity of γ, is used to show that the kernel b in (V) defined by (4.8) is completely positive, is motivated in the remark following the proof of Lemma 4.2 below.

Before giving the proof of Theorem 4 1 we state a lemma which establishes some properties of the kernel b defined by (4.8).

LEMMA 4.2. *Let β, γ, C satisfy the assumptions of Theorem* 4.1. *Then b defined by* (4.8) *is completely positive on* $[0, \infty)$, *b satisfies the assumption* (H) (*Remark* (i) *following Theorem* 2.2), *and α, k associated with b in Theorem* 2.2 *satisfy* $\alpha = c_0^{-1} > 0$ *and* $k \in L^1(0, \infty)$ *with*

$$\int_0^\infty k(\tau)\, d\tau = \frac{1}{c_0}\left[\frac{1}{C(\infty)}\bar{\gamma}(1+\beta)+\bar{\beta}\right], \tag{4.12}$$

where $\bar{\beta} = \int_0^\infty \beta(t)\, dt$, $\bar{\gamma} = \int_0^\infty \gamma(t)\, dt$. *Moreover,* $b \notin L^1(0, \infty)$ *and* $b' \in L^1(0, \infty)$.

Proof. Since the functions β and $C \in AC_{\mathrm{loc}}[0, \infty)$ it follows that the functions $r(\beta)$ and $b \in AC_{\mathrm{loc}}[0, \infty)$ (see definitions ($r(\beta)$) and (4.8) respectively). Note that $b(0) = c_0 > 0$. Define $\alpha = b(0)^{-1}$ and let k be the solution of the linear Volterra equation

$$b(0)y + b' * y = -\frac{b'}{b(0)}, \qquad 0 \leqq t < \infty. \tag{k}$$

Since $b' \in L^1_{\mathrm{loc}}[0, \infty)$, $k \in L^1_{\mathrm{loc}}(0, \infty)$, and since

$$\frac{d}{dt}(b * k)(t) = b(0)k(t) + (b' * k)(t) = -\frac{b'(t)}{b(0)},$$

one has by integration that k satisfies the linear Volterra equation

$$\alpha b(t) + (k * b)(t) = 1, \qquad 0 \leqq t < \infty. \tag{4.13}$$

Since $b(0) > 0$, one also has that k is uniquely defined by (4.13).

In order to show that b is completely positive it suffices, by Theorem 2.2, to show that k is nonnegative, nonincreasing and bounded on $[0, \infty)$. We first observe that the assumptions made on γ imply that C, $-C'$ are convex and $\log(-C')$ is convex on $(0, \infty)$. This in turn implies that $\log C$ is convex on $(0, \infty)$; see G. Gripenberg [7]. Since C is nonnegative, nonincreasing and belongs to $L^1_{\mathrm{loc}}(0, \infty)$, C is completely positive on $[0, \infty)$. Moreover C also satisfies assumption (H) (see Remark (i) following Theorem 2.2). It follows from Theorem 2.2 that there exists $\alpha_c > 0$, and $k_c \in L^1_{\mathrm{loc}}(0, \infty)$, nonnegative, nonincreasing and bounded satisfying

$$\alpha_c C(t) + (k_c * C)(t) = 1, \qquad 0 \leqq t < \infty. \tag{4.14}$$

Note that $\alpha_c = c_0^{-1} = b^{-1}(0) = \alpha$. From the definitions of b in (4.8), of $r(\beta)$, and from (1.2) and (1.3) it follows that

$$C(t) = b(t) + (\beta * b)(t), \qquad 0 \leqq t < \infty. \tag{4.15}$$

Substituting (4.15) into (4.14) yields

$$\alpha b + (k_c + \alpha\beta + k_c * \beta) * b = 1,$$

and thus (4.13) implies that

$$k(t) = k_c(t) + \alpha\beta(t) + (k_c * \beta)(t), \qquad 0 \leqq t \leqq \infty. \tag{4.16}$$

Since $k_c \in BV(0,\infty)$ we have

$$\frac{d}{dt}[\alpha\beta + k_c * \beta](t) = \alpha\beta'(t) + k_c(0)\beta(t) + \int_0^t \beta(t-\tau)\, dk_c(\tau)$$

a.e. on $[0,\infty)$. Hypothesis (4.10) and the identity (4.14) imply that $\alpha\beta'(t) + k_c(0)\beta(t) = (1/c_0)[\beta'(t) + (\gamma(0)/c_0)\beta(t)] \leqq 0$. Moreover, since k_c is nonincreasing and β is nonnegative,

$$\int_0^t \beta(t-\tau)\, dk_c(\tau) \leqq 0, \qquad 0 \leqq t < \infty.$$

Thus k is nonnegative and nonincreasing on $[0,\infty)$. Therefore, one also has $k \in BV[0,\infty)$ if $k \in L^\infty[0,1]$. But k_c and β are bounded and $\beta \in L^1(0,1)$ imply $k \in L^\infty[0,1]$ (note that here the assumption $\beta \in L^1(0,\infty)$ is not needed). From Theorem 2.2 again it follows that b is completely positive and satisfies assumption (H).

We next establish that $b \notin L^1(0,\infty)$. Since $C' = -\gamma \in L^1(0,\infty)$, and $\lim_{t\to\infty} C(t) = C(\infty) = C_0 - \int_0^\infty \gamma(s)\, ds > 0$, it follows that $C \notin L^1(0,\infty)$. If $b \in L^1(0,\infty)$, it would follow from (4.15) and the assumption $\beta \in L^1(0,\infty)$ that $C \in L^1(0,\infty)$, a contradiction. Thus $b \notin L^1(0,\infty)$.

We next prove that $b' \in L^1(0,\infty]$. Indeed, from (4.8) it follows that

$$b'(t) = C'(t) - C(0)r(\beta)(t) - (r(\beta) * C')(t).$$

But $C' = -\gamma \in L^1(0,\infty)$; moreover, $r(\beta) \in L^1(0,\infty)$, since β is nonnegative, nonincreasing, convex and $\beta \in L^1(0,\infty)$ (use the Paley–Wiener theorem and the fact that β is positive definite).

Finally, we show that $k \in L^1(0,\infty)$. From (4.16) and the fact that $\beta \in L^1(0,\infty)$, it is sufficient to prove $k_c \in L^1(0,\infty)$. From (4.14) and the fact that C is positive, nonincreasing, $C(\infty) > 0$ and k is nonnegative, we have

$$C(\infty)\int_0^t k_c(\tau)\, d\tau \leqq C * k_c \leqq 1, \qquad 0 \leqq t < \infty,$$

which proves that $k_c \in L^1(0,\infty)$. Formula (4.12) follows easily from (4.16) and the differentiated form of (4.14). This completes the proof of Lemma 4.2.

Remark. In Lemma 4.2, if $\beta(t) = \sum_{k=1}^n b_k e^{-\beta_k t}$ with $b_k > 0$ and $0 < \beta_1 < \beta_2 \cdots < \beta_n$, then condition (4.10) is satisfied if $\beta_1 \geqq \gamma(0)/c_0$ holds. Indeed, since $\log\beta$ is convex and nonincreasing, it suffices to require

$$\lim_{t\to\infty} \frac{\beta'(t)}{\beta(t)} < -\frac{\gamma(0)}{c_0}.$$

Proof of Theorem 4.1. We begin with the proof of the existence and uniqueness of strong solutions of the Volterra equation (4.9). Defining $f : [0,\infty) \times L^2(0,1) \to L^2(0,1)$ by

$$f = u_0 + 1 * (h - r(\beta) * h - u_0 r(\beta)), \tag{4.17}$$

we have

$$f' = h - r(\beta) * h - u_0 r(\beta), \qquad 0 \leqq t < \infty, \quad 0 < x < 1. \tag{4.18}$$

It follows from Lemma 4.2 that the kernel b satisfies the assumption

$$b(0) > 0, \quad b \in AC_{\mathrm{loc}}[0,\infty), \quad b' \in BV_{\mathrm{loc}}[0,\infty), \tag{H}$$

and that $f \in W^{1,2}_{\mathrm{loc}}(0,\infty; L^2(0,1))$ whenever $h \in L^2_{\mathrm{loc}}((0,\infty)\times[0,1))$. Since under assumptions (σ) $A = \partial\varphi$, where φ is defined by (4.6), the existence and uniqueness of strong solutions u with the properties asserted in part 1 of Theorem 4.1 follows from the result of Crandall and Nohel [6, Thm. 4 and Remarks in § 4].

We next establish the asymptotic results asserted in Theorem 4.1. Since $r(\beta) \in L^1(0,\infty)$ (proved in the demonstration of Lemma 4.2), and since $h = h_1 + h_2$, one has from (4.18) that

$$f' = (h_1 - r(\beta) * h_1) + (h_2 - r(\beta) * h_2) - r(\beta)u_0, \qquad 0 \leqq t < 0, \quad 0 < x < 1,$$

where $h_1 - r(\beta) * h_1 \in L^\infty(0,\infty; L^2(0,1))$ and

$$\lim_{t\to\infty}(h_1 - r(\beta) * h_1)(t) = \left(1 - \int_0^\infty r(\beta)(\tau)\, d\tau\right) h_1^\infty = s(\beta)(\infty) h_1^\infty = (1+\bar{\beta})^{-1} h_1^\infty.$$

Moreover,

$$(h_2 - r(\beta) * h_2 - r(\beta)u_0) \in L^1(0,\infty; L^2(0,1)) + L^p(0,\infty; L^2(0,1)),$$

with $1 < p < \infty$.

By using (4.18), the fact that the kernel b defined by (4.8) is by Lemma 4.2 completely positive and Theorem 2.2, we can write the Volterra equation (4.9) in the equivalent form

$$u + b * Au = u_0 + b * (\alpha f' + k * f'), \qquad 0 \leqq t < \infty. \tag{4.19}$$

To arrive at (4.19) we use the relation $\alpha b + k * b = 1$ in the right-hand side of (4.17) and recombine terms making use of (4.18). Thus (4.19) is in the basic form (V_g) of § 3 with

$$g = \alpha f' + k * f', \qquad 0 \leqq t < \infty. \tag{4.20}$$

From Lemma 4.2, $k \in L^1(0,\infty)$ and $g = g_1 + g_2$, where (with $h = h_1 + h_2$ in (4.18))

$$g_1 = \alpha(h_1 - r(\beta) * h_1) + k * (h_1 - r(\beta) * h_1), \qquad 0 \leqq t < \infty, \tag{4.21}$$

$$g_2 = g_{2,1} + g_{2,2}, \qquad 0 \leqq t < \infty, \tag{4.22}$$

with

$$g_{2,1} = -\alpha u_0 r(\beta) - \alpha r(\beta) * h_2 + k * (h_2 - r(\beta) * h_2 - u_0 r(\beta)), \qquad g_{22} = \alpha h_2.$$

Clearly $g_{21} = L^1(0,\infty; L^2(0,1)$ and $g_{22} \in L^p(0,\infty; L^2(0,1))$. From Lemma 4.2 one has that b is completely positive on $[0,\infty)$, $b \notin L^1(0,\infty)$, and $b' \in L^1(0,\infty)$. Thus all assumptions of Theorem 3.3, part 2, are satisfied. We conclude that estimate (3.7) holds, and therefore $\lim_{t\to\infty}\|u(t) - u^\infty\| = 0$, where $u^\infty = A^{-1}g_1^\infty$ with g_1^∞ given by (4.11); note that to evaluate g_1^∞ use is made of (4.12) and of Proposition 2.1.

Finally, we establish the "comparison" result asserted in Theorem 4.1, part 2. Let $P = \{u \in L^2(0,1) : u \geqq 0\}$; P is a closed convex cone in $L^2(0,1)$ and $v - u \in P$ if and only if $u \leqq v$. Moreover, it is standard that if $u \leqq v$ then $J_\lambda u \leqq J_\lambda v$ for every $\lambda > 0$, where $J_\lambda = (I + \lambda A)^{-1}$. We shall prove the result for solutions of the Volterra equation (V_1) which is equivalent to (4.9). As usual we shall prove the result for solutions of the approximating equation $(V_{1\lambda})$ of (V_1) in which A is replaced by the Yosida approximation A_λ, $\lambda > 0$, and then obtain the result by letting $\lambda \downarrow 0$.

Let $u_{0,i} \in L^2(0, 1)$, $h_i \in L^1(0, 1)$, $i = 1, 2$, satisfy $u_{01} \leqq u_{02}$ and $h_1(t) \leqq h_2(t)$ a.e. on $[0, T]$; let $u_{\lambda,i}$ be the strong solutions of the approximating equation

$$(V_{1\lambda}) \qquad u_{\lambda,i} + \beta * u_{\lambda,i} + \lambda^{-1} C * u_{\lambda,i} = \lambda^{-1} C * J_\lambda u_{\lambda,i} + u_{0,i} + 1 * h_i, \qquad i = 1, 2, \quad \lambda > 0, \quad 0 \leqq t \leqq T.$$

It follows by an elementary calculation (which uses the definitions of b, $r(\lambda^{-1} b)$ and the relation $C \equiv b + \beta * b$) that

$$u_{\lambda,i} = r(\lambda^{-1} b) * J_\lambda u_{\lambda,i} + f_{\lambda,i}, \qquad i = 1, 2, \quad \lambda > 0, \tag{4.23}$$

where $f_{\lambda,i}$ are solutions of the linear Volterra equation

$$f_{\lambda,i} + \beta * f_{\lambda,i} + \lambda^{-1} C * f_{\lambda,i} = u_{0i} + 1 * h_i, \qquad i = 1, 2. \tag{4.24}$$

Hence by a familiar calculation one has

$$f_{\lambda,1} = u_{0,i} s(\beta + \lambda^{-1} C) + h_i * s(\beta + \lambda^{-1} C), \qquad i = 1, 2,$$

and from (4.23), (4.24) the difference $u_{\lambda,2} - u_{\lambda,1}$ satisfies the Volterra equation

$$u_{\lambda,2} - u_{\lambda,1} = r(\lambda^{-1} b) * (J_\lambda u_{\lambda,2} - J_\lambda u_{\lambda,1}) + (u_{0,2} - u_{0,1}) s(\beta + \lambda^{-1} C) + (h_2 - h_1) * s(\beta + \lambda^{-1} C). \tag{4.25}$$

Since $\beta + \lambda^{-1} C$ is positive, nonincreasing, it follows from Levin's result (see § 2) that $s(\beta + \lambda^{-1} C)(t) \geqq 0$. Thus

$$z_\lambda = (u_{0,2} - u_{0,1}) s(\beta + \lambda^{-1} C) + (h_2 - h_1) * s(\beta + \lambda^{-1} C)$$

satisfies $z_\lambda(t) \geqq 0$ a.e. on $[0, \infty)$.

Next, define $v_\lambda = u_{\lambda,2} - u_{\lambda,1} - z_\lambda$; using (4.25) we have that v_λ satisfies

$$v_\lambda = r(\lambda^{-1} b) * (J_\lambda(v_\lambda + u_{\lambda,1} + z_\lambda) - J_\lambda(u_{\lambda,1})). \tag{4.26}$$

As in [3] one shows that $v_\lambda = \lim_{n \to \infty} v_{\lambda,n}$, where

$$v_{\lambda,n+1} = r(\lambda^{-1} b) * (J_\lambda(v_{\lambda,n} + u_{\lambda,1} + z_\lambda) - J_\lambda(u_{\lambda,1})), \tag{4.27}$$

where $v_{\lambda,1} \in L^1(0, T; L^2(0, 1))$ is arbitrary. Choosing $v_{\lambda,1}(t) \in P$ a.e. on $[0, T]$, $T > 0$ arbitrary, one shows easily that $v_{\lambda,n}(t) \in P$ a.e. on $[0, T]$ for all positive integers n. This also uses the fact that $r(\lambda^{-1} b) \geqq 0$, that J_λ is an increasing map with respect to the ordering $\leqq$, and that $z_\lambda(t) \in P$ a.e. on $[0, T]$. Thus $v_\lambda(t) \in P$ a.e. on $[0, T]$, and for $\lambda > 0$

$$u_{\lambda,2}(t) - u_{\lambda,1}(t) = z_\lambda(t) + v_\lambda(t) \geqq 0 \quad \text{a.e.} \quad \text{on } [0, T]. \tag{4.28}$$

Since $T > 0$ is arbitrary, (4.28) holds on $[0, \infty)$, and the conclusion follows from letting $\lambda \downarrow 0$. This completes the proof of Theorem 4.1.

Acknowledgment. The authors thank the referees for suggesting a more descriptive title, requesting clarification of the assumptions regarding the functions β and γ in § 4 (see paragraph following the statement of Theorem 4.1) and pointing out several typographical errors.

REFERENCES

[1] PH. BENILAN, *Equations d'evolution dans une espace de Banach quelconque et applications*, Thèse de Doctorat d'Etat, Université de Paris Sud, 1972.

[2] PH. CLÉMENT, *On abstract Volterra equations with kernels having a positive resolvent*, Israel J. Math., 36 (1980), pp. 193–200.

[3] PH. CLÉMENT AND J. A. NOHEL, *Abstract linear and nonlinear Volterra equations preserving positivity*, this Journal, 10 (1979), pp. 365–388.
[4] M. G. CRANDALL, *The semigroup approach to first order quasilinear equations in several space variables.* Israel J. Math., 12 (1972), pp. 108–132.
[5] M. G. CRANDALL and J. A. NOHEL, *An abstract functional differential equation and a related nonlinear Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.
[6] G. GRIPENBERG, *An abstract nonlinear Volterra equation*, Israel J. Math., to appear.
[7] ———, *On positive, nonincreasing resolvents of Volterra equations*, J. Differential Equations, 30 (1978), pp. 380–390.
[8] J. J. LEVIN, *Resolvents and bounds for linear and nonlinear Volterra equations*, Trans. Amer. Math. Soc., 228 (1977), pp. 207–222.
[9] R. C. MACCAMY, *Stability theorems for a class of functional differential equations*, SIAM. J. Appl. Math., 30 (1976), pp. 557–576.
[9a] ———, *Approximations for a class of functional differential equations*, SIAM J. Appl. Math., 23 (1970), pp. 70–83.
[9b] ———, *A model for one-dimensional, nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21–33.
[10] R. K. MILLER, *On Volterra integral equations with nonnegative integrable resolvents*, J. Math. Anal. Appl., 22 (1968), pp. 319–340.
[11] ———, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
[12] J. W. NUNZIATO, *On heat conduction in materials with memory*, Quart. Appl. Math., 29 (1971), pp. 187–204.
[13] D. V. WIDDER, *The Laplace Transform*, Princeton, University Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0004 $01.00/0

MATHEMATICAL STUDY OF THE NONLINEAR SINGULAR INTEGRAL MAGNETIC FIELD EQUATION. III*

MARK J. FRIEDMAN†

Abstract. We extend the results of Part I [SIAM J. Appl. Math., 39 (1980), pp. 14–20] on the spectrum of the singular integral operator

$$(A\mathbf{M})(x) = -\frac{1}{4\pi}\,\text{grad div}\int_\Omega \frac{\mathbf{M}(y)}{r}\,dy.$$

As an application we obtain an estimate of the lower bound of the spectrum of the magnetic field operator $R\mathbf{M} = h\mathbf{M} + A\mathbf{M}$ from $\mathbf{L}^2(\Omega)$ into the subspace J of generalized solenoidal vector-functions from $\mathbf{L}^2$. Here $\mathbf{M}$ is the magnetization vector, $h\mathbf{M} = (\mathbf{M}/(\mu(M, x)-1))$ $(M = |\mathbf{M}|)$ is the total field, $A\mathbf{M}$ is the induced field, and Ω is a simply connected domain in R_3.

7. Introduction. In this paper we keep the notation and the enumeration of Part I [2] and Part II [3].

In Part I [2] we began the investigation of the spectrum in $\mathbf{L}^2$ of the singular integral operator [2, p. 16]

$$(2.11)\qquad (A\mathbf{M})(x) = \text{grad div}\,\boldsymbol{\psi}(x) \equiv -\frac{1}{4\pi}\,\text{grad div}\int_\Omega \frac{\mathbf{M}(y)}{r}\,dy, \qquad x \in \Omega.$$

We showed there that (i) A is bounded, with $\|A\| = 1$; (ii) A is self-adjoint; and (iii) A is positive semidefinite, with $(A\mathbf{M}, \mathbf{M}) \geqq 0$. The present paper extends the results of Part I [2]. The principal result is given by Theorem 8.1, and follows from classical potential theory, elementary properties of pseudo-differential operators on a compact manifold without edge, and the decomposition (8.19) of $\mathbf{L}^2$ into a direct sum [1].

Theorem 8.1 can have various applications to the investigation of the magnetic field equation ([3] with notational change)

$$(6.1)\qquad R\mathbf{M} = h\mathbf{M} + A\mathbf{M} = \mathbf{H}_a$$

and numerical methods for its solution. As an example we obtain an estimate of the lower bound of the spectrum of $R: \mathbf{L}^2 \to J$, where J is a subspace of generalized solenoidal vector-functions from $\mathbf{L}^2$ (see (8.22) for definition). This choice is natural, since in applications we always have $\mathbf{H}_a \in J$. For simplicity we consider the isotropic case. We denote by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,S}$ the norms in $H^s = H^s(\Omega)$, $H^s(S)$ (see [5] for the definition of these spaces. We shall also use the notation $\|\cdot\|_{0,\Omega} \equiv \|\cdot\|$, $H^0 = \mathbf{L}^2$.

8. The spectrum of A and R. Let us introduce some subspaces of $\mathbf{L}^2$ [1].

$$(8.1)\qquad \mathring{J} = \{\mathbf{M}: \mathbf{M} = \text{rot}\,\mathbf{F}, \mathbf{F} \in \mathbf{H}^1, \text{div}\,\mathbf{F} = 0, \mathbf{F}\times\mathbf{n}|_S = 0\},$$

$$(8.2)\qquad \mathring{G} = \{\mathbf{M}: \mathbf{M} = \text{grad}\,\psi, \psi \in H^1, \psi|_S = 0\},$$

$$(8.3)\qquad U = \{\mathbf{M}: \mathbf{M} = \text{grad}\,\psi, \psi \in H^1, \Delta\psi = 0\}.$$

Here $\mathbf{H}^1 = \mathbf{H}^1(\Omega)$ is the space of vector-functions $\mathbf{F} = (F_1, F_2, F_3)$, $F_i \in H^1$, $i = 1, 2, 3$.

* Received by the editors February 1, 1979, and in revised form October 25, 1979. This work was performed under the auspices of the U.S. Department of Energy under contract W-7405-ENG-48.

† Center for Applied Mathematics, Cornell University, Ithaca, New York 14853.

THEOREM 8.1. *Let the boundary S of Ω be twice continuously differentiable. Then*

1. $\operatorname{Ker} A = \mathring{J}$.
2. $\operatorname{Ker}(A-I) = \mathring{G}$.
3. *Within the interval $(0,1)$ the spectrum of A is at most countable: $\lambda = \frac{1}{2}$ is the unique limit point; each value $\lambda \neq \frac{1}{2}$ is regular or has a finite multiplicity; U is an invariant subspace of A, and the eigenfunctions of A in U form a complete orthogonal system in U.*

We divide the proof into Lemmas 8.1–8.4.

LEMMA 8.1. *For any eigenvalue λ of A, the set of corresponding eigenfunctions of A which are smooth in Ω is dense in the $\mathbf{L}^2$ subspace of corresponding eigenfunctions of A.*

Proof. Let λ, $\mathbf{M}$ satisfy

$$A\mathbf{M} - \lambda \mathbf{M} = 0. \tag{8.4}$$

Setting $\mathbf{M}(x) \equiv 0$ for any $x \in R_3 - \Omega$, we rewrite (8.4) as a convolution in R_3:

$$-\frac{1}{4\pi} \operatorname{grad} \operatorname{div} \left(\mathbf{M} * \frac{1}{|y|}\right)(x) - \lambda \mathbf{M}(x) = 0, \qquad x \in \Omega. \tag{8.5}$$

Now let $\rho(x) \in C^{\infty}(R_3)$ be a function such that $\rho(x)$ has compact support in R_3, $\int_{R_3} \rho(x)\, dx = 1$. Setting $\rho_\varepsilon(x) = \varepsilon^{-3}\rho(x/\varepsilon)$ for $\varepsilon > 0$, we have, from the properties of convolutions,

$$\lambda \rho_\varepsilon * \mathbf{M} = \rho_\varepsilon * \frac{1}{4\pi}\left(\operatorname{grad} \operatorname{div}\left(\mathbf{M} * \frac{1}{|y|}\right)\right) = \frac{1}{4\pi} \operatorname{grad} \operatorname{div}\left((\rho_\varepsilon * \mathbf{M}) * \frac{1}{|y|}\right);$$

i.e., the smooth function $\mathbf{M}^\varepsilon = \rho_\varepsilon * \mathbf{M}$ satisfies (8.4). To end the proof we note that $\mathbf{M}^\varepsilon$ converges to $\mathbf{M}$ in $\mathbf{L}^2$ as $\varepsilon \to 0$.

LEMMA 8.2. $\operatorname{Ker} A = \mathring{J}$.

Proof. Consider A on the linear set $D(A) = \mathbf{C}^1(\Omega) \cap \mathbf{C}(\bar{\Omega})$ of vector-functions with components from $C^1(\Omega) \cap C(\bar{\Omega})$. From the proof of [2, Lemma 2.1] we have the identities

$$\mathbf{M}(x) = -\operatorname{rot} \operatorname{rot} \boldsymbol{\psi}(x) + \operatorname{grad} \operatorname{div} \boldsymbol{\psi}(x), \qquad x \in \Omega, \tag{8.6}$$

$$\operatorname{div} \int_\Omega \frac{\mathbf{M}(y)}{r}\, dy = -\int_S \frac{\mathbf{M}(y) \cdot \mathbf{n}_y}{r}\, dS + \int_\Omega \frac{\operatorname{div} \mathbf{M}(y)}{r}\, dy, \qquad x \in \Omega. \tag{8.7}$$

Using Green's formula, we obtain that

$$\mathbf{M}(x) - \frac{1}{4\pi} \operatorname{rot} \operatorname{rot} \int_\Omega \frac{\mathbf{M}(y)}{r}\, dy = 0, \qquad x \in \Omega.$$

This implies

$$\operatorname{div} \mathbf{M}(x) = 0, \qquad x \in \Omega. \tag{8.8}$$

Using Green's formula, we obtain that

$$(\mathbf{M} \cdot \mathbf{n}, 1)_{0,S} = 0. \tag{8.9}$$

From (8.7) and (8.8) it follows that for $\lambda = 0$, (8.4) can be written as

$$\operatorname{grad} v(x) \equiv \frac{1}{4\pi} \operatorname{grad} \int_S \frac{\sigma(y)}{r}\, dS_y \equiv 0, \qquad x \in \Omega, \tag{8.10}$$

where we use the notation $\sigma = \mathbf{M} \cdot \mathbf{n}$. Taking into account the jump conditions on S for

the derivatives of the single layer potential when x approaches S from the interior,

$$\left(\frac{\partial v}{\partial x_i}\right)\Big|_S = \frac{\sigma(x)}{2}\cos(\mathbf{n}, x_i) + \frac{1}{4\pi}\int_S \sigma(y)\frac{\partial}{\partial x_i}\left(\frac{1}{r}\right) dS_y, \qquad i = 1, 2, 3,$$

we obtain from (8.10)

$$\frac{\sigma}{2} + T\sigma \equiv \frac{\sigma}{2} + \frac{1}{4\pi}\int_S \sigma(y)\frac{\partial}{\partial n_x}\left(\frac{1}{r}\right) dS_y = 0, \qquad x \in S. \tag{8.11}$$

From (8.9) we see that $(\sigma, 1)_{0,S} = 0$. We therefore consider (8.11) in the space $C_0(S) = \{\sigma \in C(S), (\sigma, 1)_{0,S} = 0\}$. It is the homogeneous equation for the interior Neumann problem. The conjugate equation is the equation for the exterior Dirichlet problem. It is known to have only the trivial solution in $C_0(S)$ (see, for example, [6]). By the standard Fredholm theory (8.11) also has only a trivial solution in $C_0(S)$. It follows that

$$\operatorname{Ker} A \cap D(A) \subset \{\mathbf{M} \in D(A): \operatorname{div} \mathbf{M} = 0, \mathbf{M} \cdot \mathbf{n}|_S = 0\}.$$

The inverse inclusion follows immediately from the identity (8.7). By Lemma 8.1 the closure of Ker $A \cap D(A)$ in $\mathbf{L}^2$ is Ker A, and by [1, Theorem 3.2] this closure coincides with $\mathring{J}$; thus the lemma is proved.

LEMMA 8.3. Ker $(A - 1) = \mathring{G}$.

Proof. Consider again A on $D(A) = \mathbf{C}^1(\Omega) \cap \mathbf{C}(\bar{\Omega})$. For $\lambda = 1$, (8.4) is written as

$$\mathbf{M}(x) + \frac{1}{4\pi}\operatorname{grad}\operatorname{div}\int_\Omega \frac{\mathbf{M}(y)}{r}\, dy = 0, \qquad x \in \Omega. \tag{8.12}$$

This implies rot $\mathbf{M} = 0$. Let us set

$$\phi(x) = -\frac{1}{4\pi}\operatorname{div}\int_\Omega \frac{\mathbf{M}(y)}{r}\, dy, \qquad x \in R_3. \tag{8.13}$$

Then $\mathbf{M} = \operatorname{grad}\phi$ in Ω. It is easy to verify that ϕ satisfies the boundary value problem

$$\Delta\phi = \operatorname{div}\mathbf{M}, \qquad x \in \Omega, \tag{8.14}$$

$$\phi^+ = \phi^-, \qquad x \in S, \tag{8.15}$$

$$\Delta\phi = 0, \qquad x \in R_3 - \bar{\Omega}, \tag{8.16}$$

$$\frac{\partial\phi^-}{\partial n} = 0, \qquad x \in S, \tag{8.17}$$

$$\lim \phi(x) = 0 \qquad \text{for } |x| \to \infty, \tag{8.18}$$

where $^+$ and $^-$ denote, respectively, the inner and outer limits on S. Here (8.14), (8.15), (8.16) and (8.18) follow immediately from (8.7) and the properties of the space potential and the single layer one. From (8.12) $\mathbf{M}^+ \cdot \mathbf{n} - \partial\phi^+/\partial n = 0$; together with $\partial\phi^+/\partial n - \partial\phi^-/\partial n = \mathbf{M}^+ \cdot \mathbf{n}$, this gives (8.17). The problem of (8.16)–(8.18) has only the trivial solution, and therefore by (8.15), $\phi^+ = 0$. It follows that

$$\operatorname{Ker}(A - I) \cap D(A) \subseteq \{\mathbf{M} \in D(A): \mathbf{M} = \operatorname{grad}\phi, \phi^+ = 0\}.$$

On the other hand, $\mathbf{M} = \operatorname{grad}\phi$, $\phi^+ = 0$ imply rot $\mathbf{M} = 0$, $\mathbf{M}^+ \times \mathbf{n} = 0$ for $x \in \Omega$. Together with the identities (8.6),

$$\operatorname{rot}\psi = \frac{1}{4\pi}\int_S \frac{\mathbf{n}(y) \times \mathbf{M}(y)}{r}\, dS_y - \int_\Omega \frac{\operatorname{rot}\mathbf{M}(y)}{r}\, dy, \tag{2.7}$$

this gives the inverse inclusion. From Lemma 8.1, the closure of Ker $(A-I)\cap D(A)$ is Ker $(A-I)$; and from the results in [1] the former is $\mathring{G}$. Thus the lemma is proved.

Let us denote by λ_0 the minimum eigenvalue of T in $C_0(S)$ and by Λ_0 the maximum one.

LEMMA 8.4. *On the interval* $(0, 1)$ *the spectrum of* A *in* $\mathbf{L}^2$ *is at most countable*:

1) λ_0 *is the lower bound,* Λ_0 *is the upper bound,* $0<\lambda_0\leqq\Lambda_0<1$; $\lambda=\frac{1}{2}$ *is the unique limit point*; *each value* $\lambda\neq\frac{1}{2}$ *is regular or has finite multiplicity.*

2) U *is an invariant subspace of* A, *and the eigenfunctions of* A *in* U *form a complete orthogonal system in* U.

Proof. By [1],

$$\mathbf{L}^2=\mathring{J}\oplus U\oplus\mathring{G}, \tag{8.19}$$

and therefore from Lemmas 8.2 and 8.3 it follows that for studying the spectrum of A on $(0, 1)$ it is sufficient to consider A on U. We now reduce the problem of the investigation of the spectrum of A on U to that of the investigation of the spectrum of T on $H_0^{-1/2}(S)=\{\sigma\in H^{-1/2};\ \langle\sigma, 1\rangle_{H^{-1/2}(S)\times H^{1/2}(S)}=0\}$.

For $\mathbf{M}\in U$ let us define a potential u by setting grad $u=\mathbf{M}$, $\int_S u\,dS=0$. The norm in $H^1(\Omega)$ can be defined [6] as

$$\|u\|_{1,\Omega}^2=\int_\Omega(\operatorname{grad}u)^2\,dx+\left(\int_S u\,dS\right)^2. \tag{8.20}$$

For u defined above, we therefore have

$$\|u\|_{1,\Omega}^2=\int_\Omega(\operatorname{grad}u)^2\,dx=\int_\Omega(\mathbf{M})^2\,dx=\|\mathbf{M}\|^2. \tag{8.21}$$

Now let λ be a point of the spectrum of A in U. Then there exists a sequence $(\mathbf{M}_n)\subset U$ such that $\mathbf{M}_n=\operatorname{grad}u_n$, $\Delta u_n=0$, $\int_S u_n\,dS=0$, $\|\mathbf{M}_n\|=1$, $\|A\mathbf{M}_n-\lambda\mathbf{M}_n\|\to 0$ as $n\to\infty$. It follows that

$$\left\|-\frac{1}{4\pi}\operatorname{div}\int_\Omega\frac{\mathbf{M}_n(y)}{r}\,dy-\lambda u_n\right\|_{1,\Omega}\to 0$$

as $n\to\infty$. By [5], for a harmonic function u, the mapping $u\to\partial u^+/\partial n$ is continuous from $H^1(\Omega)$ to $H^{-1/2}(S)$. From the proof of Lemma 8.2, for $\sigma_n=\mathbf{M}_n^+\cdot\mathbf{n}\equiv\partial u_n^+/\partial n$ it follows that $\|T\sigma_n-(\frac{1}{2}-\lambda)\sigma_n\|_{-1/2,S}\to 0$ as $n\to\infty$. Thus there is a correspondence between the spectrum of A on $(0, 1)$ in $\mathbf{L}^2(\Omega)$ and a subset of the spectrum of T in $H^{-1/2}(S)$ (see, for example, [8, p. 364]). For each real s T is bounded from $H^s(S)$ into $H^{s+1}(S)$. In particular, T is compact in $H^s(S)$ for all real s. It follows that all eigenfunctions of T are smooth. In particular the spectrum of T in $H_0^{-1/2}(S)$ coincides with that in $C_0(S)$. The spectrum of T in $C(S)$ is a subset of $[-\frac{1}{2}, \frac{1}{2})$ (see, for example, [4]). Now from the proof of Lemma 8.2 it follows that the spectrum of T in $C_0(S)$ belongs to $(-\frac{1}{2}, \frac{1}{2})$. Together with the standard properties of a compact operator, this gives the first statement of the lemma.

To get the second statement of the lemma, we note that for $\mathbf{M}\in U$ from (2.11), (8.7), (8.8) it follows that

$$A\mathbf{M}=\frac{1}{4\pi}\operatorname{grad}\int_S\frac{\mathbf{M}(y)\cdot\mathbf{n}_y}{r}\,dS\in U,$$

and that A is self-adjoint in $\mathbf{L}^2$ [2]. As is well known [7], a compact self-adjoint operator in a Hilbert space gives rise to a complete orthogonal basis of eigenfunctions.

Let us denote by P_1, P_2, P_3 the orthogonal projectors of $\mathbf{L}^2$ onto $\mathring{J}$, U, $\mathring{G}$, respectively. By [1],

$$J \equiv \mathring{J} \oplus U = \{\mathbf{M} : \mathbf{M} = \operatorname{rot} \mathbf{F}, \mathbf{F} \in \mathbf{H}^1\}. \tag{8.22}$$

THEOREM 8.2. *Let* $R : \mathbf{L}^2 \to \mathrm{J}$ *be an operator such that*

$$(R\mathbf{M})(x) = \frac{\mathbf{M}(x)}{\mu(M, x) - 1} + (A\mathbf{M})(x), \qquad 1 < \mu_{\min} \leqq \mu \leqq \mu_{\max} < \infty.$$

Then there holds

$$(R\mathbf{M}, \mathbf{M}) \geqq \left(\frac{1}{\mu_{\max} - 1} + \frac{\mu_{\min} - 1}{\mu_{\max} - 1} \frac{1}{\mu_{\min} - \lambda_0(\mu_{\min} - 1)} \lambda_0\right) \|\mathbf{M}\|^2. \tag{8.23}$$

Proof. Let us define $\lambda = \text{const} \geqq 0$ by

$$\frac{1}{\mu_{\max} - 1} + \lambda = \inf_{\substack{\|\mathbf{M}\| = 1 \\ R\mathbf{M} \in J}} (R\mathbf{M}, \mathbf{M}). \tag{8.24}$$

By Theorem 8.1 we have the following chain of equalities and inequalities:

$$\begin{aligned} \frac{1}{\mu_{\max} - 1} + \lambda_0(\|P_2\mathbf{M}\|^2 + \|P_3\mathbf{M}\|^2) &\leqq \left(\frac{\mathbf{M}}{\mu - 1}, \mathbf{M}\right) + (AP_2\mathbf{M}, P_2\mathbf{M}) + \|P_3\mathbf{M}\|^2 \\ &= (R\mathbf{M}, \mathbf{M}) = \left(\frac{\mathbf{M}}{\mu - 1}, (P_2 + P_3)\mathbf{M}\right) \\ &\quad + (A(P_2 + P_3)\mathbf{M}, (P_2 + P_3)\mathbf{M}) \\ &\leqq \left(\frac{1}{\mu_{\min} - 1} + 1\right) \|(P_2 + P_3)\mathbf{M}\|. \end{aligned} \tag{8.25}$$

Now (8.23) follows from (8.24), (8.25) and the following inequalities:

$$\lambda_0 \inf_{\substack{\|\mathbf{M}\| = 1 \\ R\mathbf{M} \in J}} (\|P_2\mathbf{M}\|^2 + \|P_3\mathbf{M}\|^2) \leqq \lambda,$$

$$\frac{1}{\mu_{\max} - 1} + \lambda \leqq \left(\frac{1}{\mu_{\min} - 1} + 1\right) \inf_{\substack{\|\mathbf{M}\| = 1 \\ R\mathbf{M} \in J}} (\|P_2\mathbf{M}\|^2 + \|P_3\mathbf{M}\|^2).$$

REFERENCES

[1] E. B. BYKHOVSKY AND N. V. SMIRNOV, *On the orthogonal decompositions of the space of vector-functions which are squarely summable over the domain*, TrMIAN-SSSR-59 (1960), pp. 6–36 (In Russian).

[2] M. J. FRIEDMAN, *Mathematical study of the nonlinear singular integral magnetic field equation*, I, SIAM J. Appl. Math., 39 (1980), pp. 14–20.

[3] ———, *Mathematical study of the nonlinear singular integral magnetic field equation*, II, SIAM J. Numer. Anal., to appear.

[4] N. M. GUNTER, *La théorie du potentiel et ses applications aux problèmes fondamentaux de la physique mathématique*, Gauthier-Villars, Paris, 1934.

[5] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. 1, Springer, Berlin, 1972.

[6] S. G. MIKHLIN, *Linear Partial Differential Equations*, Visshaya Shkola, Moscow, 1977 (In Russian).

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

[8] N. N. KOYTORICH, B. Z. KANZELENFAUM AND A. N. SIVOV, *The Generalized Method of Fundamental Oscillations in Diffraction Theory*, Nauka, Moscow, 1977 (In Russian).

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0005 $01.00/0

UNIQUENESS OF A LIMIT CYCLE FOR A PREDATOR-PREY SYSTEM*

KUO-SHUNG CHENG†

Abstract. The uniqueness of a limit cycle for a predator-prey system is proved in this paper. We assume that in the absence of predation the prey regenerates by logistic growth and the predator feeds on the prey with a saturating functional response to prey density. Specifically, we assume that Michaelis–Menten kinetics describe how feeding rates and birth rates change with increasing prey density.

1. Introduction. S. B. Hsu, S. P. Hubbell and Paul Waltman in [2] and [3] considered the following competing-predators system:

$$
\begin{aligned}
&\dot{S}(t) = \gamma S(t)\left(1-\frac{S(t)}{K}\right)-\left(\frac{m_1}{y_1}\right)\left(\frac{X_1(t)S(t)}{a_1+S(t)}\right)-\left(\frac{m_2}{y_2}\right)\left(\frac{X_2(t)S(t)}{a_2+S(t)}\right),\\
&\dot{X}_1(t) = X_1(t)\left(\frac{m_1 S(t)}{a_1+S(t)}-D_1\right),\\
&\dot{X}_2(t) = X_2(t)\left(\frac{m_2 S(t)}{a_2+S(t)}-D_2\right),\\
&S(0)=S_0>0, \qquad X_i(0)=X_{i0}>0, \qquad i=1,2,
\end{aligned}
\tag{1}
$$

where $X_i(t)$ is the population of the ith predator at time t, $S(t)$ is the population of the prey at time t, m_i is the maximum growth (birth) rate of the ith predator, D_i is the death rate of the ith predator, y_i is the yield factor of the ith predator feeding on the prey and a_i is the half-saturation constant of the ith predator, which is the prey density at which the functional response of the predator is half maximal. The parameters γ and K are the intrinsic rate of increase and the carrying capacity for the prey population, respectively. S. B. Hsu et al. analyzed solutions of this system of ordinary differential equations and found out that their behavior depends mainly on the two-dimensional system

$$
\begin{aligned}
&\dot{S}(t) = \gamma S(t)\left(1-\frac{S(t)}{K}\right)-\left(\frac{m}{y}\right)\left(\frac{x(t)S(t)}{a+S(t)}\right),\\
&\dot{x}(t) = x(t)\left(\frac{mS(t)}{a+S(t)}-D_0\right),\\
&S(0)=S_0>0, \qquad x(0)=x_0>0,
\end{aligned}
\tag{2}
$$

where γ, K, m, y, a and D_0 are positive constants.

The results they obtained for system (2) are as follows.

(a) The solutions $S(t)$, $x(t)$ of (2) are positive and bounded.

(b) Let $b = m/D_0$ and $\lambda = a/(b-1)$ if $b>1$.

* Received by the editors June 3, 1980, and in revised form October 28, 1980. This work was partially supported by the National Science Council of the Republic of China.

† Department of Applied Mathematics, National Chiao-Tung University, Hsinchu, Taiwan 300, Republic of China.

(i) If $b \leqq 1$ or $K \leqq \lambda$, then the critical point $(K, 0)$ of (2) is asymptotically stable and

$$\lim_{t\to\infty} S(t) = K, \qquad \lim_{t\to\infty} x(t) = 0.$$

(ii) If $\lambda < K \leqq a + 2\lambda$, then the critical point (λ, x^*), $x^* = (\gamma y/m)(1-\lambda/K)(a+\lambda)$, of (2) is asymptotically stable and

$$\lim_{t\to\infty} S(t) = \lambda, \qquad \lim_{t\to\infty} x(t) = x^*.$$

(iii) If $K > a + 2\lambda$, then (λ, x^*) is unstable and there exists at least one periodic orbit in the first quadrant of the S-x plane. If there is just one periodic orbit, it is stable. If the periodic orbit is not unique, then the outer one is semistable from the outside and the inner one is semistable from the inside.

S. B. Hsu et al. [3] conjectured that the limit cycle is unique and suggested that this can be a delicate question.

It is interesting mathematically to prove the uniqueness of the limit cycle for system (2). If this can be done, then we can better understand the behavior of solutions of (1). Therefore, the purpose of this paper is to show that the limit cycle of (2) is unique under the same conditions as in [2]. From now on we shall assume that $K > a + 2\lambda$.

2. Two lemmas.

LEMMA 1. *Let Γ be a nontrivial closed orbit of system* (2). *Then*

$$\Gamma \subset \{(S, x) | 0 < S < K, 0 < x\}.$$

Let L, R, H and J be the leftmost, rightmost, highest and lowest points of Γ respectively. Then

$$L \in \{(S, x) | 0 < S < \lambda, x = f(S)\},$$

$$R \in \{(S, x) | \lambda < S < K, x = f(S)\},$$

$$H \in \{(S, x) | S = \lambda, x^* < x\},$$

$$J \in \{(S, x) | S = \lambda, 0 < x < x^*\},$$

where $f(S) = \gamma(y/m)(1 - S/K)(a + S)$, the curve of which is symmetric with respect to the vertical line $S = (K - a)/2$, and $x^* = f(\lambda)$.

The proof is simple and we omit it.

LEMMA 2. *Let Γ be a nontrivial closed orbit of* (2). *Γ meets the vertical line $S = (K - a)/2$ at the points A and B with x-coordinates $x_B > x_A$.* (See Fig. 1.) *Let the mirror image of arc $\widehat{BHLJA}$ of Γ with respect to the "mirror" $S = (K - a)/2$ be $\widehat{BH'L'J'A}$. Then arc $\widehat{H'L'J'}$ intersects arc $\widehat{BRA}$ of Γ at two points $P(S_P, x_P)$ and $Q(S_Q, x_Q)$ with $x_Q > f(S_Q)$ and $x_P < f(S_P)$. Furthermore, if $P'(S_{P'}, x_{P'})$ and $Q'(S_{Q'}, x_{Q'})$ are respectively the mirror images of P and Q with respect to the mirror $S = (K - a)/2$, then*

$$0 < \frac{S_{Q'}}{\lambda - S_{Q'}} \leqq \frac{S_Q}{S_Q - \lambda} \tag{3}$$

and

$$0 < \frac{S_{P'}}{\lambda - S_{P'}} \leqq \frac{S_P}{S_P - \lambda}. \tag{4}$$

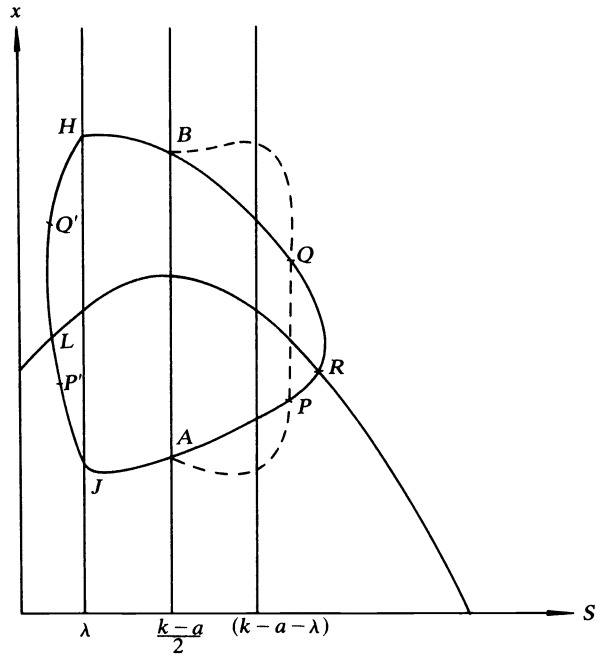


FIG. 1

Proof. Consider the function $V(S, x)$,

$$V(S,x)=\int_{\lambda}^{S}\frac{\left(\frac{m\xi}{a+\xi}-D_0\right)}{\frac{m\xi}{a+\xi}}\,d\xi+\frac{1}{y}\int_{x^*}^{x}\frac{\eta-x^*}{\eta}\,d\eta. \tag{5}$$

Then

$$\begin{aligned}\frac{dV(S(t),x(t))}{dt}&=\frac{1}{y}\left(\frac{mS}{a+S}-D_0\right)\left[\gamma\left(\frac{y}{m}\right)\left(1-\frac{S}{K}\right)(a+S)-x^*\right]\\&=\frac{1}{y}\left(\frac{mS}{a+S}-D_0\right)[f(S)-x^*].\end{aligned} \tag{6}$$

Let the period of Γ be T. We have

$$\int_0^T\frac{dV(S(t),x(t))}{dt}\,dt=0. \tag{7}$$

On the other hand,

$$\begin{aligned}\int_0^T\frac{dV}{dt}\,dt&=\frac{1}{y}\int_0^T\left(\frac{mS(t)}{a+S(t)}-D_0\right)[f(S(t))-x^*]\,dt\\&=\frac{1}{y}\oint_\Gamma[f(S)-x^*]\frac{dx}{x}.\end{aligned} \tag{8}$$

Now assume that arc $\widehat{H'L'J'}$ does not intersect arc $\widehat{BRA}$ of Γ.

It is easy to see that $S_{L'} > S_R$; that is, the region Ω_2 bounded by line $S = (K-a)/2$ and arc $\widehat{BRA}$ is properly contained in the region Ω_1' bounded by line $S = (K-a)/2$ and arc $\widehat{BH'L'J'A}$. From (7) and (8), we have

$$0 = \int_0^T \frac{dV(S(t), x(t))}{dt}\, dt = \oint_\Gamma \frac{1}{y}\frac{[f(S)-x^*]}{x}\, dx$$

$$= \frac{1}{y} \iint_{\Omega_1+\Omega_2} \frac{1}{x} f'(S)\, dS\, dx \qquad \text{(Green's theorem)}$$

$$= \frac{1}{y}\iint_{\Omega_1} \frac{f'(S)}{x}\, dS\, dx + \frac{1}{y}\iint_{\Omega_2} \frac{f'(S)}{x}\, dS\, dx \tag{9}$$

$$= -\frac{1}{y}\iint_{\Omega_1'} \frac{f'(S)}{x}\, dS\, dx + \frac{1}{y}\iint_{\Omega_2} \frac{f'(S)}{x}\, dS\, dx$$

$$= -\frac{1}{y} \iint_{\Omega_1'-\Omega_2} \frac{f'(S)}{x}\, dS\, dx > 0.$$

This is a contradiction. Hence arc $\widehat{H'L'J'}$ does intersect arc $\widehat{BRA}$. Now assume that the points $Q(S_Q, x_Q)$ and $P(S_P, x_P)$ are, respectively, the "highest" and "lowest" intersection point. If $Q(S_Q, x_Q) = P(S_P, x_P)$, then arc $\widehat{H'L'J'}$ intersects arc $\widehat{BRA}$ only at a single point. In this case, the arguments leading to the conclusion that the region Ω_2 is properly contained in Ω_1' still hold. Yet this contradicts (9). Hence $Q(S_Q, x_Q) \neq P(S_P, x_P)$. Assume that $x_Q > f(S_Q)$. Let $(dx/dS)_Q'$ and $(dx/dS)_Q$ be the slopes of arcs $\widehat{BH'L'J'A}$ and $\widehat{BRA}$ at point Q respectively. It is obvious that

$$0 > \left(\frac{dx}{dS}\right)_Q \geqq \left(\frac{dx}{dS}\right)_Q' . \tag{10}$$

But

$$\left(\frac{dx}{dS}\right)_Q = \frac{x_Q\left(\dfrac{mS_Q}{a+S_Q} - D_0\right)}{\gamma S_Q\left(1-\dfrac{S_Q}{K}\right) - \left(\dfrac{m}{y}\right)\left(\dfrac{x_Q S_Q}{a+S_Q}\right)}$$

$$= -(m-D_0)\left(\frac{y}{m}\right)\frac{x_Q(S_Q-\lambda)}{s_Q(x_Q - f(S_Q))},$$

and

$$\left(\frac{dx}{dS}\right)_Q' = -(m-D_0)\left(\frac{y}{m}\right)\frac{x_{Q'}(\lambda - S_{Q'})}{S_{Q'}(x_{Q'} - f(S_{Q'}))}$$

$$= -(m-D_0)\left(\frac{y}{m}\right)\frac{x_Q(\lambda - S_{Q'})}{S_{Q'}(x_Q - f(S_Q))}.$$

(Recall that $x_Q = x_{Q'}, f(S_Q) = f(S_{Q'})$.) Thus from (10) we have

$$0 < \frac{S_{Q'}}{\lambda - S_{Q'}} \leqq \frac{S_Q}{S_Q - \lambda}. \tag{11}$$

Now consider the quadratic function $G(S')$,

$$G(S') = (S-\lambda)(\lambda - S')\left[\frac{S'}{\lambda - S'} - \frac{S}{S-\lambda}\right], \tag{12}$$

where $S = K - a - S'$. A straightforward calculation shows that

$$G(S') = -2S'^2 + 2S'(K-a) - \lambda(K-a).$$

The two positive roots of $G(S') = 0$ are

$$S_{\pm} = \frac{K-a}{2} \pm \sqrt{\left(\frac{K-a}{2}\right)^2 - \lambda\left(\frac{K-a}{2}\right)}.$$

Hence $G(S') < 0$ if $S' < S_-$ or $S' > S_+$, and $G(S') > 0$ if $S_- < S' < S_+$. Since $S_{Q'} < (K-a)/2 < S_+$, from (11) and (12) we conclude that

$$S_{Q'} \leqq S_- \quad \text{or} \quad S_Q \geqq S_+. \tag{13}$$

The arc $\widehat{QR}$ satisfies the following differential equations:

$$\left(\frac{dx}{dS}\right)_{\widehat{QR}} = -(m - D_0)\left(\frac{y}{m}\right)\frac{x(S-\lambda)}{S(x - f(S))}, \qquad x(S_Q) = x_Q, \quad S \geqq S_Q. \tag{14}$$

and $\widehat{QL'}$ satisfies

$$\left(\frac{dx}{dS}\right)_{\widehat{QL'}} = -(m - D_0)\left(\frac{y}{m}\right)\frac{x(\lambda - S')}{S'(x - f(S))}, \qquad x(S_Q) = x_Q, \quad S \geqq S_Q, \tag{15}$$

where $S' = K - a - S$. Since $G(S') < 0$ for $S' < S_-$, we have from (13), (14) and (15) that

$$0 > \left(\frac{dx}{dS}\right)_{\widehat{QR}} > \left(\frac{dx}{dS}\right)_{\widehat{QL'}}. \tag{16}$$

for the same x and $S > S_Q$. Hence from a well-known comparison theorem we get

$$x(S)|_{\widehat{QR}} > x(S)|_{\widehat{QL'}} \quad \text{for } S_Q < S < S_{L'}. \tag{17}$$

From (17) we conclude that arc $\widehat{H'L'}$ can intersect arc $\widehat{BR}$ at most at one point.

Similarly, we obtain that

$$0 < \frac{S_{P'}}{\lambda - S_{P'}} \leqq \frac{S_P}{S_P - \lambda}, \tag{18}$$

$$S_{P'} \leqq S_- \quad \text{or} \quad S_P \geqq S_+, \tag{19}$$

and arc $\widehat{J'L'}$ can intersect arc $\widehat{AR}$ at most at one point. This completes the proof of the lemma. □

3. Uniqueness of the limit cycle. Now we come to our main result.

THEOREM 1. *If $K > a + 2\lambda$, then system (2) possesses a unique limit cycle which is stable.*

Proof. Let Γ be any nontrivial closed orbit of (2), and let its four extreme points be $L(S_L, f(S_L))$, $R(S_R, f(S_R))$, $H(\lambda, x_H)$ and $J(\lambda, x_J)$. Assume that Γ intersects the line $S=(K-a)/2$ at points $A((K-a)/2, x_A)$ and $B((K-a)/2, x_B)$, with $x_B > f((K-a)/2) > x_A$. From Lemma 2, the mirror image of arc $\widehat{BHLJA}$ of Γ, $\widehat{BH'L'J'A}$, intersects the arc $\widehat{BRA}$ of Γ at points $P(S_P, x_P)$ and $Q(S_Q, x_Q)$. Let the mirror images of points P and Q with respect to the line $S=(K-a)/2$ be $P'(S_{P'}, x_{P'})$ and $Q'(S_{Q'}, x_{Q'})$ respectively. It is obvious that $S_{P'} = K-a-S_P$ and $S_{Q'} = K-a-S_Q$.

Now consider

$$g(S,x)=\frac{m}{y}\frac{S(f(S)-x)}{a+S}, \qquad h(S,x)=(m-D_0)\frac{x(S-\lambda)}{a+S}. \tag{20}$$

The divergence of the vector field $(g(S,x), h(S,x))$ defined by (2) is

$$\begin{aligned}\operatorname{Div}(g,h) &= \frac{\partial g}{\partial S}+\frac{\partial h}{\partial x} \\ &= \frac{m}{y}\frac{Sf'(S)}{(a+S)}+\frac{m}{y}\frac{a(f(S)-x)}{(a+S)^2}+(m-D_0)\frac{(S-\lambda)}{a+S}.\end{aligned} \tag{21}$$

From (2) it is easy to see that

$$\oint_\Gamma \left(\frac{m}{y}\right)\frac{a}{(a+S)^2}(f(S)-x)\,dt=\oint_\Gamma \frac{a\,dS}{S(a+S)}=0, \tag{22}$$

$$\oint_\Gamma (m-D_0)\frac{S-\lambda}{a+S}\,dt=0. \tag{23}$$

Hence

$$\oint_\Gamma \operatorname{Div}(g,h)\,dt=\left(\int_{\widehat{AP}}+\int_{\widehat{PRQ}}+\int_{\widehat{QB}}+\int_{\widehat{BQ'}}+\int_{\widehat{Q'LP'}}+\int_{\widehat{P'A}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right]dt. \tag{24}$$

Now

$$\begin{aligned}\left(\int_{\widehat{P'A}}+\int_{\widehat{AP}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right]dt &= \int_{S_{P'}}^{S_A}\frac{f'(S)}{f(S)-x_1(S)}\,dS+\int_{S_A}^{S_P}\frac{f'(S)}{f(S)-x_2(S)}\,dS \\ &= \int_{S_A}^{S_P}\frac{f'(S)}{f(S)-x_2(S)}\,dS \\ &\quad +\int_{S_A}^{S_P}\frac{f'(K-a-S)}{f(K-a-S)-x_1(K-a-S)}\,dS \\ &= \int_{S_A}^{S_P} f'(S)\left[\frac{x_2(S)-x_1(K-a-S)}{(f(S)-x_2(S))(f(S)-x_1(K-a-S))}\right]dS.\end{aligned} \tag{25}$$

The notation in (25) is self-evident. From Lemma 2, we know that $x_2(S) > x_1(K-a-S)$ for $S_A < S < S_P$. Thus we have (recall that $f'(S)<0$ for $S_A<S<S_P$)

$$\left(\int_{\widehat{P'A}}+\int_{\widehat{AP}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right]dt<0. \tag{26}$$

Similarly, we have

$$\left(\int_{\widehat{QB}}+\int_{\widehat{BQ'}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right]dt$$

$$(27)\qquad =\int_{S_B}^{S_Q} f'(S)\left[\frac{x_3(K-a-S)-x_4(S)}{(x_4(S)-f(S))(x_3(K-a-S)-f(S))}\right]dS<0.$$

Let $\bar{\Omega}$ be the region bounded by arc $\widehat{Q'LP'}$ and line segment $\overline{P'Q'}$, and $\bar{\bar{\Omega}}$ be the region bounded by arc $\widehat{PRQ}$ and line segment $\overline{QP}$. We have

$$\int_{\widehat{Q'LP'}}\left(\frac{m}{y}\frac{Sf'(S)}{a+S}\right)dt$$

$$=\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\left[\int_{\widehat{Q'LP'}}\frac{Sf'(S)}{x(S-\lambda)}dx\right]$$

$$(28)\quad =\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\left[\left(\int_{\widehat{Q'LP'}}+\int_{\overline{P'Q'}}+\int_{\overline{Q'P'}}\right)\left(\frac{Sf'(S)}{x(S-\lambda)}\right)dx\right]$$

$$=\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\left[\iint_{\bar{\Omega}}\frac{-1}{x}\cdot\frac{(\gamma/K)[2(S-\lambda)^2+\lambda(K-a-2\lambda)]}{(S-\lambda)^2}dS\,dx\right.$$

$$\left.+\int_{Q'P'}\frac{Sf'(S)}{x(S-\lambda)}dx\right]\ \text{(Green's theorem)}$$

$$<\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\int_{x_{P'}}^{x_{Q'}}\frac{S_1(x)f'(S_1(x))}{x(\lambda-S_1(x))}dx.$$

Similarly, we have

$$\int_{\widehat{PRQ}}\left(\frac{m}{y}\frac{Sf'(S)}{a+S}\right)dt$$

$$(29)\quad =\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\left[\iint_{\bar{\bar{\Omega}}}\frac{-1}{x}\cdot\frac{\gamma/K[2(S-\lambda)^2+\lambda(K-a-2\lambda)]}{(S-\lambda)^2}dS\,dx+\int_{\overline{PQ}}\frac{Sf'(S)}{x(S-\lambda)}dx\right]$$

$$<\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\int_{x_P}^{x_Q}\frac{S_2(x)f'(S_2(x))}{x(S_2(x)-\lambda)}dx,$$

where $S_1(x)$ and $S_2(x)$ represent the line segments $\overline{P'Q'}$ and $\overline{PQ}$ respectively. From (28), (29) and the identity $S_2(x)=K-a-S_1(x)$, we have

$$\left(\int_{\widehat{Q'LP'}}+\int_{\widehat{PRQ}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right]dt$$

$$(30)\qquad <\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\left[\int_{x_{P'}}^{x_{Q'}}\frac{S_1(x)f'(S_1(x))}{x(\lambda-S_1(x))}dx-\int_{x_{P'}}^{x_{Q'}}\frac{(K-a-S_1(x))f'(S_1(x))}{x[K-a-\lambda-S_1(x)]}dx\right]$$

$$=\left(\frac{m}{y}\right)\left(\frac{1}{m-D_0}\right)\int_{x_{P'}}^{x_{Q'}}\frac{f'(S_1(x))}{x}\cdot\frac{G(S_1(x))}{[\lambda-S_1(x)][K-a-\lambda-S_1(x)]}dx,$$

where the polynomial G is defined in (12).

From (13) and (19) we have

$$S_1(x) \leqq \max\{S_{P'}, S_{Q'}\} \leqq S_-, \qquad x_{P'} \leqq x \leqq x_{Q'}$$

and hence

$$G(S_1(x)) \leqq 0 \quad \text{for } x_{P'} \leqq x \leqq x_{Q'}. \tag{31}$$

From (30) and (31) we finally have

$$\left(\int_{\widehat{Q'LP'}} + \int_{\widehat{PRQ}}\right)\left[\frac{m}{y}\frac{Sf'(S)}{a+S}\right] dt < 0. \tag{32}$$

From (26), (27) and (32) we get

$$\oint_\Gamma \operatorname{Div}(g, h)\, dt < 0. \tag{33}$$

This means that the closed orbit Γ is stable. But two adjacent periodic orbits cannot be positively stable on the sides facing each other [1, p. 397, Thm. 3.4]. Hence the uniqueness of the limit cycle of system (2) is proved. □

4. Acknowledgments. The author would like to express appreciation to S. B. Hsu, S. S. Lin and F. S. Tsen for their helpful discussions, and to the referee for useful suggestions. The author would like to note that exactly the same problem had been discussed by Bingxi Li in a preprint [4] which unfortunately contained a serious error. Some ideas of this paper came from Li's preprint.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, New York, 1955.
[2] S. B. HSU, S. P. HUBBELL AND PAUL WALTMAN, *Competing predators*, SIAM J. Appl. Math., 35 (1978), pp. 617–625.
[3] ———, *A contribution to the theory of competing predators*, Ecological Monographs, 48 (1978), pp. 337–349.
[4] B. LI, *Uniqueness of a limit cycle for a competing-predator system*, Abstracts Amer. Math. Soc., 1 (1980), p. 319.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0006 $01.00/0

CONVEX APPROXIMATION BY SPLINES*

R. K. BEATSON†

Abstract. Jackson type estimates are obtained for the approximation of convex functions by convex splines with equally spaced knots. The results are of the same order as the Jackson type estimates for unconstrained approximation by splines with equally spaced knots.

1. Introduction. The aim of this paper is to obtain Jackson type estimates for approximating convex functions by convex splines. Let k and N be positive integers and let $S(k, N)$ denote the space of all splines of order k with simple knots $\{i/N\}_{i=0}^{N}$; i.e., $s \in S(k, N)$ if and only if $s^{(k-2)}$ is continuous on $[0, 1]$ and on each interval $[i/N, (i+1)/N]$, $i = 0, 1, \cdots, N-1$, s is a polynomial of degree $\leqq k-1$. If A is a collection of functions defined on $[0, 1]$, then A^* will denote the subcollection of functions in A which are convex (not necessarily strictly convex) on $[0, 1]$. If f is a convex function on $[0, 1]$ then we define the error in convex approximation by splines to be

$$E_N^*(f, k) = \inf \{\|f - s\|: s \in S^*(k, N)\},$$

where $\|\cdot\|$ is the supremum norm on $[0, 1]$. We will establish the following result.

THEOREM 1. *Let $k \geqq 2$ be a positive integer. There is a constant $C > 0$ depending on k alone such that, if f is convex on $[0, 1]$ and $f \in C^j[0, 1]$ for some $0 \leqq j \leqq k-1$, then*

$$E_N^*(f, k) \leqq CN^{-j}\omega(f^{(j)}, N^{-1}), \qquad N = 1, 2, \cdots. \tag{1.1}$$

The analogous theorem for monotone approximation is due to De Vore [5]. De Vore also proved a similar theorem about monotone approximation by polynomials which had been a known open problem since the 1968 paper of Lorentz and Zeller [8]. More recently Chui, Smith and Ward [3] have given a different and more transparent proof of De Vore's result concerning splines, and also of the analogous L^p result. Some of the techniques used in this paper are taken from their work.

The proof uses "local techniques" and proceeds in two steps. First the function is approximated by a convex piecewise polynomial. Then the convex piecewise polynomial is smoothed into a convex spline. The reader is urged to draw his or her own pictures to bring out the geometric nature of many of the arguments.

2. Convex polynomial interpolation. One possible approach to convex approximation by splines (perhaps with multiple knots) is to look for an interpolating spline which is convex. Since f is convex if f' is increasing, the natural approach is to require the spline to interpolate to f and f' at the knots. The Markov type inequality contained in Lemma 2.1 shows that any such approach must fail. Using it and an elementary argument one can easily build, for any k, a convex function $f \in C^k[0, 1]$ such that for each $N = 2^m$, $m = 0, 1, \cdots$, there is no convex function s reducing to a polynomial of degree $\leqq k-1$ in $[0, 1/N]$ which satisfies $s(i/N) = f(i/N)$ and $s'(i/N) = f'(i/N)$ for $i = 0, 1$.

Lemma 2.1 also has a positive interpretation. This is that a convex polynomial must satisfy certain strong inequalities and therefore must be easy to approximate. This viewpoint is crucial in § 4 and is embodied in a more convenient form in Lemma 2.2.

* Received by the editors May 14, 1980, and in revised form October 9, 1980.
† Department of Mathematics, University of Texas at Austin, Austin, Texas 78712.

LEMMA 2.1. *Let p be a polynomial of degree $\leqq 2n (n \geqq 1)$ convex on $[-1, 1]$ with $p(-1) = p'(-1) = 0$ and $p'(1) = 1$. Then*

$$1 - x_{n,n} \leqq p(1) \leqq 1 - x_{1,n},$$

where $x_{n,n}, x_{1,n}$ are the largest and smallest zeros respectively of the n-th Legendre polynomial. These bounds are achieved.

Proof. From the hypotheses we have

$$1 = p'(1) = \int_{-1}^{1} p''(t)\, dt \tag{2.1}$$

and

$$p(1) = \int_{-1}^{1} (1-t)p''(t)\, dt. \tag{2.2}$$

Now let $-1 < x_{1,n} < \cdots < x_{n,n} < 1$ be the zeros of the nth Legendre polynomial, L_n, in ascending order. Let $A_i(n)$ be the weight associated with $x_{i,n}$ in the Gauss–Legendre quadrature formula with nodes at the zeros of L_n. Then

$$p(1) = \int_{-1}^{1} (1-t)p''(t)\, dt = \sum_{k=1}^{n} A_k(n)(1 - x_{k,n})p''(x_{k,n}),$$

and hence

$$(1 - x_{n,n}) \sum_{k=1}^{n} A_k(n)p''(x_{k,n}) \leqq p(1) \leqq (1 - x_{1,n}) \sum_{k=1}^{n} A_k(n)p''(x_{k,n}). \tag{2.3}$$

But

$$1 = p'(1) = \int_{-1}^{1} p''(t)\, dt = \sum_{k=1}^{n} A_k(n)p''(x_{k,n}).$$

Substituting this into (2.3) we obtain the bounds of the lemma.

It remains to show that these bounds are actually achieved. To see this in the case of the lower bound, consider

$$p''(x) = \lambda \left[\frac{L_n(x)}{(x - x_{n,n})} \right]^2,$$

where λ is a normalizing constant chosen so that $p'(1) = 1$. $p''(x)$ vanishes at all the zeros of L_n except for $x_{n,n}$. Hence, using (2.1) and (2.2), we have

$$\begin{aligned} p(1) = \sum_{k=1}^{n} A_k(n)(1 - x_{k,n})p''(x_{k,n}) &= A_n(n)(1 - x_{n,n})p''(x_{n,n}) \\ &= (1 - x_{n,n}) \sum_{k=1}^{n} A_k(n)p''(x_{k,n}) \\ &= (1 - x_{n,n})\, p'(1) = 1 - x_{n,n}. \end{aligned}$$

Thus the lower bound is achieved. The proof that the upper bound is achieved is similar. □

LEMMA 2.2. *Let $k \geqq 2$ be an integer. There exists a δ_k, $0 < \delta_k < 1$ with the following property. If p is a polynomial of degree $\leqq k$ convex on $[0, 1]$ with $p(0) = p'(0) = 0$, then*

$$(1 - \delta_k) \min_{[1-\delta_k, 1]} p'(x) \geqq \max_{[1-\delta_k, 1]} p(x). \tag{2.4}$$

Proof. It is sufficient to prove the lemma when k is even, say $k=2n$. Now if $p(1)=0$, then, because of the convexity, p is identically zero and the lemma is true. Hence we may assume without loss of generality that $p(1)>0$, and by homogeneity that $p(1)=1$. Now from the previous lemma (after a change of variable) we have

$$1<\frac{2}{1-x_{1,n}}\leqq p'(1)\leqq\frac{2}{1-x_{n,n}}. \tag{2.5}$$

From Markov's inequality $\|p''\|\leqq 4k^4$. Thus for any $x\in[0,1]$

$$(1-x)p'(1-x)\geqq(1-x)\left[\frac{2}{1-x_{1,n}}-x4k^4\right]=q(x).$$

Here $q(0)=2/(1-x_{1,n})>1$ and $q(1)=0$. We choose δ_k as the least positive x so that $q(x)=1$. With this choice of δ_k

$$(1-\delta_k)p'(1-\delta_k)\geqq 1=p(1). \qquad \square$$

3. Convex approximation by piecewise polynomials. In this section we prove Jackson type theorems for approximation of convex functions f by globally convex piecewise polynomials. $\pi(k,N)$ will denote the collection of all continuous functions on $[0,1]$ whose restrictions on each subinterval $[i/N,(i+1)/N]$, $i=0,\cdots,N-1$, are polynomials of degree $\leqq k$. Hence, $\pi^*(k,N)$ is the subcollection of functions in $\pi(k,N)$ which are convex on $[0,1]$. Let

$$D_N^*(f,k)=\inf\{\|f-g\|\colon g\in\pi^*(k,N)\}.$$

We will prove:

THEOREM 3.1. *Let $k\geqq 2$ be an integer. Then*

$$D_N^*(f,k)\leqq 2N^{-j}\omega(f^{(j)},N^{-1})$$

for all $f\in C^j{}^*[0,1]$, $0\leqq j\leqq k$.

The method of proof is to find convex polynomial approximations for each subinterval satisfying certain endpoint conditions. The endpoint conditions are chosen to force the polynomial pieces to join up in a globally convex manner. The most obvious endpoint conditions are interpolation to f and f', but as we remarked in the previous section this approach must fail. Thus we are led to relax the interpolation conditions.

LEMMA 3.2. *Let $j\geqq 2$. For every $f\in C^j{}^*[0,1]$, there is a polynomial $q\in\pi_j$ such that*

$$q \text{ is convex on } [0,1], \tag{3.1}$$

$$q(0)=f(0) \quad \text{and} \quad q(1)=f(1), \tag{3.2}$$

$$f'(0)\leqq q'(0)\leqq q'(1)\leqq f'(1) \tag{3.3}$$

and

$$\|(f-q)^{(i)}\|\leqq 2\omega(f^{(j)},1) \quad \text{for } i=0,1,\cdots,j. \tag{3.4}$$

Interestingly this degree of approximation result is proven by using properties of best restricted-range approximations.

Let u and l be extended real-valued functions defined on $[0,1]$ with $l(x)\leqq u(x)$ throughout $[0,1]$. Suppose $g\in C[0,1]$ and let

$$W=\{p\in\pi_n\colon l(x)\leqq p(x)\leqq u(x),\ \forall x\in[0,1]\}.$$

If W is nonempty we will call $p\in W$ $(=W_n)$ a best restricted approximation to g if

$$\|g-p\|=\inf\{\|g-q\|\colon q\in W\}.$$

W is clearly a closed finite dimensional set in $C[0, 1]$. Hence, if W is nonempty, the existence of a best restricted approximation to g from W follows from a well-known compactness argument. We will show that any best restricted approximation to a smooth g interpolates to g at least $n+1$ times. Professor W. W. Hager has kindly pointed out that the special case of this result, which we use later, appears in his 1974 Ph.D. dissertation and was published in [7]. The interested reader should consult Sippel [9] for alternation and uniqueness theorems, which hold under stronger hypotheses on l, u and g.

LEMMA 3.3. *Let $g \in C^n[0, 1]$ satisfy $l(x) \leqq g(x) \leqq u(x)$ for all $x \in [0, 1]$, and let W be nonempty. Let p be a best approximation to g from W. Then p interpolates to g at least $n+1$ times in a Hermite sense. That is, there exist m points $0 \leqq z_1 < \cdots < z_m \leqq 1$ and m positive integers d_i so that*

$$p^{(j)}(z_i) = g^{(j)}(z_i), \qquad j = 0, \cdots, d_i - 1, \quad i = 1, \cdots, m$$

and

$$\sum_{i=1}^{m} d_i = n + 1.$$

Proof. Let $D = \{z \in [0, 1]: (g-p)(z) = 0\}$. The lemma is certainly true when D has $n+1$ or more distinct members. Hence we may assume D has m, $0 \leqq m \leqq n$, members. If $m = 0$ then $g - p$ has a fixed sign on $[0, 1]$ and we may add a constant to p to obtain an approximation $\hat{p} \in W$ lying strictly between $p(x)$ and $g(x)$ for all x. Then $\hat{p}$ is a better approximation to g than p, a contradiction.

We may now assume $1 \leqq m \leqq n$. Label the points in D $z_1, \cdots, z_m$. For each z_i let d_i be the largest integer with $1 \leqq d_i \leqq n+1$ such that $(g-p)^{(j)}(z_i) = 0$, $j = 0, \cdots, d_i - 1$. If $\sum_{i=1}^{m} d_i \geqq n+1$, then the lemma is true. Assume, on the contrary, that $\sum_{i=1}^{m} d_i \leqq n$. Then, since in particular each $d_i \leqq n$, we have

$$(g-p)^{(d_i)}(z_i) \neq 0, \qquad i = 1, \cdots, m.$$

Define $q(x) = \prod_{i=1}^{m} (x - z_i)^{d_i}$. Then $q \in \pi_n$ has the same zeros as $(g-p)$ on $[0, 1]$ and with exactly the same multiplicities. Hence we may choose $\bar{q} \in \{q, -q\}$ so that $\bar{q}(x)$ has the same sign as $(g-p)(x)$ throughout $[0, 1]$. Then

$$\bar{q}^{(j)}(z_i) = (g-p)^{(j)}(z_i) = 0, \qquad j = 0, \cdots, d_i - 1, \quad i = 1, \cdots, m$$

and

$$\operatorname{sign}(\bar{q}^{(d_i)}(z_i)) = \operatorname{sign}((g-p)^{(d_i)}(z_i)) \neq 0, \qquad i = 1, \cdots, m.$$

A simple compactness argument shows that, for some sufficiently small $\lambda > 0$,

$$0 \leqq \lambda |\bar{q}(x)| \leqq |(g-p)(x)| \quad \forall x \in [0, 1].$$

It follows that, for each $x \in [0, 1]$, $(p + \lambda\bar{q})(x)$ lies between $p(x)$ and $g(x)$, with $(p + \lambda\bar{q})(x) = p(x)$ only when $x \in D$. Hence $(p + \lambda\bar{q}) \in W$ and $\|g - (p + \lambda\bar{q})\| < \|g - p\|$. This contradiction establishes the lemma. □

Proof of Lemma 3.2. Since we can add a linear function to f and to the approximation q without changing anything in the statement, we may assume $f(0) = f'(0) = 0$.

Now let h be a polynomial in π_{j-2} satisfying

$$0 \leqq h(x) \leqq f''(x) \quad \forall x \in [0, 1],$$

which minimizes $\|f'' - h\|$. Then by Lemma 3.3 h interpolates to f'' at least $j-1$ times in a Hermite sense. It follows, by Rolle's theorem, that there exist points $\xi_2, \cdots, \xi_j$ in

$[0, 1]$ such that

$$f^{(i)}(\xi_i) = h^{(i-2)}(\xi_i), \qquad i = 2, \cdots, j.$$

Define $p \in \pi_j$ by

$$p(x) = \int_0^x \int_0^t h(s)\, ds\, dt = \int_0^x (x-t)h(t)\, dt.$$

Then with $\xi_0 = \xi_1 = 0$ there exist $\xi_0, \cdots, \xi_j$ in $[0, 1]$ such that

$$f^{(i)}(\xi_i) = p^{(i)}(\xi_i), \qquad i = 0, 1, \cdots, j.$$

Now, for each $i = 0, 1, \cdots, j-1$,

$$\|f^{(i)} - p^{(i)}\| = \left\| \int_{\xi_i}^x (f^{(i+1)} - p^{(i+1)})(t)\, dt \right\| \leqq \|f^{(i+1)} - p^{(i+1)}\|,$$

so that arguing inductively using the fact that $p^{(j)}(x)$ is a constant we find

$$\|f^{(i)} - p^{(i)}\| \leqq \omega(f^{(j)}, 1), \qquad i = j, j-1, \cdots, 0. \tag{3.5}$$

Now consider the equations

$$f'(1) - p'(1) = \int_0^1 f''(t) - p''(t)\, dt,$$

$$f(1) - p(1) = \int_0^1 (1-t)(f''(t) - p''(t))\, dt.$$

Since $0 \leqq 1 - t \leqq 1$ and $0 \leqq p''(t) \leqq f''(t)$ for $t \in [0, 1]$, we find

$$0 \leqq f(1) - p(1) \leqq f'(1) - p'(1).$$

Hence, since $p(0) = p'(0) = f(0) = f'(0) = 0$, the polynomial

$$q(x) = p(x) + (f(1) - p(1))x$$

has

$$q(0) = f(0) = 0, \qquad q(1) = f(1),$$

and

$$0 = f'(0) \leqq q'(0) \leqq q'(1) \leqq f'(1).$$

The lemma follows since

$$\|(f-q)^{(i)}\| \leqq \|(f-p)^{(i)}\| + |f(1) - q(1)| \leqq 2\omega(f^{(j)}, 1), \qquad i = 0, 1, \cdots j,$$

by (3.5). □

Proof of Theorem 3.1. Since $\pi^*(j, N) \subset \pi^*(k, N)$ it is sufficient to show that

$$D_N^*(f, j) \leqq 2N^{-j}\omega(f^{(j)}, N^{-1}), \qquad j = 2, \cdots, k \tag{3.6}$$

and

$$D_N^*(f, 1) \leqq 2N^{-j}\omega(f^{(j)}, N^{-1}), \qquad j = 0, 1. \tag{3.7}$$

The last inequality is immediate from the known properties of the polygonal approximation. To show (3.6) we construct a piecewise polynomial using Lemma 3.2. Firstly blow $[0, 1]$ up to $[0, N]$ by letting $\hat{f}(N^{-1}x) = f(x)$. Then on each subinterval $[i, i+1]$,

$i = 0, \cdots, N-1$, approximate $\hat{f}$ by the polynomial whose existence is guaranteed by Lemma 3.2. Finally invert the change of variable to return to $[0, 1]$. The convexity of the resulting piecewise polynomial, s, is clear. The degree of approximation is also clear when we note

$$\omega(\hat{f}^{(j)}, [0, N], 1) = N^{-j}\omega(f^{(j)}, [0, 1], N^{-1}).$$

Here we have used an obvious notation to denote the interval over which each modulus of continuity is defined. □

4. Convex approximation by splines. In this section we will prove Theorem 1. First we will smooth our convex piecewise polynomial approximant to C^1. Then this new approximant will be smoothed to C^{k-2}. In what follows $\|h\|_I$ is the essential supremum of $|h(x)|$ over I.

LEMMA 4.1. *Let $k \geqq 3$ and $M \geqq 2$. Let $f \in C^*[0, 2M]$ be a spline of order k with knots at $0, 2, 4, \cdots, 2M$. Then there exists a spline $s \in C^1*[0, 2M]$ of order k with knots at $0, 1, 2, 3, \cdots, 2M$, such that*

$$\|(f-s)^{(j)}\|_{[0,2M]} \leqq \max_{1 \leqq i \leqq M-1} (f'(2i^+) - f'(2i^-)), \qquad 0 \leqq j \leqq k-1$$

and

$$s(x) = f(x) \quad \text{for } x \in [0, 1) \cup (2M-1, 2M].$$

Proof. Let $f_i(x) = f(x-2i)$, $i = 1, \cdots, M-1$. Then

$$f_i(x) = n_i(x) + (f_i'(0^+) - f_i'(0^-))(x)_+^1,$$

where n_i is convex since f_i', although possibly discontinuous, is increasing. Indeed, $n_i \in C^*[-2i, 2M-2i] \cap C^1*[-2, 2]$. Let

$$s_i(x) = n_i(x) + (f_i'(0^+) - f_i'(0^-))\frac{(x+1)^2}{4}.$$

Then s_i is a convex spline of order k with

$$s_i(\pm 1) = f_i(\pm 1), \qquad s_i'(\pm 1) = f_i'(\pm 1)$$

and

$$\|f_i^{(j)} - s_i^{(j)}\|_{[-1,1]} \leqq f_i'(0^+) - f_i'(0^-), \qquad 0 \leqq j \leqq k-1. \tag{4.1}$$

Now define $s(x)$ piecewise by

$$s(x) = s_i(x-2i) \quad \text{for } x \in [2i-1, 2i+1]$$

for $1 \leqq i \leqq M-1$, and $s(x) = f(x)$ for x otherwise. Then s is a spline of order k, C^1 and convex on each subinterval $(2i-1, 2i+1)$. The double interpolation at the odd integers forces the pieces to join up in a globally C^1 and globally convex manner. Finally, from (4.1),

$$\|(f-s)^{(j)}\|_{[0,2M]} \leqq \max_{1 \leqq i \leqq M-1} (f'(2i^+) - f'(2i^-)), \qquad j = 0, 1, \cdots, k-1. \qquad \square$$

Before smoothing to C^{k-2} we need some notation and some technical lemmas.

LEMMA 4.2. *Let $j \geqq 2$ be an integer. There exists a positive number $\delta = \delta(j)$ such that if q is any polynomial of degree $\leqq j$ with $q(0) = q'(0) = 0$ and $\|q\|_{[0,1]} = 1$ there is a subinterval I of $[0, 1]$ of length $1/(2(j-1))$ on which $|q''(x)| \geqq \delta$.*

Proof. Define

$$Q = \{q \in \pi_j : q(0) = q'(0) = 0 \text{ and } \|q\| = 1\}.$$

Consider $q \in Q$. q'' has at most $j-2$ zeros in $[0, 1]$. Hence q'' is nonzero on at least one open subinterval of length $1/(j-1)$. Let $[a, b]$ be a closed subinterval of length $1/(2(j-1))$ of this subinterval. Then by continuity and compactness $|q''|$ is bounded away from zero on $[a, b]$. It follows that with $I_a = [a, a + 1/(2(j-1))]$ and

$$\delta(q) = \sup_{I_a \subset [0,1]} \min_{x \in I_a} |q''(x)|,$$

$\delta(q) > 0$ for all $q \in Q$. Now define

$$\delta = \inf_{q \in Q} \delta(q).$$

If $\delta = 0$ then there must be a sequence $\{q_k\}_1^\infty \subset Q$ so $\delta(q_k) \to 0$. But then by compactness of Q,$\{q_k\}$ has a convergent subsequence $q_{k_i} \to q_0 \in Q$. But then $q''_{k_i} \to q''_0$ uniformly on $[0, 1]$ so that $\delta(q_0) = 0$. This contradiction shows $\delta > 0$. □

Let $k \geqq 4$ be an integer. Let $d = d(k) > k$ be an integer to be chosen later. Let $\mathbf{t} = \{t_i\}_{i=-\infty}^\infty$, where $t_i = i$ for $i \in [-d-k+1, d+k]$, and $\mathbf{s} = \{s_i\}_{i=-\infty}^\infty$, where $s_0 = s_1 = \cdots = s_{k-3} = 0$ and $s_{j+k-3} = j$ for $j \in [1, d+3]$, be two knot sequences. Let $N_{i,k,\mathbf{t}}$ and $N_{i,k,\mathbf{s}}$ be the normalized B-splines of order k corresponding to the knots at $\mathbf{t}$ and $\mathbf{s}$ as indicated by the third subscript. The normalization is as in the article of de Boor [1]. Denote by $\mathbf{z}$ the set of integers and by $N_{i,k}$, $N_{i,k,\mathbf{z}}$. We note that $N_{i,k} = N_{i,k,\mathbf{t}}$ for $i \in [-d, d]$, and that for $\{N_{i,k}\}$ two standard normalizations coincide, so that both

$$\sum_{j=i-k+1}^{i} N_{j,k}(t) = 1 \quad \text{on } [i, i+1)$$

and

$$\int_{-\infty}^{\infty} N_{i,k}(t)\, dt = \int_i^{i+k} N_{i,k}(t)\, dt = 1.$$

Let X be the subspace spanned by the set

$$A = \{N_{i,k,\mathbf{t}}\}_{-\infty}^{\infty} \cup \{N_{i,k,\mathbf{s}}\}_{i=0}^{k-4}.$$

Let Y be the subspace spanned by the set $\{N_{i,k,\mathbf{t}}\}_{-\infty}^{\infty}$. All the functions in Y are in $C^{k-2}[-d, d]$, while those in X are in $C^{k-2}([-d, 0) \cup (0, d])$ but may have discontinuities in their second and higher derivatives at zero. [The number of continuity conditions satisfied by $N_{i,k,\mathbf{r}}$ at ξ plus the number of knots from $\{r_i, \cdots, r_{i+k}\}$ coincident at ξ equals k.] A well-known basis theorem of Curry and Schoenberg (see, for example, [2, pp. 113–118]) shows that any function $f \in C^1[-d, d]$ whose restriction to $(-d, 0)$, $(0, d)$ are C^{k-2} splines with knots at the integers coincides, on $[-d, d]$, with a function in X. Following [3] define a "smoothing operator" T mapping X into Y by

$$T\Big(\sum_{i=-\infty}^{\infty} \alpha_i N_{i,k,\mathbf{t}} + \sum_{i=0}^{k-4} \beta_i N_{i,k,\mathbf{s}}\Big) = \sum_{i=-\infty}^{\infty} \alpha_i N_{i,k,\mathbf{t}},$$

and set $E = I - T$, where I is the identity operator. Note that for any $s \in X$, Es has support at most $[0, k-1]$. The following lemma is analogous to [3, Lemma 4.1]. Throughout the rest of the section $C_0, C_1, C_2, \cdots$ denote positive constants depending only on k, unless otherwise stated.

LEMMA 4.3. *There is a positive constant C_1 depending only on $k \geqq 4$ such that for any $s \in X$*

$$\|(Es)^{(j)}\|_{\mathbb{R}} \leqq C_1 \sum_{i=2}^{k-2} |s^{(i)}(0^+) - s^{(i)}(0^-)|,$$

for $0 \leqq j \leqq k-1$.

Proof. First, if

$$s = \sum_{i=0}^{k-4} \beta_i N_{i,k,\mathbf{s}} + \sum_{-\infty}^{\infty} \alpha_i N_{i,k,\mathbf{t}},$$

the β_i satisfy

$$\sum_{i=0}^{j-2} \beta_i [N_{i,k,\mathbf{s}}^{(j)}(0^+) - N_{i,k,\mathbf{s}}^{(j)}(0^-)] = s^{(j)}(0^+) - s^{(j)}(0^-), \qquad j = 2, \cdots, k-2.$$

Hence applying Cramer's rule to this triangular system we get

$$|\beta_i| < C_0 \sum_{j=2}^{k-2} |s^{(j)}(0^+) - s^{(j)}(0^-)|, \qquad i = 0, \cdots, k-4.$$

Now

$$\|(Es)^{(j)}\|_{\mathbb{R}} = \left\| \sum_{i=0}^{k-4} \beta_i N_{i,k,\mathbf{s}}^{(j)} \right\|_{\mathbb{R}} \leqq \sum_{i=0}^{k-4} |\beta_i| \, \|N_{i,k,\mathbf{s}}^{(j)}\|_{\mathbb{R}} \leqq C_1 \sum_{i=2}^{k-2} |s^{(j)}(0^+) - s^{(j)}(0^-)|. \qquad \square$$

LEMMA 4.4. *For each $k \geqq 4$ there exists a positive integer d and a constant C_2 with the following property. For every $f \in X^*$ whose restrictions to $(-d, 0)$ and $(0, d)$ are $(k-1)$st degree polynomials, there is an $s \in Y^*$ such that $s = f$ on $(-\infty, d)$ and (d, ∞) and*

$$\|f - s\|_{\mathbb{R}} \leqq C_2 \sum_{j=2}^{k-2} |f^{(j)}(0^+) - f^{(j)}(0^-)|.$$

Proof. Throughout this proof it will be assumed unless stated otherwise that $i \in [-d-k+1, d-1]$. d will be chosen as the least positive integer with $d \geqq 6k^2$ and $\delta_{k-1} d \geqq (3k-5)3k$, where δ_{k-1} is defined in Lemma 2.2.

Let F be the collection of functions f satisfying the hypotheses of the lemma and also satisfying $f(0) = f'(0) = 0$ and $\sum_{j=2}^{k-2} |f^{(j)}(0^+) - f^{(j)}(0^-)| \leqq 1$. It is sufficient to prove that for every $f \in F$ there is an $s \in Y^*$ such that $s(x) = f(x)$ for all $x \notin [-d, d]$ and $\|s - f\|_{\mathbb{R}} \leqq C_2$ for some constant C_2 depending only on k. We divide the proof into two cases, (i) $\|f\|_{[-d,d]} \leqq \alpha$ and (ii) $\|f\|_{[-d,d]} > \alpha$, where $\alpha > 0$ is to be chosen later.

(i) Suppose $f \in F$ and $\|f\|_{[-d,d]} \leqq \alpha$. Note that f is convex and nonnegative on $(-\infty, \infty)$, increasing on $(0, \infty)$ and decreasing on $(-\infty, 0)$. Let

$$Tf = \sum \alpha_i N_{i,k,\mathbf{t}}.$$

$Tf = f$ on the complement of $[0, k-1]$. Since f reduces to a polynomial of degree $\leqq k-1$ on $(-d, 0)$ and $(0, d)$ and $(Tf)^{(j)}(x) = \sum (\Delta^j \alpha_{i-j}) N_{i,k-j,\mathbf{t}}(x)$, $j = 0, 1, \cdots, k-1$, we find $\Delta^k \alpha_i = 0$ for $i \notin [-k, 0)$. Here Δ denotes the usual forward difference operator. It follows that there are two $(k-1)$st degree polynomials q_1 and q_2 so that $q_1(i) = \alpha_i$ for $i \geqq 0$ and $q_2(i) = \alpha_i$ for $i < 0$. Recall that the value of $\sum \beta_i N_{i,m,\mathbf{r}}$ at a point in $[r_e, r_{e+1}]$ is a convex combination of only m B-spline coefficients, namely $\beta_{e-m+1}, \cdots, \beta_e$. Using this and that $Tf(x) = f(x)$ for $x \notin [0, k-1]$, we see that in each segment of length k $\{\alpha_l, \cdots, \alpha_{l+k-1}\}$ of $\{\alpha_i\}$ where $l \geqq 0$ there must be indices i_1, i_2, i_3 such that $\alpha_{i_1} \geqq 0$, $\Delta\alpha_{i_2} \geqq 0$ and $\Delta^2 \alpha_{i_3} \geqq 0$. Furthermore, since q_1, q_1', q_1'' have at most $k-1$, $k-2$,

$k-3$ sign changes respectively, we find that in any segment of $(3k-5)(2k+1)$ indices in $[0, d)$ there exists a subsegment of length $2k+1$, $U=\{l^*, \cdots, l^*+2k\}$ with $\alpha_i \geqq 0$, $\Delta\alpha_i \geqq 0$, $\Delta^2\alpha_i \geqq 0$ for all $i \in U$. We have chosen d so large that such a segment U can be found within the set $[(1-\delta_{k-1})d, d)$. Similar arguments show that there exists a segment of $2k+1$ indices $L=\{l_*, \cdots, l_*+2k)$ within the set $[-d, -d(1-\delta_{k-1})$ with $\alpha_i \geqq 0$, $\Delta\alpha_i \leqq 0$, $\Delta^2\alpha_i \geqq 0$ for $i \in L$.

Now considering the monotony of f and f' and of the B-spline coefficients with indices in U, we find

$$\alpha_{l^*+k-1} \geqq 0,$$
$$\alpha_{l^*+k} \leqq f(l^*+2k) \leqq \|f\|_{[(1-\delta_{k-1})d,d]}$$

and

$$\Delta\alpha_{l^*+k-1} \geqq f'(l^*+k) \geqq \min_{[(1-\delta_{k-1})d,d]} f'(x).$$

From Lemma 2.2 it follows that the linear function p^*, with $p^*(l^*+k-1)=\alpha_{l^*+k-1}$ and $p^*(l^*+k)=\alpha_{l^*+k}$, intercepts the x-axis at some $x^* \geqq 0$. Similarly the linear function p_*, with $p_*(l_*+k-1)=\alpha_{l_*+k-1}$ and $p_*(l_*+k)=\alpha_{l_*+k}$, intercepts the x-axis at $x_* \leqq 0$.

We define the spline s via its B-spline coefficients. $s=\sum \beta_i N_{i,k,\mathbf{t}}$, where

$$\beta_i = \begin{cases} \alpha_i & \text{for } i \notin [l_*+k+1, l^*+k-2], \\ p_*(i) & \text{for } l_*+k+1 \leqq i < x_*, \\ 0 & \text{for } x_* \leqq i \leqq x^*, \\ p^*(i) & \text{for } x^* < i \leqq l^*+k-2. \end{cases}$$

Then $s''(x)=\sum(\Delta^2\beta_{i-2})N_{i,k-2,\mathbf{t}}$, and s is convex on $[l_*+k-1, l^*+2k]$ since the corresponding segment of B-spline coefficients is convex. But $s(x)=Tf(x)=f(x)$ on $(-\infty, l_*+k+1) \cup (l^*+2k-1, \infty)$. Hence s is globally convex.

It remains to show the estimate of $\|f-s\|_{\mathbb{R}}$. First, the size of each B-spline coefficient depends only on the size of the function nearby. More precisely (see de Boor [2, pp. 154–155]) $|\alpha_i| \leqq C_3\|f\|_{[-d,d]} \leqq C_3\alpha$, $i \in [-d, d-k]$. Hence $|\beta_i| \leqq C_3\alpha$ for $i \in [-d, d-k]$ and certainly $\beta_i=\alpha_i$ on the complement of this set. Hence $\|Tf-s\|_{\mathbb{R}} \leqq \sum |\alpha_i-\beta_i| \, \|N_{i,k,\mathbf{t}}\|_{\mathbb{R}} \leqq C_4\alpha$. It follows that $\|f-s\|_{\mathbb{R}} \leqq \|f-Tf\|_{\mathbb{R}} + \|Tf-s\|_{\mathbb{R}} \leqq C_5$, where C_5 depends only on k and α. Here we have used Lemma 4.3.

(ii) Suppose now $f \in F$ and $\|f\|_{[-d,d]} > \alpha$. Let $[l, l+2(k-2)]$ be an interval in $[-d, d] \setminus (0, k-1)$ to be chosen later. Now recall that $(Tf)(x)=f(x)$ outside $[0, k-1]$, and set

$$s(x)=(Tf)(x)+\int_{-\infty}^{x}\int_{-\infty}^{t} C_1 \sum_{i=-k+3}^{k-2} N_{i,k-2}(v) - (y_1 N_{l,k-2}(v) + y_2 N_{l+k-2,k-2}(v)) \, dv \, dt$$
$$=(Tf)(x)+M(x),$$

where C_1 is the constant introduced in Lemma 4.3 and y_1, y_2 are to be chosen so that $M(x)$ vanishes outside of $[-d, d]$. With such a choice of $M(x)$ $s(x)$ will agree with $f(x)=(Tf)(x)$ for $x \notin [-d, d]$.

The form of M implies $M(x)=0$ for $x \leqq -d$ and $M''(x)=0$ for $x \geqq d$. Hence y_1 and y_2 are to be chosen so that $M(d)=M'(d)=0$. Since $\int_{-\infty}^{\infty} N_{i,k-2}(t)\,dt=1$ and

$\int_{-\infty}^{\infty} tN_{i,k-2}(t)\,dt = i+(k-2)/2$, the equations for y_1 and y_2 are

$$\begin{bmatrix} 1 & 1 \\ l+\dfrac{k-2}{2} & l+\dfrac{3k-6}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} C_1(2k-4) \\ C_1 \sum\limits_{i=-k+3}^{k-2} \left(i+\dfrac{k-2}{2}\right) \end{bmatrix},$$

or $A\mathbf{y} = \mathbf{b}$. There is always a unique solution since determinant $(A) = k-2$. Also the elements of A and $\mathbf{b}$ are bounded by a constant depending only on k. Hence by Cramer's rule y_1 and y_2 are bounded by a constant C_6 depending only on k, and

$$\|s-f\|_{[-d,d]} \leqq \|f-Tf\|_{[-d,d]} + \|M\|_{[-d,d]} \leqq C_7.$$

It remains to show that if $\alpha = \alpha(k)$ is chosen sufficiently large s will be convex. Recalling our choice of d and using Lemma 4.2, we choose α so large that if $\|f\|_{[-d,d]} > \alpha$ then the set $[l, l+2(k-2)]$ can be chosen so that $f''(x) \geqq C_6$ everywhere on the set. Then, since $|y_1|, |y_2| \leqq C_6$, $s''(x)$ is nonnegative on $[l, l+2(k-2)]$. Since $\sum_{i=-k+3}^{k-2} N_{i,k-2}(x) = 1$ for $x \in [0, k-1]$ and $\|f''-(Tf)''\|_{[0,k-1]} \leqq C_1$, $s''(x)$ is also nonnegative on $[0, k-1]$. But $s''(x) = f''(x)$ for all other x's. Hence s is convex on $\mathbb{R}$. □

We are now ready to prove the main result.

Proof of Theorem 1. For $k=2$ Theorem 1 follows from well-known properties of the polygonal approximation. Assume in what follows that $k \geqq 3$ and $f \in C^j{}^*[0,1]$, $0 \leqq j \leqq k-1$. It is sufficient to consider $N > 12d$, where d is the d defined in Lemma 4.4, since, for $N \leqq 12d$, the spline will follow from the polynomial result of Lemma 3.2.

Set $\hat{f}(t) = f(t/N)$ and let M be the integer part of $N/6d$ so that $N-6d < 6Md \leqq N$. If $j \geqq 2$, then by applying Lemma 3.2 on each of the intervals $I_1 = [0, 6d]$, $I_2 = [6d, 12d]$, $\cdots$, $I_{M-1} = [6d(M-2), 6d(M-1)]$ and $I_M = [6d(M-1), N]$ we find a convex function $\hat{g}$ whose restriction to each of the intervals $I_1, \cdots, I_M$ is a polynomial of degree $\leqq j$ with

$$\|(\hat{g}-\hat{f})^{(i)}\|_{[0,N]} \leqq C_8 \omega(\hat{f}^{(j)}, [0,N], 1), \tag{4.2}$$

for $0 \leqq i \leqq j$. Here we have used that all the intervals I_e have lengths between $6d$ and $12d$ to relate $\hat{f}^{(j)}$ to $\bar{f}^{(j)}$, where $\bar{f}$ is the function derived from $\hat{f}$ by the linear change of variable taking I_e onto $[0,1]$. If $j < 2$ then the polygonal approximation $\hat{g}$, with knots $0, 6d, 12d, \cdots, 6d(M-1), N$, is convex with $\hat{f}$ and satisfies

$$\|(\hat{g}-\hat{f})^{(i)}\|_{[0,N]} \leqq C_9 \omega(\hat{f}^{(j)}, [0,N], 1) \tag{4.3}$$

for $i = 0, j$. From (4.2) and (4.3), and since $\|\hat{g}^{(i)}\|_{[0,N]} = 0$, $i > e$, if $\hat{g}$ is a piecewise polynomial of degree $\leqq e$ we have

$$\sum_{i=1}^{k-2} |\hat{g}^{(i)}(6ld^+) - \hat{g}^{(i)}(6ld^-)| \leqq C_{10} \omega(\hat{f}^{(j)}, [0,N], 1), \tag{4.4}$$

for all $l = 1, \cdots, M-1$.

We proceed to smooth the $C^*[0,N]$ approximation to $C^1{}^*[0,N]$. Applying Lemma 4.1 to $\hat{g}$ on $[0, 6dM]$ we get a C^1 convex function $\hat{h}$ whose restriction to each interval $J_1 = [0, 3d]$, $J_2 = [3d, 6d]$, $\cdots$, $J_{2M-1} = [6d(M-1), 6d(M-1)+3d]$, and $J_{2M} = [6d(M-1)+3d, N]$, is a polynomial of degree $\leqq k-1$ such that

$$\begin{aligned} \|(\hat{h}-\hat{g})^{(i)}\|_{[0,N]} &\leqq C_{11} \max_{1 \leqq e \leqq M-1} (\hat{g}'(6de^+) - \hat{g}'(6de^-)) \\ &\leqq C_{12} \omega(\hat{f}^{(j)}, [0,N], 1), \qquad 0 \leqq i \leqq k-1. \end{aligned} \tag{4.5}$$

Now we are ready to smooth from $C^1*[0, N]$ to $C^{k-2}*[0, N]$. If $k \geqq 4$, apply Lemma 4.4 to $\hat{h}$ on each of the intervals $J_l \cup J_{l+1}$, $l = 1, 2, \cdots, 2M-1$ in turn. This yields a convex spline of order k, $\hat{s}$, with simple knots at the integers such that

$$\hat{s}(x) = \hat{h}(x), \qquad x \notin \bigcup_{l=1}^{2M-1} [(3l-1)d, (3l+1)d] \tag{4.6}$$

and

$$\begin{aligned} \|\hat{s} - \hat{h}\|_{[(3l-1)d,(3l+1)d]} &\leqq C_2 \sum_{i=2}^{k-2} |\hat{h}^{(i)}(3ld^+) - \hat{h}^{(i)}(3ld^-)| \\ &\leqq C_{13}\omega(\hat{f}^{(j)}, [0, N], 1), \end{aligned} \tag{4.7}$$

$1 \leqq l \leqq 2M-1$, by (4.4) and (4.5). If $k = 3$, take $\hat{s} = \hat{h}$. Then

$$\|\hat{f} - \hat{s}\|_{[0,N]} \leqq \|\hat{f} - \hat{g}\|_{[0,N]} + \|\hat{g} - \hat{h}\|_{[0,N]} + \|\hat{h} - \hat{s}\|_{[0,N]} \leqq C_{14}\omega(\hat{f}^{(j)}, [0, N], 1),$$

by (4.2), (4.3), (4.5), (4.6) and (4.7). Finally, putting $s(x) = \hat{s}(Nx)$ and noting $\omega(\hat{f}^{(j)}, [0, N], 1) = N^{-j}\omega(f^{(j)}, [0, 1], N^{-1})$, we obtain Theorem 1. □

5. Remarks. The concept of building a constrained piecewise polynomial approximation and then smoothing it to a constrained spline approximation is due to Chui, Smith and Ward [3]. Their ideas, and the techniques used in the present paper, are applicable to other problems. For example the author has obtained a Jackson type theorem for approximation of nonnegative functions f by splines s satisfying $0 \leqq s(x) \leqq f(x)$, $x \in [0, 1]$, with related methods. Another problem to which such methods may apply is the problem of approximating j-convex functions ($j > 2$) by j-convex splines.

Acknowledgment. I would like to thank Professors C. K. Chui, P. W. Smith and J. D. Ward for several helpful conversations.

REFERENCES

[1] C. DE BOOR, *Splines as linear combinations of B-splines*, in Approximation Theory II, G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 1–47.

[2] ———, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[3] C. K. CHUI, P. W. SMITH AND J. D. WARD, *Degree of L^p-approximation by monotone splines*, this Journal, 11 (1980), pp. 436–447.

[4] ———, *Monotone approximation by spline functions*, Proceedings of the 1979 Conference on Approximation Theory, Bonn, to appear.

[5] R. A. DE VORE, *Monotone approximation by splines*, this Journal, 8 (1977), pp. 891–905.

[6] ———, *Monotone approximation by polynomials*, this Journal, 8 (1977), pp. 906–921.

[7] W. W. HAGER, *Convex control and dual approximations, Part* II, Control and Cybernetics, 8 (1979), pp. 73–86

[8] G. G. LORENTZ AND K. L. ZELLER, *Degree of approximation by monotone polynomials* 1, J. Approximation Theory, 1 (1968), pp. 501–504.

[9] W. SIPPEL, *Approximation by functions with restricted ranges*, in Approximation Theory, G. G. Lorentz, ed., Academic Press, New York, 1973, pp. 481–484.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0007 $01.00/0

AN INTEGRODIFFERENTIAL EQUATION FOR PLANE WAVES PROPAGATING INTO A RANDOM FLUID: ASYMPTOTIC BEHAVIOR*

M. J. LEITMAN†

Abstract. Beran and McCoy [J. Math. Phys., 17(1976), pp. 1186–1189], [J. Acoust. Soc. Amer., 56(1974), pp. 1667–1672] have developed a mathematical model for the propagation of acoustic waves in water which incorporates the scattering effect of random microscopic variations in density (sound speed) into the classical model of geometric optics. This work characterizes the behavior of the acoustic intensity spectral density as a function of distance from the source.

The problem is formulated as a Cauchy problem in L^p-space. It is shown that nonnegative initial profiles produce nonnegative solution profiles which conserve intensity (area under the nonnegative solution profile). The scattering effect is shown to cause dispersal (mean-square decay) and consequent loss in resolution of the wave as it moves away from the source. This loss is expected to occur more slowly than any exponential, which is the case if there were no scattering (the optical model).

Some comments on the approximation of solutions are included. In particular, the last phenomenon is lost in most approximations.

Finally the connection between solutions to this problem and spatially inhomogeneous Markov processes is established. Specifically, the original Cauchy problem constitutes the Kolmogorov equation associated with the process.

1. Introduction and main results. Beran and McCoy [7], [8], [13] have developed a mathematical model for the propagation of acoustic waves in water which incorporates the scattering effect of microscopic variations in density (sound speed) into the classical model of geometric optics. This work describes the dispersion and consequent loss in coherence of the wave as a function of distance from the source.[1]

If we let the function $\mu \to x_t(\mu)$, $-\infty < \mu < \infty$, denote the acoustic intensity spectral density at a distance $t \geqq 0$ from the source, then Beran and McCoy [13, eq. (23)] show that under suitable physical assumptions x_t satisfies an initial value problem of the following form:

$$\frac{d}{dt}x_t(\mu) = \int_{-\infty}^{\infty} \Psi(\mu, \nu)(x_t(\nu) - x_t(\mu))\, d\nu, \qquad t \geqq 0,$$

$$x_0(\mu) = \mathring{x}(\mu), \qquad -\infty < \mu < \infty,$$

where Ψ is a kernel determined by the physics of the problem and $\mathring{x}$ is a prescribed nonnegative state.

We require the kernel Ψ to satisfy the following standing hypotheses:[2]

(A1) $\Psi(\mu, \nu) > 0$ (positivity);

(A2) $\Psi(\mu, \nu) = \Psi(\nu, \mu)$ (symmetry);

(A3) $\Psi(\mu, \nu) \leqq \bar{\Psi} < \infty$ (boundedness);

(A4) $\varphi(\mu) \overset{\text{def}}{=} \int \Psi(\mu, \nu)\, d\nu \leqq \bar{\varphi} < \infty$ (sectional integrability and boundedness);

(A5) $\lim_{M\to\infty} \frac{1}{M} \int_{|\nu|\geqq M} \int_{|\mu|\leqq M} \Psi(\mu, \nu)\, d\mu\, d\nu = 0.$

* Received by the editors March 20, 1979, and in final revised form October 10, 1980.

† Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio, 44106.

[1] Computational aspects of this problem and a short treatment of the asymptotics may be found in a paper by Beran, Leitman and Schwartz [15]. Also, announcement and discussion of some of the results contained here appear in [16].

[2] All equalities and inequalities between functions of μ hold almost everywhere with respect to Lebesgue measure. All integrals are Lebesgue integrals over the entire real line, unless specifically stated otherwise.

Properties (A1)–(A4) arise more or less naturally from the physics of the model; (A5) is a technical assumption which is satisfied in the cases of interest. From the point of view of existence, uniqueness and asymptotic stability of solutions to the initial value problem, some of these hypotheses are overly restrictive. Satisfactory results obtain under weaker assumptions.

The initial value problem just described has the form of a transport equation. It is not unique to the model deduced by Beran and McCoy. Indeed, under the hypotheses (A1)–(A4), it is a "master" or "kinetic" equation of the type which has been derived by the statistical treatment of other models in mathematical physics which incorporate random effects. For example, see the work of Besieris and Tappert [5], Papanicolaou [4], Papanicolaou and Kohler [6], [9], [10], [14] and Papanicolaou and Keller [3] for additional discussion of the derivation and validity of equations of this type. The present context suffices to provide a vehicle to consider the mathematical features of the problem.

It is important throughout to bear in mind two specific examples. The *optical model*, without scattering, has a kernel of convolution type:

$$\Psi(\mu, \nu) = f(\mu - \nu).$$

In this case (A1)–(A4) are equivalent to $f \geqq 0, f \in L^1 \cap L^\infty$ and f even: moreover, (A5) is a consequence of (A1)–(A4). The *acoustic model*, with scattering, has a kernel of nonconvolution form typified by

$$\Psi(\mu, \nu) = [1 + (\mu^2 - \nu^2)^2]^{-1}.$$

Henceforth we suppose that $\mu \in (-\infty, \infty)$, $t \in [0, \infty)$ and write L^p for $L^p((-\infty, \infty))$, $1 \leqq p \leqq \infty$. If $x \in L^p$ and $y \in L^q$, where p and q are conjugate, we write

$$\langle x, y \rangle = \int x(\mu) y(\mu)\, d\mu.$$

It is also convenient to define an operator A in the class of functions on $(-\infty, \infty)$ by

$$(Ax)(\mu) = \int \Psi(\mu, \nu)(x(\nu) - x(\mu))\, d\nu.$$

In terms of the operator A, the initial value problem has the form of a Cauchy problem:

$$\frac{d}{dt} x_t = A x_t, \qquad t \geqq 0,$$

$$x_0 = \mathring{x}.$$

Regarding the solutions of this problem we assert the following:

(T1) *Each initial state* $\mathring{x} \in L^p$ *determines a unique solution* $x_t = T_t \mathring{x}$, *where* $T_t \equiv \exp tA$, $t \geqq 0$, *is an analytic semigroup of bounded linear operators in* L^p, $1 \leqq p \leqq \infty$.

(T2) *For each* $t \geqq 0$, T_t *is a positive linear operator in* L^p, $1 \leqq p \leqq \infty$, *so that*

$$x \geqq 0 \Rightarrow T_t x \geqq 0.$$

(T3) *For each* $t \geqq 0$, $\|T_t\|_p = 1$, $1 \leqq p \leqq \infty$, *and, more specifically,*

(i) *if* $p = 1$, $x \geqq 0$, $x \not\equiv 0$ *then* $0 < \|T_t x\|_1 = \|x\|_1$;

(ii) *if* $p = 2$, $x \not\equiv 0$ *then* $0 < \|T_t x\|_2 \leqq \|x\|_2$;

(iii)[3] *if* $p = \infty$, $x \equiv 1^*$ *then* $T_t 1^* = 1^*$.

[3] 1^* denotes the constant function on $(-\infty, \infty)$ with value 1.

(T4) *For* $x \not\equiv 0$, $t \to \|T_t x\|_2$ *is strictly decreasing, and*

$$\lim_{t\to\infty} \|T_t x\|_2 = 0 \quad \forall x \in L^2.$$

(T5) *There is at least one* $\mathring{x} \in L^2$, $\mathring{x} \geqq 0$, *such that*

$$\lim_{t\to\infty} \frac{d}{dt} \ln \|T_t \mathring{x}\|_2 = 0. \tag{**}$$

We thus see that nonnegative initial profiles produce nonnegative solution profiles (T2), and that intensity (area under nonnegative solution profiles) is conserved (T3(i)). However, profiles flatten out and coherence is lost in the sense that $x_t = T_t \mathring{x}$ decreases to zero in the mean square as $t \to \infty$ (T3(ii)), (T4). The most significant assertion, (T5), implies that the rate at which coherence is lost is expected to be very slow. For if (**) holds, then every positive initial state lies arbitrarily close to one for which the rate of decay, in the mean square, is slower than exponential.[4]

Before proving these assertions we make some additional remarks. If no scattering is assumed, so that A has a convolution kernel, then (**) holds for *every* positive initial state $\mathring{x} \in L^2$. This is a straightforward consequence of Plancherel's theorem. We are thus motivated to make the following

Conjecture. Under the hypotheses (A1)–(A5), solutions to the Cauchy problem (*) satisfy (**) for every nonnegative initial state $\mathring{x} \geqq 0$, $\mathring{x} \not\equiv 0$ in L^2.

The nonnegativity of the initial state seems to be essential here. To see why, we suppose that A has a nonnegative eigenfunction $\hat{x} \geqq 0$ in L^2. In the sequel (P2) we show that A is negative definite and symmetric in L^2, consequently the eigenvalue $\hat{\lambda}$ associated with $\hat{x}$ must be negative. Taking $\mathring{x} = \hat{x}$, we get $T_t \hat{x} = e^{\hat{\lambda} t} \hat{x}$, so that $\hat{x}$ clearly violates (**) in (T5). Note that $\hat{x} \geqq 0$ is also an eigenvector of T_t corresponding to the eigenvalue $e^{\hat{\lambda} t}$. Now we also show in the sequel (T2), (T3) that T_t is positive, symmetric, and $\|T_t\|_2 = 1$. A result of Coffman, Duffin and Mizel [12] then guarantees that $e^{\hat{\lambda} t} = 1$ and, hence, $\hat{\lambda} t = 0$. This is impossible, so that A cannot have a nonnegative eigenfunction.[5] Of course, this fact alone does not validate the conjecture; but if there is a counterexample, it cannot be provided by an eigenfunction of A.

2. Proofs of the main assertions. Verification of (T1)–(T5) depends upon establishing certain properties of the generator A. We do this in a series of four propositions (P1)–(P4).

It will be convenient to use F to denote the usual integral operator induced by the kernel Ψ; that is,

$$(Fx)(\mu) = \int \Psi(\mu, \nu) x(\nu)\, d\nu.$$

Thus, in terms of F, $Ax = Fx - \varphi x$,

(P1) (A1)–(A4) $\Rightarrow A : L^p \to L^p$, $1 \leqq p \leqq \infty$, *is a bounded linear operator whose norm satisfies*

$$\|A\|_p \leqq 2\bar{\varphi}.$$

Proof of (P1). Since φ is bounded (A4), it suffices to show that $F : L^p \to L^p$, and that $\|F\|_p \leqq \bar{\varphi}$, for $1 \leqq p \leqq \infty$. This result is easily verified in the case where $p = 1$ or

[4] The condition (**) in (T5) neither implies nor is implied by the assertion: for $\varepsilon > 0$ there is a $\hat{t}(\varepsilon) \geqq 0$ such that $\|T_t \mathring{x}\|_2 \leqq e^{-\varepsilon t}$, $t \geqq \hat{t}(\varepsilon)$.

[5] If A has a nonnegative eigenfunction $\hat{x}$ in L^1 then $A\hat{x} = \hat{\lambda}\hat{x}$. As will be seen (P3(ii)) $\int (Ax)(\mu)\, d\mu = 0$, so that $\hat{\lambda} \int |x(\mu)|\, d\mu = 0$. Thus, $\hat{\lambda} = 0$. But we will also show (P3(i)), that $A\hat{x} = 0 \Rightarrow \hat{x} = 0$ in L^1. Thus A cannot have a nonnegative eigenfunction in L^1 either.

$p=\infty$. When $1<p<\infty$ we can appeal to the Riesz convexity theorem [2] or proceed directly as follows. Since F is positive (A1) we need only consider $x \geqq 0$ in L^p. Then, for $q=p/(p-1)$,

$$\begin{aligned}
\|Fx\|_p^p &= \int \left(\int \Psi(\mu,\nu)x(\nu)\,d\nu \right)^p d\mu \\
&= \int \left(\int \Psi^{1/q}(\mu,\nu)\Psi^{1/p}(\mu,\nu)x(\nu)\,d\nu \right)^p d\mu \\
&\leqq \int \left(\left(\int \Psi(\mu,\nu)\,d\nu \right)^{1/q} \left(\int \Psi(\mu,\nu)x_{(\nu)}^p\,d\nu \right)^{1/p} \right)^p d\mu \\
&\leqq \iint x^p(\nu)\Psi(\mu,\nu)\varphi^{p/q}(\mu)\,d\nu\,d\mu \\
&\leqq \bar{\varphi}^{p/q} \int \left(\int \Psi(\mu,\nu)\,d\mu \right) x^p(\nu)\,d\nu \\
&\leqq \bar{\varphi}^{p/q+1} \int x^p(\nu)\,d\nu.
\end{aligned}$$

Thus $\|Fx\|_p \leqq \bar{\varphi}\|x\|_p$, and (P1) is proved. □

For case $p=2$, the above result is an integral version of Shur's theorem for matrices. More generally we have the following result. If there are positive constants c, d and positive functions f, g such that $Ff \leqq cg^{q-1}$ and $Fg \leqq df^{p-1}$, then $F: L^p \to L^p$ and $\|F\|_p \leqq c^{p-1}d$. This follows by the same argument as used in the proof of (P1).

(P2) (A1)–(A4) *imply, for* $A: L^2 \to L^2$:

(i) $\langle Ax, y\rangle = \langle x, Ay\rangle$;

(ii) $x \neq 0 \Rightarrow \langle Ax, x\rangle < 0$;

(iii) *if, in addition,* (A5) *holds, then*

$$\sup_{\|x\|_2=1} \langle Ax, x\rangle = 0.$$

Thus (A1)–(A4) *imply that* A *is a symmetric, negative definite linear operator in* L^2, *which contains the point zero in its continuous spectrum whenever* (A5) *holds as well.*

Proof of (P2). First we establish a useful formula for $\langle Ax, y\rangle$. Using the symmetry of Ψ (A2) an easy calculation yields

$$\langle Ax, y\rangle = -\frac{1}{2}\iint \Psi(\mu,\nu)[x(\mu)-x(\nu)][y(\mu)-y(\nu)]\,d\nu\,d\mu.$$

Clearly A is symmetric in L^2. Setting $y=x$ in $\langle Ax, y\rangle$ we get

$$\langle Ax, x\rangle = -\frac{1}{2}\iint \Psi(\mu,\nu)[x(\mu)-x(\nu)]^2\,d\nu\,d\mu.$$

Thus $\langle Ax, x\rangle \leqq 0$ for every $x \in L^2$. Furthermore, since Ψ is positive (A1), it follows that $\langle Ax, x\rangle = 0$ if and only if the function $(\mu,\nu) \to [x(\mu)-x(\nu)]$ vanishes on $(-\infty,\infty)\times(-\infty,\infty)$, in which case the function $\mu \to x(\mu)$ is constant on $(-\infty,\infty)$. Since $x \in L^2$, this constant must be zero. Thus A is negative definite in L^2.

Finally, suppose that (A5) holds as well as (A1)–(A4). Define x_M in L^2 for $M>0$ by $x_M = (2M)^{-1/2}\mathscr{X}_{[-M,M]}$.[6] Then $\|x_M\|_2 = 1$, for each $M>0$, and (A5) guarantees that $\lim_{M\to\infty}\langle Ax_M, x_M\rangle = 0$. Hence zero is in the continuous spectrum of A. □

Since the assumption (A5) plays an important role in (P2), the following comments are relevant. The typical acoustic kernel given earlier satisfies (A5). Every convolution (optical) kernel which satisfies (A1)–(A4) also satisfies (A5). And there are kernels

[6] $\mathscr{X}_{[-M,M]}$ denotes the characteristic function of the interval $[-M, M]$ in $(-\infty,\infty)$.

which satisfy (A1)–(A4) which do not satisfy (A5). Indeed, if $\mathscr{C}$ denotes the set $\mathscr{C}=\{(\mu,\nu):|\mu\pm 2\nu|\leqq 1, |2\mu\pm\nu|\leqq 1\}$, then $\mathscr{K}_{\mathscr{C}}$ satisfies (A2)–(A4) but not (A5). Just add $\mathscr{K}_{\mathscr{C}}$ to any kernel Ψ which satisfies (A1)–(A4) to obtain such an example.

(P3) (A1)–(A4) *implies*

(i) $x\in L^p$, $1\leqq p<\infty$, $Ax=0\Rightarrow x=0$;

(ii) $x\in L^1\Rightarrow\int (Ax)(\mu)\,d\mu=0$;

(iii) $A1^*=0$.

Proof of (P3).

(i) First suppose $p=1$. Now $Ax=0$ means $\varphi x=Fx$, and hence $\varphi|x|\leqq F(|x|)$. Since Ψ is bounded (A3), it follows that $\varphi x\in L^\infty$, and hence $\varphi x^2\in L^1$. Then

$$\begin{aligned}\int \varphi(\mu)|x(\mu)|^2\,d\mu &= \int\int \Psi(\mu,\nu)|x(\mu)|^2\,d\nu\,d\mu\\ &\leqq \int\int \Psi(\mu,\nu)|x(\nu)|\,|x(\mu)|\,d\nu\,d\mu.\end{aligned}$$

The symmetry of Ψ (A2) then yields

$$\int\int \Psi(\mu,\nu)[|x(\mu)|^2-|x(\mu)|\,|x(\nu)|]\,d\nu\,d\mu\leqq 0$$

and

$$\int\int \Psi(\mu,\nu)[|x(\nu)|^2-|x(\mu)|\,|x(\nu)|]\,d\nu\,d\mu\leqq 0.$$

Adding, we get

$$\int\int \Psi(\mu,\nu)[|x(\mu)|-|x(\nu)|]^2\,d\nu\,d\mu\leqq 0.$$

As in the proof of (P2) we find that $x=0$ in L^1.

Now suppose $x\in L^p$, $(x\not\equiv 0)$, $1<p<\infty$. Then $Ax=0$ implies

$$\begin{aligned}\varphi(\mu)|x(\mu)|&\leqq\int \Psi(\mu,\nu)|x(\nu)|\,d\nu\\ &<\left(\int \Psi(\mu,\nu)\,d\nu\right)^{1/q}\left(\int \Psi(\mu,\nu)|x(\nu)|^p\,d\nu\right)^{1/p}\\ &<\varphi^{1/q}(\mu)\left(\int \Psi(\mu,\nu)|x(\nu)|^p\,d\nu\right)^{1/p}.\end{aligned}$$

This yields $\varphi|x|^p<F(|x|^p)$, which is the same inequality as in the case $p=1$ with $|x|\in L^1$ replaced by $|x|^p\in L^1$. Then $x=0$ in L^p.

(ii) If $x\in L^1$ then φx, Fx, and Ax are all in L^1. The result then follows by Fubini's theorem.

(iii) Since $F1^*=\varphi 1^*$, we get $A1^*=0$. Thus (P3) is proved. □

Incidentally, computations similar to those we have used yield

$$\|Ax\|_p^p<\bar{\varphi}^{p-1}\int\int \Psi(\mu,\nu)[x(\nu)-x(\mu)]^p\,d\nu\,d\mu$$

for $1\leqq p<\infty$, provided $x\not\equiv 0$. In particular, if $p=2$,

$$\|Ax\|_2^2<-2\bar{\varphi}\langle Ax,x\rangle$$

as expected.

(P4)[7] (A1)–(A4) $\Rightarrow J_\lambda \equiv (I-(1/\lambda)A)^{-1}$, $\lambda>0$, *is a strictly positive contraction in* L^p, $1\leqq p\leqq\infty$: *for* $\lambda>0$,

$$x\geqq 0,\quad x\not\equiv 0\Rightarrow J_\lambda x>0,\quad \|J_\lambda\|_p\leqq 1.$$

More specifically,

(i) *if* $p=1$, $\|J_\lambda\|_1=1$, *and* $x\geqq 0\Rightarrow\|J_\lambda x\|_1=\|x\|_1$;
(ii) *if* $p=\infty$, $\|J_\lambda\|_\infty=1$, *and* $J_\lambda 1^*=1^*$; *and*
(iii) *if* $p=2$, $x\neq 0\Rightarrow\|J_\lambda x\|_2<\|x\|_2$, *and* $\|J_\lambda\|_2=1$ *whenever* (A5) *also holds.*
Finally, (A1)–(A5) $\Rightarrow\|J_\lambda\|_p=1$, $1\leqq p\leqq\infty$, *for* $\lambda>0$.

Proof of (P4). First we verify the cases $p=1, 2, \infty$; the general case $1\leqq p\leqq\infty$ will follow. Note that, for $\lambda>0$, $x=J_\lambda y$ must be a solution of

$$(***)\qquad x-\frac{1}{\lambda}Ax=y\quad\text{or equivalently}\quad\left(1+\frac{1}{\lambda}\varphi\right)x=y+\frac{1}{\lambda}Fx.$$

(i) *Case* $p=1$. Since $\varphi/(\lambda+\varphi)\leqq\bar\varphi/(\lambda+\bar\varphi)$, it follows that the composition $F(1/(\lambda+\varphi))I$ is a strict contraction on L^1: $\|F(1/(\lambda+\varphi))I\|_1\leqq\bar\varphi/(\lambda+\bar\varphi)<1$. Hence J_λ given by

$$J_\lambda=\frac{\lambda}{\lambda+\varphi}\sum_{k=0}^{\infty}\left[F\left(\frac{1}{\lambda+\varphi}\right)I\right]^k$$

is well defined on L^1 and $x=J_\lambda y$ solves (***). Since Ψ is positive we see that

$$(\lambda+\varphi)|x|\leqq\lambda|y|+F(|x|).$$

Fubini's theorem and (A4) then yield

$$\lambda\int|x(\mu)|\,d\mu+\int\varphi(\mu)|x(\mu)|\,d\mu\leqq\lambda\int|y(\mu)|\,d\mu+\int\varphi(\mu)|x(\mu)|\,d\mu.$$

Thus $\|J_\lambda y\|_1\leqq\|y\|_1$ for all $y\in L^1$, and $\|J_\lambda\|_1\leqq 1$. To see that J_λ is strictly positive, note that $F(1/(\lambda+\varphi))I$ is strictly positive and use the series for J_λ. Alternatively, use (P3(ii)) to get

$$\int x(\mu)\,d\mu=\int y(\mu)\,d\mu.$$

But $y\geqq 0$ and $\|J_\lambda\|_1\leqq 1$ imply

$$\int|x(\mu)|\,d\mu\leqq\int y(\mu)\,d\mu.$$

Hence $x\geqq 0$. Since Ψ and $(1+(1/\lambda)\varphi)$ are positive, so is $x=J_\lambda y>0$.

To see that $\|J_\lambda\|_1=1$, it suffices to consider $y\geqq 0$. Since J_λ is positive and

$$\int(J_\lambda y)(\mu)\,d\mu=\int y(\mu)\,d\mu,$$

we get $\|J_\lambda y\|_1=\|y\|_1$ for $y\geqq 0$. This means that $\|J_\lambda\|_1\geqq 1$. Thus $\|J_\lambda\|_1=1$ as claimed.

(ii) *Case* $p=\infty$. Since $\varphi/(\lambda+\varphi)\leqq\bar\varphi/(\lambda+\bar\varphi)$, it follows that the composition $(1/(\lambda+\varphi))F$ is a strict contraction on L^∞: $\|(1/(\lambda+\varphi))F\|_\infty\leqq\bar\varphi/(\lambda+\bar\varphi)<1$. Then J_λ as

[7] Note that the Laplace transform $\lambda\mapsto\hat x_\lambda$ of the solution $t\mapsto x_t$ corresponding to the initial data $\hat x$ is given in terms of the resolvent R_λ of A by $\hat x_\lambda=R_\lambda\hat x$, where $J_\lambda=\lambda R_\lambda$.

given by

$$J_\lambda = \left(\sum_{k=0}^{\infty}\left[\left(\frac{1}{\lambda+\varphi}\right)\Psi\right]^k\right)\frac{\lambda}{\lambda+\varphi}$$

is well defined in L^∞ and $x = J_\lambda y$ solves (***).

Again, since Ψ is positive, (A4) implies

$$\left(1+\frac{1}{\lambda}\varphi\right)|x| \leqq \|y\|_\infty + \frac{1}{\lambda}\varphi\|x\|_\infty.$$

Equivalently,

$$|x| \leqq \|y\|_\infty + \frac{1}{\lambda}\bar{\varphi}[\|x\|_\infty - |x|].$$

This, in turn, yields $\|x\|_\infty \leqq \|y\|_\infty$. Hence $\|J_\lambda y\|_\infty \leqq \|y\|_\infty$, for all $y \in L^\infty$ and $\|J_\lambda\|_\infty \leqq 1$. The strict positivity of J_λ in L^∞ follows from the strict positivity of $(1/(\lambda+\varphi))F$ and the series for J_λ. Now $J_\lambda 1^* = 1^*$ from (P3(iii)). Hence $\|J_\lambda\|_\infty \geqq 1$. Thus $\|J_\lambda\|_\infty = 1$ as claimed.

Observe that a simple inductive argument yields

$$\left(\frac{1}{\lambda+\varphi}\right)\left[F\left(\frac{1}{\lambda+\varphi}\right)I\right]^k = \left[\left(\frac{1}{\lambda+\varphi}\right)F\right]^k\left(\frac{1}{\lambda+\varphi}\right)I$$

for $k = 0, 1, 2, \cdots$. Hence the two series expressions for J_λ in L^1 and L^∞ are (formally) the same, and agree on $L^1 \cap L^\infty$.

(iii) *Case* $p = 2$. Since $(I-(1/\lambda)A)$ is symmetric and positive definite in L^2, its range is dense. Let $y \in L^2$ be in this range, and let $x \in L^2$ be such that (***) is satisfied. Then

$$\langle x, x\rangle = \langle y, x\rangle + \frac{1}{\lambda}\langle Ax, x\rangle.$$

From (P2(ii)) we have that $y \not\equiv 0$ implies

$$\langle x, x\rangle < \langle x, y\rangle,$$

and $y \equiv 0$ if and only if $x \equiv 0$. Hence, either $\|x\|_2 < \|y\|_2$ or $\|x\|_2 = \|y\|_2 = 0$. We conclude that J_λ exists in L^2, $\|J_\lambda y\| < \|y\|_2$ $(y \not\equiv 0)$ and $\|J_\lambda\|_2 \leqq 1$.

To see that J_λ is strictly positive on L^2, proceed as follows. Let $x \in L^2$ be written $x = x_1 + x_\infty$, where $x_1 \in L^1 \cap L^2$ and $x_\infty \in L^\infty \cap L^2$. Just take $x_1(\mu) = x(\mu)$ if $|x(\mu)| > 1$, $x_1(\mu) = 0$ if $|x(\mu)| < 1$, and $x_\infty = x - x_1$. Suppose $x \geqq 0$ so that $x_1, x_\infty \geqq 0$. By our previous arguments $J_\lambda x_1 \geqq 0$ in $L^1 \cap L^2$ and $J_\lambda x_\infty \geqq 0$ in $L^\infty \cap L^2$. Thus $J_\lambda x \geqq 0$ in L^2. If $x \geqq 0$ and $x \not\equiv 0$, then either $J_\lambda x_1 > 0$ or $J_\lambda x_\infty > 0$ so that $J_\lambda x > 0$. Hence J_λ is strictly positive. Alternatively, the strict positivity follows from the positivity as in the cases $p = 1, \infty$ due to the strict positivity of F.

To complete this case, assume that (A5) holds as well as (A1)–(A4). Then, for $\lambda > 0$,

$$\inf_{\langle x,x\rangle=1}\left\langle\left(I-\frac{1}{\lambda}A\right)x, x\right\rangle = 1, \quad \sup_{\langle x,x\rangle=1}\left\langle\left(I-\frac{1}{\lambda}A\right)x, x\right\rangle = 1+\frac{1}{\lambda}\|A\|_2.$$

It then follows that the spectrum of J_λ, as a symmetric operator in L^2, lies in the real interval $[1/(1+(1/\lambda)\|A\|_2), 1]$, and includes both endpoints. Hence, $\|J_\lambda\|_2 = 1$. (Note that $\|J_\lambda\|_2 = 1$ if and only if zero is in the spectrum of A.)

So far we have verified (P4) for $p = 1, 2, \infty$. The conclusion follows for all p, $1 \leqq p \leqq \infty$, by the Riesz convexity theorem. Since we already have $\|J_\lambda\|_1 = \|J_\lambda\|_\infty = 1$ on L^1 and L^∞, we need only verify that (***) has a measurable solution x for each $y \in L^p$, $1 < p < \infty$. Indeed, write $y = y_1 + y_\infty$ as before, where $y_1 \in L^1 \cap L^p$ and $y_\infty \in L^\infty \cap L^p$. Then $J_\lambda y_1 \in L^1$ and $J_\lambda y_\infty \in L^\infty$ are well defined by our previous results. Now define J_λ on L^p by $J_\lambda y = J_\lambda y_1 + J_\lambda y_\infty$ and observe that $x = J_\lambda y$ solves (***); moreover, $x = J_\lambda y \in L^1 + L^\infty$ and is surely measurable. Thus we see that, for $1 < p < \infty$, J_λ is a well-defined strictly positive operator on L_p, such that $\|J_\lambda\|_p \leqq 1$. Finally, since $p \to \|J_\lambda\|_p$ is a convex function for $1 \leqq p \leqq \infty$ and $\|J_\lambda\|_1 = \|J_\lambda\|_\infty = 1$, it follows that $\|J_\lambda\|_p = 1$ for *all* p if $\|J_\lambda\|_{\hat{p}} = 1$ for some $\hat{p}$, $1 < \hat{p} < \infty$. But in part (iii) of this proof we show that this is the case for $\hat{p} = 2$ whenever (A5) holds. This completes the proof of (P4). □

The results described above for $1 \leqq p \leqq \infty$ depended upon the Riesz convexity theorem. Thus (P4) is a sort of "little Riesz theorem" for L^p spaces whose underlying measure space is not finite (see [11]).

We now turn to the proofs of the main assertions (T1)–(T5).

Proof of (T1). This result is an immediate consequence of (P1) and Hille's first exponential formula [1]

$$T_t = \exp(tA) = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k. \qquad \square$$

Proof of (T2). This result is an immediate consequence of the exponential formula [1]

$$T_t = \lim_{n \to \infty} [J_{(t/n)}]^n$$

and the positivity of J_λ established in (P4). □

Proof of (T3). The assertions all follow by the exponential formula used in the proof of (T2). Indeed, T_t inherits directly all the properties of J_λ (except the *strict* positivity). □

If only (A1)–(A4) are assumed, for $t > 0$ the conclusion $\|T_t\|_p = 1$, $1 \leqq p \leqq \infty$, must be replaced by $\|T_t\|_p = 1$ for $p = 1, \infty$ and $\|T_t\|_p \leqq 1$ for $1 < p < \infty$. Note that (A5) was used only to guarantee that zero was in the spectrum of A as an operator in L^2.

Proof of (T4). We already have from (T3(ii)) and the semigroup property that $t \to \|T_t x\|_2$ is decreasing for every $x \neq 0$. To see that it is strictly decreasing we have, from the proof of (P2), that for $x_t = T_t \mathring{x}$, $t \geqq 0$,

$$\frac{d}{dt}\|x_y\|_2^2 = 2\langle Ax_t, x_t\rangle < 0,$$

provided $x_t \not\equiv 0$. Now if $x_{t_0} \equiv 0$ for some $t_0 > 0$, then $x_t \equiv 0$ for all $t \geqq t_0$. This cannot happen since the semigroup is analytic.

Next we show that $\lim_{t \to \infty} \|x_t\|_2 = 0$. Here we use the spectral theorem for symmetric operators in L^2: The operator $-A$ is a positive definite symmetric operator in L^2 (P2), and hence posseses a representation in terms of a resolution of the identity

$$-A = \int \lambda \, dE(\lambda),$$

where $\{E(\lambda): -\infty < \lambda < \infty\}$ is a family of projections in L^2, called the resolution of the identity for $-A$. For completeness we include its relevant properties:

(i) $\lambda \to E(\lambda)$ is of bounded variation and (normalized) left-continuous on $(-\infty, \infty)$.

(ii) λ_0 is an eigenvalue of $-A$ if and only if

$$E(\lambda_0) \neq \lim_{\lambda \downarrow \lambda_0} E(\lambda).$$

(iii) λ_0 is in the continuous spectrum of $-A$ if and only if $\lambda \to E(\lambda)$ is not constant in a neighborhood of λ_0 and

$$E(\lambda_0) = \lim_{\lambda \downarrow \lambda_0} E(\lambda).$$

(iv) $E(\lambda) = 0$, $\lambda \leqq 0$;
$E(\lambda) = 1$, $\lambda > \|A\|_2$.

(v) $E(\lambda)E(\mu) = E(\lambda)$ whenever $\lambda \leqq \mu$.

In terms of E we have the following well-known formula [1]:

$$x_t = T_t \mathring{x} = \int e^{-t\lambda}\, d[E(\lambda)\mathring{x}], \qquad t \geqq 0.$$

It then follows that $\lim_{t\to\infty} T_t\mathring{x}$ exists in L^2 and is given by

$$\lim_{t\to\infty} T_t\mathring{x} = \lim_{\lambda \downarrow 0} E(\lambda)\mathring{x} \equiv E(0^+)\mathring{x}.$$

But (P2) asserts that $\lambda = 0$ is *not* an eigenvalue of $-A$. Hence $E(0^+) = E(0) = 0$. This proves (T4). □

As noted above, (P2) asserts that (A1)–(A4) imply that $\lambda = 0$ is not an eigenvalue of $-A$. If, in addition, (A5) holds, then $\lambda = 0$ is in the continuous spectrum of $-A$ but is still not an eigenvalue. Thus the assumption of (A5) is not necessary for the validity of (T4).

Proof of (T5). To establish (T5), we appeal to the spectral formula used in the proof of (T4). From this formula we get

$$\|T_t\mathring{x}\|_2^2 = \int e^{-2\lambda t}\, d(\|E(\lambda)\mathring{x}\|_2^2),$$

which in turn implies, for $\mathring{x} \not\equiv 0$,

$$-\frac{d}{dt} \ln \|T_t\mathring{x}\|_2 = \frac{\int \lambda\, e^{-2\lambda t}\, d(\|E(\lambda)\mathring{x}\|_2^2)}{\int e^{-2\lambda t}\, d(\|E(\lambda)\mathring{x}\|_2^2)}.$$

We need the following result: There is at least one $\mathring{x} \geqq 0$ in L^2 such that $\|E(\lambda)\mathring{x}\|_2 > 0$ for all $\lambda > 0$. Deferring the proof of this assertion, which follows in the form of a lemma, suppose $\mathring{x} \geqq 0$ has been so chosen. Then, for every $\varepsilon > 0$,

$$\int_0^\varepsilon e^{2(\varepsilon-\lambda)t}\, d(\|E(\lambda)\mathring{x}\|_2^2) > 0;$$

moreover, this expression is unbounded as $t \to \infty$. Now by replacing $\int$ by $\int_0^\varepsilon + \int_\varepsilon^\infty$ in the above expression for $-(d/dt) \ln \|T_t\mathring{x}\|_2$, we see that

$$0 \leqq -\frac{d}{dt} \ln \|T_t\mathring{x}\|_2 \leqq \varepsilon + \frac{2\bar{\varphi}}{\int_0^\varepsilon e^{2(\varepsilon-\lambda)t}\, d(\|E(\lambda)\mathring{x}\|_2^2)}.$$

Hence

$$0 \leqq \overline{\lim_{t\to\infty}} -\frac{d}{dt} \ln \|T_t\mathring{x}\|_2 \leqq \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, (**) is established, and (T5) is proved. □

This proof depends strongly on the fact that $\lambda = 0$ is in the continuous spectrum of $-A$, which was guaranteed by the assumption (A5).

Now if (A5) does not hold, $\lambda = 0$ may not be in the spectrum of $-A$. Then

$$\inf_{\langle x,x\rangle=1} \langle -Ax, x\rangle = \lambda_0 > 0$$

and

$$\frac{d}{dt}\|T_t\mathring{x}\|_2 \leqq -\lambda_0\|T_t\mathring{x}\|_2$$

for *all* $\mathring{x} \in L^2$. This in turn implies that

$$\|T_t\mathring{x}\|_2 \leqq e^{-\lambda_0 t}\|\mathring{x}\|_2,$$

so that solutions decay exponentially fast for any initial $\mathring{x} \in L^2$. For the acoustic and optical cases we consider, this situation does not obtain.

LEMMA. *Let* $\{E(\lambda): \lambda > 0\}$ *be a family of nonzero projections in* L^2 *such that*

$$\lambda \leqq \mu \Rightarrow E(\lambda)E(\mu) = E(\lambda).$$

Then there exists an $\mathring{x} \geqq 0$ *in* L^2 *such that*

$$E(\lambda)\mathring{x} \neq 0 \quad \textit{for all } \lambda > 0.$$

Proof of the lemma. Assume the result is false; that is, for every $x \geqq 0$ in L^2 there is a $\hat{\lambda}(x) > 0$ such that $E(\hat{\lambda}(x))x = 0$. We proceed inductively to obtain a contradiction.

Fix $\lambda_0 > 0$ and choose $x_0 \geqq 0$ in L^2 so that $E(\lambda_0)x_0 \neq 0$. (Such a choice is always possible since $E(\lambda_0) \neq 0$ and every function in L^2 is the difference of two nonnegative functions.) Now $E(\hat{\lambda}(x_0))x_0 = 0$ and, necessarily, $\hat{\lambda}(x_0) < \lambda_0$. Define $\lambda_1 = \frac{1}{2}\hat{\lambda}(x_0)$, so that $\lambda_1 < \frac{1}{2}\lambda_0$. Next choose $x_1 \geqq 0$ in L^2 so that $E(\lambda_1)x_1 \neq 0$ and $\|x_1\|_2 < \frac{1}{2}\|E(\lambda_0)x_0\|_2$. Define $\lambda_2 = \frac{1}{2}\hat{\lambda}(x_1)$, so that $\lambda_2 < \frac{1}{2}\lambda_1$ and $E(\lambda_2)x_1 = 0$.

Continuing inductively we produce a sequence of functions $\{x_n\}$ in L^2 and a sequence of positive numbers $\{\lambda_n\}$ such that, for $n = 0, 1, 2, \cdots$,

(i) $x_n \geqq 0$ $(x_n \not\equiv 0)$;
(ii) $E(\lambda_n)x_n \not\equiv 0$;
(iii) $E(\lambda_{n+1})x_n = 0$;
(iv) $\|x_{n+1}\|_2 < \frac{1}{2}\|E(\lambda_n)x_n\|_2 < \frac{1}{2}\|x_n\|_2$;
(v) $\lambda_{n+1} = \frac{1}{2}\hat{\lambda}(x_n) < \frac{1}{2}\lambda_n$.

Define $\mathring{x} \geqq 0$ in L^2 by $\mathring{x} = \sum_{k=0}^{\infty} x_k$. Using (iv) above, obtain

$$\|\mathring{x}\|_2 \leqq \sum_{k=0}^{\infty} \|x_k\|_2 \leqq \sum_{k=0}^{\infty} (\tfrac{1}{2})^k\|x_0\|_2 \leqq 2\|x_0\|, \quad \text{so that } \mathring{x} \in L^2.$$

For arbitrary $n = 0, 1, 2, \cdots$, consider $E(\lambda_n)\mathring{x}$. From (ii) and (iii) above,

$$E(\lambda_n)\mathring{x} = \sum_{k=0}^{\infty} E(\lambda_n)x_k = \sum_{k=n}^{\infty} E(\lambda_n)x_k = E(\lambda_n)x_n + \sum_{l=1}^{\infty} E(\lambda_n)x_{n+l},$$

where $E(\lambda_n)x_n \neq 0$. But an inductive argument using (iv) implies

$$\left\| \sum_{l=1}^{\infty} E(\lambda_n)x_{n+l} \right\|_2 < \sum_{l=1}^{\infty} (\tfrac{1}{2})^l \|E(\lambda_n)x_n\| = \|E(\lambda_n)x_n\|_2.$$

Hence $\|E(\lambda_n)\mathring{x}\|_2 > 0$, for every $n = 0, 1, 2, \cdots$. Finally, since $\lambda_n \downarrow 0$ as $n \to \infty$, it follows

that $E(\lambda)\mathring{x} \not\equiv 0$ for all $\lambda > 0$. This contradicts our assumption, and the lemma is proved. □

3. Approximation of solutions. With a view toward approximating solutions to the original problem, we replace the kernel Ψ by $\Psi^{(M)} = \mathscr{X}_{M\times M}\Psi$, where $\mathscr{X}_{M\times M}$ is the characteristic function of the square $\{(\mu, \nu): |\mu|, |\nu| \leqq M\}$. The approximate problem thus obtained possesses all the features of the original problem except that its solution semigroup $\{T_t^{(M)} : t \geqq 0\}$ satisfies

$$\lim_{t\to\infty} T_t^{(M)} = P^{(M)},$$

where $P^{(M)}$ is the positive projection in L^2 given by

$$(P^{(M)}x)(\mu) = \begin{cases} \dfrac{1}{2M}\displaystyle\int_{-M}^{M} x(\nu)\, d\nu, & \mu \in [-M, M], \\ x(\mu), & \mu \notin [-M, M]. \end{cases}$$

For $\mathring{x} \in L^2$ the Trotter–Kato theorem guarantees that

$$T_t^{(M)}\mathring{x} \to T_t\mathring{x}$$

as $M \to \infty$ uniformly for $t \in [0, \alpha]$, $\alpha > 0$. Of course, $P^{(M)}\mathring{x} \to 0$ as $M \to \infty$.

Now if it happens that $\Psi^{(M)} = \inf_{|\nu|,|\mu|\leqq M} \Psi^{(M)}(\mu, \nu) > 0$, as it does in our typical example, then the decay rate is exponential. Indeed, *for every* $\mathring{x} \in L^2$,

$$\|T_t^{(M)}\mathring{x} - P^{(M)}\mathring{x}\|_2 \leqq e^{-2M\Psi^{(M)}t}\|\mathring{x} - P^{(M)}\mathring{x}\|_2.$$

Furthermore, if $\lim_{M\to\infty} 2M\Psi^{(M)} = 0$, this exponential rate becomes slower as the degree of approximation improves. This is the case for all convolution (optical) kernels as well as those for which $\varphi \in L^1$. Other approximation schemes also exhibit this phenomenon.

4. Concluding remarks. In view of the fact that the equation arises from a multiple scattering problem, we might expect a connection between its solution and stochastic processes. Indeed, for $t \geqq 0$, $\mu \in (-\infty, \infty)$ and any Borel set $E \subset (-\infty, \infty)$, define

$$\mathscr{P}(t, \mu, E) = (T_t\mathscr{X}_E)(\mu),$$

where $\mathscr{X}_E$ is the characteristic function of the set E. It is not too hard to show that $\mathscr{P}$ is a Markov process which is temporally homogeneous and spatially inhomogeneous (except in the optical case). Our original Cauchy problem thus corresponds to the Kolmogorov equation associated with the Markov process $\mathscr{P}$.

We conclude with a few observations. Our analysis depended in no essential way upon the boundedness of A. What is essential in our analysis is the averaging property, namely

$$\int (Ax)(\mu)\, d\mu = 0.$$

Furthermore, our analysis used symmetry and spectral theory rather heavily. If the operators are not symmetric or normal, similar results should be available by exploiting the positivity of the semigroup, or by using results like the Weyl–von Neumann theorem. If the kernel Ψ of A is merely nonnegative instead of positive, the nature of the set on which Ψ vanishes becomes critical. Analysis of the asymptotic behavior in this situation is much more delicate. Finally, generalization to include nonlinear hereditary effects also seems feasible, say by integrating against the solution semigroup.

Acknowledgment. The author is indebted to Professors V. J. Mizel, J. J. McCoy, M. J. Beran, M. Pinsky and G. C. Papanicolaou for many valuable conversations.

REFERENCES

[1] E. HILLE and R. PHILLIPS, *Functional Analysis and Semi-groups.* American Mathematical Society, Providence, RI, 1957.

[2] N. DUNFORD and J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1958.

[3] G. PAPANICOLAOU and J. B. KELLER, *Stochastic differential equation with applications to random harmonic oscillators and wave propagation in random media*, SIAM J. Appl. Math., 21 (1971), pp. 287–305.

[4] G. C. PAPANICOLAOU, *A kinetic theory for power transfer in stochastic systems*, J. Math. Phys., 13 (1972), pp. 1912–1918.

[5] I. M. BESIERIS and F. D. TAPPERT, *Kinetic equation for the quantized motion of a particle in a randomly perturbed potential field*, J. Math. Phys., 14 (1973), pp. 1829–1836.

[6] G. C. PAPANICOLAOU and W. KOHLER, *Power statistics for wave propagation in one dimension and comparison with radiative transport theory*, J. Math. Phys., 14 (1973), pp. 1733–1745.

[7] M. J. BERAN and J. J. MCCOY, *Propagation through an anisotropic random medium*, J. Math. Phys., 15 (1974), pp. 1901–1912.

[8] ———, *Propagation from a finite beam or source through an anisotropic random medium*, J. Acoust. Soc. Amer., 56 (1974), pp. 1667–1672.

[9] W. KOHLER and G. C. PAPANICOLAOU, *Power statistics for wave propagation in one dimension and comparison with radiative transport theory.* II, J. Math. Phys., 15 (1974), pp. 2186–2197.

[10] G. C. PAPANICOLAOU and W. KOHLER, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure and Appl. Math., 27 (1974), pp. 641–668.

[11] H. H. SCHAEFFER, *Banach Lattices and Positive Operators*, Springer-Verlag, New York, 1974.

[12] C. COFFMAN, D. DUFFIN and V. J. MIZEL, *Positivity of weak solutions of non-uniformly elliptic equations*, Annal. Matematica Pura Applic., 104 (1975), pp. 209–238.

[13] M. J. BERAN and J. J. MCCOY, *Propagation through anisotropic random medium. An integro-differential formulation*, J. Math. Phys., 17 (1976), pp. 1186–1189.

[14] W. KOHLER and G. G. PAPANICOLAOU, *Wave propagation in a randomly inhomogeneous ocean*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Physics 70, Joseph B. Keller and John S. Papadikis, eds., Springer-Verlag, New York, 1972.

[15] M. J. BERAN, M. J. LEITMAN and N. SCHWARTZ, *Scattering in the depth direction for an anisotropic random medium*, J. Math. Phys., 19 (1978), pp. 2121–2123.

[16] M. J. LEITMAN, *On plane waves propagating into a random fluid: Asymptotic behavior*, Proceedings of the International Symposium on Volterra Integral Equations, Helsinki, August 1978; Lecture Notes in Mathematics 737, Springer-Verlag, Berlin, 1979.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0008 $01.00/0

EIGENFUNCTION EXPANSIONS FOR A SINGULAR EIGENVALUE PROBLEM WITH EIGENPARAMETER IN THE BOUNDARY CONDITION*

DON HINTON†

Abstract. We study the singular boundary value problem

$$\tau(u)=\frac{1}{r}\{-(pu')'+qu\}=\lambda u, \qquad a\leqq x<b<\infty,$$

$$-\beta_1 u(a)+\beta_2(pu')(a)=\lambda[\beta_1' u(a)-\beta_2'(pu')(a)].$$

Conditions are placed on the coefficients which ensure the spectrum is bounded below and the essential spectrum is empty. An eigenfunction expansion theory is developed for a class $\mathcal{D}_1$ of functions. For this class the convergence is uniform and absolute on compact intervals. When the eigenvalues are all nonnegative, the class $\mathcal{D}_1$ is shown to be the domain of $A^{1/2}$, where A is the self-adjoint operator associated with the boundary value problem.

We consider here the singular eigenvalue problem

$$\tau(u)=\frac{1}{r}\{-(pu')'+qu\}=\lambda u, \qquad a\leqq x<b, \tag{1}$$

$$-[\beta_1 u(a)-\beta_2(pu')(a)]=\lambda[\beta_1' u(a)-\beta_2'(pu')(a)], \tag{2}$$

where we assume throughout that:

(i) The functions r, p and q are real continuous functions on the interval $[a, b)$ with r and p positive; further, p is assumed to be continuously differentiable.

(ii) The numbers β_1, β_2, β_1' and β_2' are real and satisfy

$$\rho=\beta_2'\beta_1-\beta_1'\beta_2>0. \tag{3}$$

The singularity at b may be finite or infinite.

This paper is a continuation of [9], where the regular problem was considered. As in [9], we obtain a convergence theory, uniform and absolute on compact intervals, for the eigenfunction expansions associated with (1)–(3). Again when the eigenvalues are nonnegative, the class of functions considered is the domain of the operator $A^{1/2}$, where A is the self-adjoint operator associated with (1)–(3).

Ultimately, we obtain our convergence theory along the same lines as [9]; however, it is first necessary to prove some technical lemmas which do not arise in the regular problem. These lemmas illustrate some of the difficulties associated with eigenfunction expressions arising from singular problems. When the eigenvalues are nonnegative, the convergence of the eigenfunction expansion is in the metric $\|F\|_A^2=\langle AF, F\rangle=\|A^{1/2}F\|^2$, i.e., the energy norm (see Theorem 3). Hence the convergence is somewhat stronger than indicated above. An alternative calculation of the metric is provided by the quadratic functional J_1.

The eigenvalue problem (1)–(3) arises in a large number of vibrational and heat conduction problems. The singular case represents a problem in which some physical dimension is effectively infinite. The case of cooling of a semi-infinite bar when the end is placed in contact with a finite amount of liquid is considered in [7]. It is assumed in [7]

* Received by the editors June 4, 1980, and in revised form October 17, 1980. This research was supported in part by the National Science Foundation under grant NSF MCS77-28268.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916.

that heat flows from the bar only into the liquid and is convected from the liquid to a surrounding medium; Newton's law of cooling is assumed at the liquid-solid interface. A numerical analysis of a singular problem with a singularity at 0 may be found in [8]. It considers the cooling of a cylindrical rod which is dropped into a container containing a finite amount of liquid. As a final example, consider the longitudinal displacement $y(x, t)$ of a semi-infinite bar with a mass m attached at $x = 0$. It is assumed that the cross-sectional area A of the bar is constant, but that the density $\delta(x)$ and modulus of elasticity $E(x)$ are variable. Then following the analysis of [2, p. 7] and [1, p. 250], the displacement satisfies

$$\delta(x)\frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x}\left(E(x)\frac{\partial y}{\partial x}\right), \qquad t > 0,$$

and the mass at $x = 0$ gives the condition

$$my_{tt}(0, t) = AE(0)y_x(0, t).$$

If we assume appropriate initial conditions, then a separation of variables leads to the eigenvalue problem (the eigenparameter used is $-\lambda$)

$$-\frac{d}{dx}\left(E(x)\frac{dX}{dx}\right) = \lambda\delta(x)X, \qquad -m^{-1}AE(0)X'(0) = \lambda X(0),$$

which is of the type (1)–(3) with $\beta_1 = 0$, $\beta_2 = -AE(0)/m$, $\beta_1' = 1$ and $\beta_2' = 0$.

Additional applications, extensive references, and a general discussion of the literature of boundary value problems with eigenparameter in the boundary condition may be found in [6], [7], [16].

Recent work of Fulton [7] also treats singular problems of the kind (1)–(3). In [7], the left endpoint a is assumed to be either regular or singular of limit-circle type; the right endpoint b is assumed to be either limit circle or limit point type. We consider here a to be regular and b to be singular in the limit point case. The analysis of [7] is divided into two cases: b is limit circle and b is limit point. In the first case the spectrum is necessarily discrete and a detailed eigenfunction expansion theorem is obtained. The part of this theorem which relates to our work says that if $F = (F_1, F_2)$ is in the domain of the self-adjoint operator A associated with the eigenvalue problem, then the series for F_1 converges uniformly and absolutely on compact intervals; further, the series for F_1 may be termwise differentiated, with uniform and absolute convergence also holding on compact intervals. In the limit point case of [7], an analogous expansion theorem is obtained, but the series is replaced by an integral because of the possible presence of a continuous spectrum. In this case a more prominent role is played by the spectral function $\rho(\lambda)$ and Titchmarsh–Weyl m-coefficient $m(\lambda)$.

We consider here only the case where τ has an empty essential spectrum and show that this leads to the absence also of an essential spectrum in (1)–(3). As compared to [7] we obtain uniform and absolute convergence on compact sets for functions which are in the domain of $A^{1/2}$. The eigenfunction series may be termwise differentiated with the convergence of the differentiated series being in an $\mathscr{L}^2$ space. Thus we obtain uniform and absolute convergence on compact sets for a larger class of functions than in [7], but with weaker convergence of the differentiated series. Other parts of the expansion theorems of [7] deal with series or integrals that we do not consider.

For a Hilbert space formulation of (1)–(3), we follow [6], [7], [9], which uses a two component Hilbert space to realize the operator-theoretic formulation given by Walter [16].

We define the domain Δ of the operator τ by

$$\Delta = \{f \in \mathscr{L}_r^2(a, b): f, pf' \in \mathrm{AC}_{\mathrm{loc}}, \tau(f) \in \mathscr{L}_r^2(a, b)\}.$$

$\mathrm{AC}_{\mathrm{loc}}$ denotes locally absolutely continuous, and $\mathscr{L}_r^2(a, b)$ is the complex Hilbert space of Lebesgue measurable functions f satisfying $\int_a^b r|f|^2 < \infty$. In the terminology of differential operator theory, τ is a maximal operator [12, p. 10].

The Hilbert space H is defined by $H = \mathscr{L}_r^2(a, b) \oplus C$, where C is the complex numbers. The inner product in H is given by

$$\langle (F_1, F_2), (G_1, G_2) \rangle = \int_a^b r F_1 \bar{G}_1 + \frac{1}{\rho} F_2 \bar{G}_2.$$

Let $D(A)$ be the set of all $(F_1, F_2) \in H$ which satisfy

(i) $F_1 \in \Delta$,

(ii) $F_2 = \beta_1' F_1(a) - \beta_2'(pF_1')(a)$,

and define $A: D(A) \to H$ by

$$A(F_1, F_2) = (\tau(F_1), -\beta_1 F_1(a) + \beta_2(pF_1')(a)).$$

Thus $F_1 \in \Delta$ satisfies (1)–(2) if and only if

$$F = (F_1, \beta_1' F_1(a) - \beta_2'(pF_1')(a)) \in D(A) \quad \text{and} \quad AF = \lambda F.$$

Clearly $D(A)$ is dense in H. The fact that A is a closed operator follows from the closure of τ[12, p. 15], [14, § 17.4] (these references consider only weights $r(x) = 1$; their arguments apply to general weight functions), and the fact that $f_n \to f$ and $\tau f_n \to \tau f$ imply that $\{f_n\}$ and $\{f_n'\}$ converge uniformly to f and f', respectively, on compact sets [3, p. 1296].

Recall that τ is said to be in the *limit-point* case if

$$1 = \dim \{f \in \mathscr{L}_r^2(a, b): \tau(f) = \lambda f\},$$

where $\operatorname{Im} \lambda \neq 0$. It is a result of Weyl that this dimension is independent of λ for $\operatorname{Im} \lambda \neq 0$.

LEMMA 1. *The operator A is symmetric if and only if τ is limit-point.*

Proof. Defining

$$\langle (F_1, F_2), (G_1, G_2) \rangle_x = \int_a^x r F_1 \bar{G}_1 + \frac{1}{\rho} F_2 \bar{G}_2,$$

we have after some calculation that for $F = (F_1, F_2)$, $G = (G_1, G_2) \in D(A)$,

$$\begin{aligned} \langle F, AG \rangle_x - \langle AF, G \rangle_x = {} & -F_1 p \bar{G}_1' + pF_1' \bar{G}_1 \Big|_a^x \\ & + \frac{1}{\rho}[\beta_1' F_1(a) - \beta_2'(pF_1')(a)][-\beta_1 \overline{G_1(a)} + \beta_2(\overline{pG_1'})(a)] \\ & - \frac{1}{\rho}[-\beta_1 F_1(a) + \beta_2(pF_1')(a)][\beta_1' \overline{G_1(a)} - \beta_2'(\overline{pG_1'})(a)] \\ = {} & [-F_1 p \bar{G}_1' + pF_1' \bar{G}_1](x). \end{aligned}$$

Hence A is symmetric if and only if

$$\lim_{x \to b} [-F_1 p \bar{G}_1' + pF_1' \bar{G}_1](x) = 0, \qquad F_1, G_1 \in \Delta. \tag{4}$$

However, this is equivalent to τ being limit-point [12, p. 19], [14, § 18.3, 4].

THEOREM 1. *The operator A is self-adjoint if and only if τ is limit-point.*

Proof. By Lemma 1, we need only show that τ being limit-point implies A is self-adjoint. Since A is already symmetric, the proof will be complete if we show that each of $A+iI$, $A-iI$ has range H[14, p. 33]. We consider the operator $A+iI$; a similar argument applies to $A-iI$.

Let $G=(G_1, G_2)\in H$. To solve $(A+iI)(F)=G$, we must show there is an $F_1\in\Delta$ such that

$$\hat{\tau}(F_1)=G_1 \quad \text{and} \quad k(F_1)=G_2, \tag{5}$$

where $\hat{\tau}=\tau+iI$ and

$$k(F_1)=-\beta_1F_1(a)+\beta_2(pF_1')(a)+i[\beta_1'F_1(a)-\beta_2'(pF_1')(a)].$$

Since τ is a maximal operator, there is an $\hat{F}_1\in\Delta$ such that $\hat{\tau}(\hat{F}_1)=G_1$ [12, p. 15]. Also, τ being limit-point implies that there is a $\phi\in\Delta$ such that

$$\tau(\phi)=-i\phi \quad \text{and} \quad \int_a^b r|\phi|^2=1.$$

Equation (4) ensures that

$$\lim_{x\to b}[-\phi p\bar{\phi}'+p\phi'\bar{\phi}](x)=0. \tag{6}$$

We assume without loss of generality that $\phi'(a)$ is real.

For $F_1=\hat{F}_1+c\phi$, we have $\hat{\tau}(F_1)=\hat{\tau}(\hat{F}_1)=G_1$ and $k(F_1)=k(\hat{F}_1)+ck(\phi)$; hence if $k(\phi)\neq 0$, (5) will hold for $c=[G_2-k(\hat{F}_1)]/k(\phi)$ and the proof will be complete. If $k(\phi)=0$, then

$$\phi(a)=\frac{\beta_2-i\beta_2'}{\beta_1-i\beta_1'}(p\phi')(a)=\frac{\beta_1\beta_2+\beta_1'\beta_2'-i\rho}{\beta_1^2+(\beta_1')^2}(p\phi')(a). \tag{7}$$

From $\tau(\phi)\bar{\phi}=-i\phi\bar{\phi}$ and an integration by parts, we have

$$-i\int_a^x r|\phi|^2=-p\phi'\bar{\phi}\Big|_a^x+\int_a^x p|\phi'|^2+q|\phi|^2. \tag{8}$$

Taking the imaginary part of (8) and applying (6) and (7), we obtain

$$-\int_a^b r|\phi|^2=\operatorname{Im}[p\phi'\bar{\phi}](a)=[p(a)\phi'(a)]^2\frac{\rho}{\beta_1^2+(\beta_1')^2}.$$

Since $\rho>0$, this is a contradiction and the proof is complete. □

Theorem 1 may also be deduced from the results of [7]. We have included a short proof for completeness.

To further relate A to τ we recall the definition of essential spectrum [3, p. 1393]. A complex number λ is said to be in the essential spectrum of a closed operator T if $T-\lambda I$ does not have closed range. For a symmetric operator T which does not have eigenvalues of infinite multiplicity, this is equivalent to the condition [3, p. 1435] that there be a sequence $\{f_n\}$ in the domain of T such that (i) $\|f_n\|=1$; (ii) $\{f_n\}$ does not contain a convergent subsequence; and (iii) $\|Tf_n-\lambda f_n\|\to 0$ as $n\to\infty$.

THEOREM 2. *The essential spectrum of τ equals the essential spectrum of A.*

Proof. Suppose λ is in the essential spectrum of τ. Let τ_0 be the minimal operator associated with τ, i.e., the closure of τ in $\mathscr{L}_r^2(a, b)$ when restricted to those $f \in \Delta$ which have compact support in (a, b). Since the essential spectrum of τ equals that of τ_0 [3, p. 1394], we have then a sequence $\{f_n\}$ in the domain of τ_0 such that $\|f_n\| = 1$, $\{f_n\}$ has no convergent subsequence, and $(\tau - \lambda)f_n \to 0$ as $n \to \infty$. Since $f_n(a) = f_n'(a) = 0$ [12, p. 11], the sequence $F_n = (f_n, 0)$ is in domain A, $\|F_n\| = \|f_n\| = 1$, $\{F_n\}$ has no convergence subsequence, and $(A - \lambda)F_n = ((\tau - \lambda)f_n, 0) \to 0$ as $n \to \infty$. Thus λ is in the essential spectrum of A.

Now let λ be in the essential spectrum of A and $F_n = (f_n, \beta_1' f_n(a) - \beta_2'(pf_n')(a))$ be a sequence in $D(A)$ with $\|F_n\| = 1$; $\{F_n\}$ contains no convergent subsequence, and $(A - \lambda)F_n \to 0$ as $n \to \infty$. From $\|F_n\| = 1$ we have

$$|\beta_1' f_n(a) - \beta_2'(pf_n')(a)| \leqq \sqrt{\rho}, \tag{9}$$

and from $(A - \lambda)F_n \to 0$ as $n \to \infty$ we have, as $n \to \infty$,

$$-\beta_1 f_n(a) + \beta_2(pf_n')(a) - \lambda[\beta_1' f_n(a) - \beta_2'(pf_n')(a)] \to 0. \tag{10}$$

If we set

$$a_n = \beta_1' f_n(a) - \beta_2'(pf_n')(a), \qquad b_n = -\beta_1 f_n(a) + \beta_2(pf_n')(a),$$

we have

$$(pf_n')(a) = -\rho^{-1}(a_n\beta_1 + b_n\beta_1'), \qquad f_n(a) = -\rho^{-1}(a_n\beta_2 + b_n\beta_2');$$

hence (9) and (10) imply that $(f_n(a), (pf_n')(a))$ is a bounded sequence. It is sufficient to suppose it converges. This implies then that $\{f_n\}$ contains no convergent subsequence in $\mathscr{L}_r^2(a, b)$, since F_n contains no convergent subsequence in H. Thus $\{f_n\}$ satisfies $\|f_n\| \leqq 1$, $\{f_n\}$ contains no convergent subsequence, and $(\tau - \lambda)f_n \to 0$. Hence the sequence $\{f_n/\|f_n\|\}$ satisfies the requirements for λ being in the essential spectrum of τ.

To develop an eigenfunction expansion theory we consider the following conditions:

(P)
- (i) There is a number B such that $q(x) \geqq Br(x)$ for $a \leqq x < b$.
- (ii) τ is in the limit-point case.
- (iii) The spectrum of A is purely discrete.

For many weight functions r, P(ii) is implied by P(i); in particular this is true for $r(x) = 1$ and $[a, b) = [a, \infty)$ [12, p. 23]. However, since P(i) does not imply P(ii) in general, we include it as a hypothesis. Also we note that when (ii) holds the operator A is self-adjoint; hence P(iii) is equivalent to the essential spectrum of A being empty (note that eigenvalues of A are of multiplicity at most two). Actually, in the limit-point case, $\tau(u) = \lambda u$ has at most one linearly independent $\mathscr{L}_r^2(a, b)$ solution for each λ; hence eigenvalues of A are then of multiplicity one.

As a consequence of Theorem 2 and Lemma 3 below, it follows that (under P(i)–P(ii)) a criterion which implies that the spectrum of a self-adjoint extension of τ_0 is purely discrete and bounded below is also such a criterion for A. One of the best known such conditions is that of Friedrich [5],

$$\lim_{x \to b} \frac{1}{r(x)}\left[q(x) + \frac{1}{4p(x)h(x)^2}\right] = \infty,$$

where

$$h(x)=\begin{cases}\displaystyle\int_a^x \frac{1}{p} & \text{if } \displaystyle\int_a^b \frac{1}{p}=\infty,\\ \displaystyle\int_x^b \frac{1}{p} & \text{if } \displaystyle\int_a^b \frac{1}{p}<\infty.\end{cases}$$

Other criteria which imply that the spectrum of self-adjoint extensions of τ_0 is bounded below and purely discrete may be found in [11], [13], [15] and the references contained therein.

The set $\mathscr{D}$ is defined as all those $(F_1, F_2)\in H$ which satisfy:

(i) F_1 is locally absolutely continuous with $\int_a^b[p|F_1'|^2+q|F_1|^2]<\infty$.

(ii) If $\beta_2'=0$, then $F_2=\beta_1'F_1(a)$, and if $\beta_2'\neq 0$, then F_1 is differentiable at a and $F_2=\beta_1'F_1(a)-\beta_2'(pF_1')(a)$.

The function J is defined on $\mathscr{D}\times\mathscr{D}$ by

$\beta_2'=0$:

$$J((F_1,F_2),(G_1,G_2))=\int_a^b [F_1'\bar{G}_1'+qF_1\bar{G}_1]-\frac{1}{\rho}\beta_1'\beta_1F_1(a)\overline{G_1(a)}.$$

$\beta_2'\neq 0$:

$$\begin{aligned}J((F_1,F_2),(G_1,G_2))&=\int_a^b [pF_1'\bar{G}_1'+qF_1\bar{G}_1]\\ &\quad-\frac{1}{\rho}[\beta_1\beta_1'F_1(a)\overline{G_1(a)}-\beta_1\beta_2'F(a)\overline{(pG_1')}(a)\\ &\qquad-\beta_1\beta_2'(pF_1')(a)\overline{G_1(a)}+\beta_2\beta_2'(pF_1')(a)\overline{(pG_1')(a)}]\end{aligned}$$

Note that P(i) implies that the convergence in (i) above is absolute.

For the singular problem we use certain approximation properties which follow from [10, Lemma 1]. In the notation of [10], the boundary condition used is $y'(a)=0$; hence $A_1=0$, $A_2=1$.

LEMMA 2. *If p, q and r are as in* (1) *and* (P)(i)–(ii) *hold, then*:

(i) $D(A)\subset\mathscr{D}$.

(ii) *If $f\in\mathscr{L}_r^2(a,b)$ is locally absolutely continuous and satisfies*

$$\int_a^b p|f'|^2+q|f|^2<\infty \tag{11}$$

and $g\in\Delta$, then $(fp\bar{g}')(x)\to 0$ as $x\to b$.

(iii) *If f is as in* (ii) *and $\varepsilon>0$, then there is an $f_\varepsilon\in\Delta$ with compact support such that $f_\varepsilon(a)=f(a)$ and*

$$\int_a^b \{p|f'-f_\varepsilon'|^2+[|q|+r]|f-f_\varepsilon|^2\}<\varepsilon. \tag{12}$$

Proof. From [8, Lemma 1], τ satisfies the Dirichlet condition; hence $D(A)\subset\mathscr{D}$. From [10, Lemma 1, (iii)],

$$\int_a^b rf\overline{\tau(g)}=\int_a^b [pf'\bar{g}'+qf\bar{g}] \tag{13}$$

for all f satisfying (11) and all $g \in \Delta$ with $g'(a) = 0$. An integration by parts of the left-hand side of (13) yields that for such f and g, $(fp\bar{g}')(x) \to 0$ as $x \to b$. Since this limit does not depend on the behavior of g at a, we may conclude that the limit holds for all $g \in \Delta$. Finally, part (iii) above is a restatement of part (iv) of [10, Lemma 1]. □

We note also that in part (iii) above we may apply a standard argument to see that $f'_\varepsilon(a)$ may be arbitrarily specified as well. It will be convenient to do this below.

LEMMA 3. *Assume* P(i)–(ii). *Then*:

(i) $J(F, G) = \langle F, AG\rangle$ *for* $F \in \mathcal{D}$ *and* $G \in D(A)$.

(ii) A *is bounded below.*

(iii) *For* $F \in \mathcal{D}$, $J(F, F) \geqq b\langle F, F\rangle$ *where* b *is a lower bound for the spectrum of* A.

Proof. For $F = (F_1, F_2) \in \mathcal{D}$ and $G = (G_1, G_2) \in D(A)$, an integration by parts and application of part (ii) of Lemma 2 gives

$$\begin{aligned}\langle F, AG\rangle &= (F_1 p\bar{G}'_1)(a) + \int_a^b [pF'_1\bar{G}'_1 + qF_1\bar{G}_1] \\ &\quad + \frac{1}{\rho}[\beta'_1 F_1(a) - \beta'_2(pF'_1)(a)][-\beta_1\bar{G}_1(a) + \beta_2(p\bar{G}'_2)(a)] \\ &= \int_a^b [p\bar{F}'_1\bar{G}_1 + qF_1\bar{G}_1] \\ &\quad - \frac{1}{\rho}[\beta'_1\beta_1 F_1(a)\bar{G}_1(a) - \beta_1\beta'_2 F_1(a)(p\bar{G}'_1)(a) \\ &\qquad - \beta_1\beta'_2(pF'_1)\bar{G}_1(a) + \beta_2\beta'_2(pF_1)(a)(p\bar{G}'_1)(a)] \\ &= J(F, G).\end{aligned}$$

To show that A is bounded below, we must prove the existence of a number k such that for all $G = (G_1, G_2) \in D(A)$,

$$\begin{aligned}\langle G, AG\rangle = J(G, G) &\geqq k\langle G, G\rangle \\ &= k\left[\int_a^b r|G_1|^2 + \frac{1}{\rho}|\beta'_1 G_1(a) - \beta'_2(pG'_1)(a)|^2\right];\end{aligned}$$

this inequality may be written as

$$\begin{aligned}&\int_a^b p|G'_1|^2 + (q - kr)|G_1|^2 + \alpha|G_1(a)|^2 \\ &\qquad + \beta[G_1(a)(p\bar{G}'_1)(a) + \bar{G}_1(a)(pG'_1)(a)] + \gamma|(pG'_1)(a)|^2 \geqq 0,\end{aligned} \tag{14}$$

where

$$\alpha = \frac{1}{\rho}[-k(\beta'_1)^2 - \beta_1\beta'_1], \quad \beta = \frac{1}{\rho}[k\beta'_1\beta'_2 + \beta_1\beta'_2], \quad \gamma = \frac{1}{\rho}[-k(\beta'_2)^2 - \beta_2\beta'_2].$$

If $\beta'_2 = 0$, then we easily ensure that (14) holds by making $\alpha > 0$ ($\beta'_1 \neq 0$ if $\beta'_2 = 0$) and $-k \geqq B$, where B is as in P(ii). If $\beta'_2 \neq 0$, then we have, for $\delta > 0$,

$$\begin{aligned}(\alpha + \delta\rho^{-1})\gamma - \beta^2 &= (-k)[\delta(\beta'_2)^2 - \beta'_1\beta'_2(\beta'_1\beta_2 - \beta_1\beta'_2)]\rho^{-2} \\ &\quad + [\beta_1\beta'_2(\beta_2\beta'_1 - \beta_1\beta'_2) - \delta\beta_2\beta'_2]\rho^{-2}.\end{aligned}$$

If $\delta > 0$ is chosen so that

$$\delta(\beta'_2)^2 - \beta'_1\beta'_2(\beta'_1\beta_2 - \beta_1\beta'_2) > 0,$$

there is a k_0 such that for $k \leqq k_0$

$$\alpha + \delta\rho^{-1} > 0 \quad \text{and} \quad (\alpha + \delta\rho^{-1})\gamma - \beta^2 > 0.$$

For such k we have then that

$$(\alpha + \delta\rho^{-1})|G_1(a)|^2 + \beta[G_1(a)(p\bar{G}_1')(a) + \bar{G}_1(a)(pG_1')(a)] + \gamma|(pG_1')(a)|^2 \geqq 0.$$

Then the left-hand side of (14) is bounded below by

$$\int_a^b p|G_1'|^2 + (q - kr)|G_1|^2 - \delta\rho^{-1}|G_1(a)|^2. \tag{15}$$

It only remains to show that (15) is nonnegative for $-k$ sufficiently large. From [7, p. 39] (we assume without loss of generality that $b > a + 1$),

$$|G_1(a)|^2 \leqq \int_a^{a+1} |G_1|^2 + 2\int_a^{a+1} |G_1'\bar{G}_1|.$$

Now for $\varepsilon > 0$, $2|G_1'\bar{G}_1| \leqq \varepsilon^{-2}|G_1|^2 + \varepsilon^2|G_1'|^2$; hence

$$\begin{aligned} |G_1(a)|^2 &\leqq (1 + \varepsilon^{-2}) \int_a^{a+1} |G_1|^2 + \varepsilon^2 \int_a^{a+1} |G_1'|^2 \\ &\leqq (1 + \varepsilon^{-2})c_1 \int_a^b r|G_1|^2 + \varepsilon^2 c_2 \int_a^b p|G_1'|^2, \end{aligned}$$

where $c_1 = \max 1/r(x)$ on $[a, a+1]$ and $c_2 = \max 1/p(x)$ on $[a, a+1]$. Taking $\varepsilon^2 = (\rho/2)\delta c_2$ and $-k \geqq B + c_1(1 + \varepsilon^{-2})\delta/\rho$ completes the proof that (15) is nonnegative; hence (14) holds.

To establish (iii), for each $\varepsilon > 0$ let f_ε satisfy (12) with $f = F_1$ where $F = (F_1, F_2)$. Further let $f_\varepsilon'(a) = F_1'(a)$ if F_1 is differentiable at a and $f_\varepsilon'(a) = 0$ otherwise. Define

$$F_\varepsilon = (f_\varepsilon, \beta_1' f_\varepsilon(a) - \beta_2'(pf_\varepsilon')(a)) \in D(A), \tag{15'}$$

and note that, by Lemma 2,

$$J(F, F) = \lim_{\varepsilon \to 0} J(F_\varepsilon, F_\varepsilon) \quad \text{and} \quad \langle F, F \rangle = \lim_{\varepsilon \to 0} \langle F_\varepsilon, F_\varepsilon \rangle.$$

However, by parts (i) and (ii),

$$J(F_\varepsilon, F_\varepsilon) = \langle F_\varepsilon, AF_\varepsilon \rangle \geqq b\langle F_\varepsilon, F_\varepsilon \rangle,$$

where b is a lower bound for A. Part (iii) now follows by letting $\varepsilon \to 0$. □

The above proof shows that the boundedness below of A is somewhat more complicated than for the Sturm–Liouville case. We might expect that A would be bounded below if τ_0 is; however this is not clear.

When P(i) − P(iii) holds, the spectrum of A consists of eigenvalues $\lambda_1 < \lambda_2 < \ldots$ with corresponding eigenfunctions $\{\Gamma_n\}$ with $\Gamma_n = (\Gamma_{n1}, \Gamma_{n2})$. We use this notation for the remainder of the paper.

LEMMA 4. *Assume* P(i)–P(iii). *If* $F \in \mathcal{D}$, *then*

$$\sum_{n=1}^{\infty} \lambda_n |\langle F, \Gamma_n \rangle|^2 \leqq J(F, F).$$

Proof. The proof is identical to that of [9, Lemma 2]. □

We now construct the resolvent kernel for A in the limit-point case. For λ not in the spectrum of A, define ϕ_λ and χ_λ to be the solutions of $\tau(u)=\lambda u$ which satisfy the conditions

$$\chi_\lambda(a)=\beta_2+\beta_2'\lambda, \quad (p\chi_\lambda')(a)=\beta_1+\beta_1'\lambda, \quad \int_a^b |\phi_\lambda|^2=1.$$

The function ϕ_λ is unique up to multiples $c\phi_\lambda$ with $|c|=1$. The function χ_λ satisfies the boundary condition (2), and ϕ_λ does not satisfy the boundary condition (2) since λ is not an eigenvalue. Hence χ_λ and ϕ_λ are linearly independent. The Wronskian

$$\begin{aligned} w(\lambda) &= -\phi_\lambda(x)(p\chi_\lambda')(x)+(p\phi_\lambda')(x)\chi_\lambda(x) \\ &= -\phi_\lambda(a)(\beta_1+\beta_1'\lambda)+(p\phi_\lambda')(a)(\beta_2+\beta_2'\lambda) \end{aligned}$$

is therefore independent of x. To solve $(\lambda-A)F=(G_1, G_2)$ with $F=(F_1, F_2)$ requires that

$$(\lambda-\tau)F_1=G_1, \tag{16}$$

$$\lambda[\beta_1'F_1(a)-\beta_2'(pF_1')(a)]-[-\beta_1F_1(a)+\beta_2(pF_1')(a)]=G_2. \tag{17}$$

The variation-of-constants formula applied to (16) yields

$$F_1(x)=c_1\phi_\lambda(x)+c_2\chi_\lambda(x)+\int_a^x \frac{\phi_\lambda(x)\chi_\lambda(t)-\chi_\lambda(x)\phi_\lambda(t)}{w(\lambda)} r(t)G_1(t)\,dt. \tag{18}$$

Calculation of $F_1(a)$ and $F_1'(a)$ from (18) and substitution into (17) yields that $c_1=-G_2/w(\lambda)$. Since we are in the limit-point case, (4) implies that

$$\lim_{x\to b}[F_1(x)(p\phi_\lambda')(x)-(pF_1')(x)\phi_\lambda(x)]=0. \tag{19}$$

Calculation of $F_1(x)$ and $F_1'(x)$ from (18) and substitution into (19) yields that

$$c_2=\frac{1}{w(\lambda)}\int_a^b \phi_\lambda(t)G_1(t)r(t)\,dt.$$

The above values of c_1 and c_2 allow us to write (18) in the form

$$F_1(x)=-\frac{G_2\phi_\lambda(x)}{w(\lambda)}+\int_a^b G(x, t, \lambda)r(t)G_1(t)\,dt, \tag{20}$$

where

$$G(x, t, \lambda)=\begin{cases} \dfrac{\phi_\lambda(x)\chi_\lambda(t)}{w(\lambda)}, & a\leqq t\leqq x, \\ \dfrac{\chi_\lambda(x)\phi_\lambda(t)}{w(\lambda)}, & a<x<t. \end{cases}$$

Note that when G is real, (20) can be written as

$$F_1(x)=\langle \tilde{G}(x, \cdot, \lambda), G\rangle,$$

where

$$\tilde{G}(x, t, \lambda)=\left(G(x, t, \lambda), -\frac{\rho\phi_\lambda(x)}{w(\lambda)}\right).$$

In the formula of [9] which corresponds to (20), i.e., [9, (4), p. 34], we inadvertently omitted the $r(t)$ in the integral and the plus sign before the integral.

The remainder of our analysis proceeds in a manner similar to that of [9]. For $a < x$, a short calculation gives $\tilde{G}(x, \cdot, \lambda) \in \mathscr{D}$; moreover, $(\lambda - A)\Gamma_n = (\lambda - \lambda_n)\Gamma_n$ implies by (20) that

$$\Gamma_{n1}(x) = (\lambda - \lambda_n)\langle \tilde{G}(x, \cdot, \lambda), \Gamma_n \rangle, \qquad a \leqq x. \tag{21}$$

Hence, by Lemma 4,

$$\begin{aligned} \sum_{n=1}^{\infty} \lambda_n \left| \frac{\Gamma_{n1}(x)}{\lambda - \lambda_n} \right|^2 &= \sum_{n=1}^{\infty} \lambda_n |\langle \tilde{G}(x, \cdot, \lambda), \Gamma_n \rangle|^2 \\ &\leqq B_1(x, \lambda), \end{aligned} \tag{22}$$

where

$$B_1(x, \lambda) = J(\tilde{G}(x, \cdot, \lambda), \tilde{G}(x, \cdot, \lambda)).$$

The definition of J shows that $B_1(\cdot, \lambda)$ is uniformly bounded on compact sets of $[a, b)$.

Writing (21) in the form of (20), we see that the form of the function G yields th

$$(\lambda - \lambda_n)^{-1}\Gamma'_{n1}(x) = \frac{d}{dx}\langle \tilde{G}(x, \cdot, \lambda), \Gamma_n \rangle = \langle \tilde{G}_x(x, \cdot, \lambda), \Gamma_n \rangle,$$

where $\tilde{G}_x(x, \cdot, \lambda) = (G_x(x, \cdot, \lambda), -\rho\phi'_\lambda(x)/w(\lambda))$. Since $\tilde{G}_x(x, \cdot, \lambda) \in H$ for Bessel's equality gives

$$\sum_{n=1}^{\infty} \left| \frac{\Gamma'_{n1}(x)}{\lambda - \lambda_n} \right|^2 = \langle \tilde{G}_x(x, \cdot, \lambda), \tilde{G}_x(x, \cdot, \lambda) \rangle. \tag{23}$$

Defining $B_2(x, \lambda) = \langle \tilde{G}_x(x, \cdot, \lambda), \tilde{G}_x(x, \cdot, \lambda) \rangle$, we see from the definition of $\tilde{G}$ $B_2(\cdot, \lambda)$ is uniformly bounded on compact subsets of $[a, b)$. If $F \in H$, then we ado notation

$$c(F) = p(a) \sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle \Gamma'_{n1}(a)$$

if the series converges. The set $\mathscr{D}_1$ is defined by $\mathscr{D}_1 = \mathscr{D}$ if $\beta'_2 = 0$, and if $\beta'_2 \neq 0$, $\mathscr{D}_1$ set of all $F = (F_1, F_2) \in H$ such that:

(i) F_1 is locally absolutely continuous with

$$\int_a^b p|F'_1|^2 + q|F_1|^2 < \infty.$$

(ii) $\sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle \Gamma'_{n1}(a)$ converges and $F_2 = \beta'_1 F_1(b) - \beta'_2 c(F)$. Define J_1 on $\mathscr{D}$ by $J_1 = J$ if $\beta'_2 = 0$, and otherwise by·

$$\begin{aligned} J((F_1, F_2), (G_1, G_2)) = &\int_a^b [pF'_1\bar{G}'_1 + qF_1\bar{G}_1] \\ &- \frac{1}{\rho}[\beta_1\beta'_1 F_1(a)\overline{G_1(a)} \\ &- \beta_1\beta'_2 F_1(a)\overline{c(G)} - \beta_1\beta'_2 c(F)\overline{G_1(a)} + \beta_2\beta'_2 c(F)\overline{c(G} \end{aligned}$$

It will follow from Theorem 3 that if $\beta'_2 \neq 0$, $\mathscr{D} \subset \mathscr{D}_1$ and J_1 is an extension of

For $F = (F_1, F_2) \in H$, the completeness of the eigenvectors gives

$$F_2 = \sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle [\beta'_1 \Gamma_{n1}(a) - \beta'_2 (p\Gamma'_{n1})(a)]. \tag{24}$$

Thus for $\beta_2' \neq 0$ the series $\sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle \Gamma_{n1}'(a)$ is convergent if $\sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle \Gamma_{n1}(a)$ is convergent to $F_1(a)$, in which case

$$\beta_1' F_1(a) - \beta_2' c(F) = \sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle [\beta_1' \Gamma_{n1}(a) - \beta_2' (p\Gamma_{n1}')(a)] = F_2$$

by (24). Hence (ii) in the definition of $\mathscr{D}_1$ for $\beta_2' \neq 0$ will follow from establishing

$$F_1(a) = \sum_{n=1}^{\infty} \langle F, \Gamma_n \rangle \Gamma_{n1}(a). \tag{25}$$

□

LEMMA 5. *Assume* P(i)-(iii). *If* $F = (F_1, F_2) \in \mathscr{D}_1$ *and* $G = (G_1, G_2) \in D(A)$, *then*
(i) $G \in \mathscr{D}_1$ *and* $J_1(F, G) = \langle F, AG \rangle$.
(ii) $\sum_{n=1}^{\infty} \lambda_n |\langle F, \Gamma_n \rangle|^2 < \infty$

Proof. It is only necessary to consider $\beta_2' \neq 0$. First we establish

$$G_1(a) = \sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}(a), \qquad G_1'(a) = \sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}'(a). \tag{26}$$

Now $G \in D(A)$ implies that

$$\langle AG, AG \rangle = \sum_{n=1}^{\infty} |\langle AG, \Gamma_n \rangle|^2 = \sum_{n=1}^{\infty} \lambda_n^2 |\langle G, \Gamma_n \rangle|^2 < \infty.$$

The convergence of this series, (22) and (23) and Schwarz's inequality imply that the two series

$$\sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}(x), \qquad \sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}'(x) \tag{27}$$

converge uniformly and absolutely on compact subsets of $[a, b)$. From the completeness of the eigenvectors, the series $\sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}$ converges in $\mathscr{L}_r^2(a, \infty)$ to G_1; hence we may conclude that the series in (27) converge to $G_1(x)$ and $G_1'(x)$ respectively. This establishes that $G \in \mathscr{D}_1$. The proof that $\langle F, AG \rangle = J_1(F, G)$ is the same as part (i) of Lemma 3 but with $(pG_1')(a)$ replaced by

$$c(G) = p(a) \sum_{n=1}^{\infty} \langle G, \Gamma_n \rangle \Gamma_{n1}'(a).$$

For the proof of part (ii), let $F_k = \sum_{n=1}^{k} \langle F, \Gamma_n \rangle \Gamma_n$. Then $F_k \in D(A)$ and, as in [9, Lemma 2],

$$J_1(F - F_k, F - F_k) = J_1(F, F) - \sum_{n=1}^{k} |\langle F, \Gamma_n \rangle|^2. \tag{28}$$

It follows from the definition of J_1 that

$$\begin{aligned} &J_1(F - F_k, F - F_k) \\ &\quad = \int_a^b p|F_1' - F_{1k}'|^2 + q|F_1 - F_{1k}|^2 \\ &\quad - \frac{1}{\rho}\Big\{\beta_1\beta_1'|F_1(a) - F_{1k}(a)|^2 - 2\beta_1\beta_2' p(a) \\ &\qquad \operatorname{Re}\Big([F_1(a) - F_{k1}(a)] \sum_{n=k+1}^{\infty} \langle F, \Gamma_n \rangle \Gamma_{n1}'(a)\Big) + \beta_2\beta_2' \Big| \sum_{n=k+1}^{\infty} \langle F, \Gamma_n \rangle \Gamma_{n1}'(a)\Big|^2\Big\}. \end{aligned} \tag{29}$$

From [9, p. 39] we have, for a member $f \in \mathscr{L}_r^2(a, b)$ which is locally absolutely continuous (without loss of generality, take $b > a+1$),

$$|f(a)|^2 \leqq \int_a^{a+1} |f|^2 + 2\left(\int_a^{a+1} |f|^2\right)^{1/2}\left(\int_a^{a+1} |f'|^2\right)^{1/2}$$
$$\leqq c_1 \int_a^b r|f|^2 + 2\sqrt{c_1 c_2}\left(\int_a^b r|f|^2\right)^{1/2}\left(\int_a^b p|f'|^2\right)^{1/2},$$

where $c_1 = \max 1/r(x)$ on $[a, a+1]$ and $c_2 = \max 1/p(x)$ on $[a, a+1]$. Using $f = F_1 - F_{k1}$ and $F_{k1} \to F_1$ in $L_r^2(a, b)$, we are able to conclude from (29) that

$$\liminf_{k\to\infty} J_1(F - F_k, F - F_k) > -\infty.$$

Using this inequality in (28) completes the proof. □

THEOREM 3. *Assume* P(i)–P(iii). *If* $F = (F_1, F_2) \in H$ *and* $\sum_{n=1}^{\infty} \lambda_n |\langle F, \Gamma_n\rangle|^2 < \infty$, *then*:

(i) $F_1(x) = \sum_{n=1}^{\infty} \langle F, \Gamma_n\rangle \Gamma_{n1}(x)$, *absolutely and uniformly on compact subsets of* $[a, b)$.

(ii) F_1 *is locally absolutely continuous and* $F_1' = \sum_{n=1}^{\infty} \langle F, \Gamma_n\rangle \Gamma_n'$ *with convergence in* $\mathscr{L}_{\sqrt{p}}^2(a, b)$.

(iii) $F \in \mathscr{D}_1$ *and* $J_1(F, F) = \sum_{n=1}^{\infty} \lambda_n |\langle F, \Gamma_n\rangle|^2$.

(iv) *If* $\beta_2' \neq 0$ *and* $F \in \mathscr{D}$, *then* $c(F) = (pF_1')(a)$.

The proof is essentially the same as that of [9, Theorem 1]. □

For $\lambda_1 \geqq 0$, then by the spectral theorem for self-adjoint operators, $\mathscr{D}_1$ is the domain of $A^{1/2}$. In the regular case, for $r = 1$ and q locally of bounded variation, asymptotic formulae made the description of the class $\mathscr{D}_1$ somewhat simpler (cf. [9, p. 41]). Such formulae do not yet seem to be available in the singular case. However, in any case, we have $\mathscr{D} \subset \mathscr{D}_1$ (recall $\mathscr{D} = \mathscr{D}_1$ if $\beta_2' = 0$), and $\mathscr{D}$ imposes only a mild condition of F_1 at a (in the $\beta_2' \neq 0$ case).

Acknowledgment. The author expresses his appreciation to the referee for suggestions which led to the improvement of the manuscript.

REFERENCES

[1] R. V. CHURCHILL, *Operational Mathematics*, McGraw-Hill, New York, 1958.

[2] R. V. CHURCHILL AND J. W. BROWN, *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1978.

[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, II, Interscience, New York, 1963.

[4] W. N. EVERITT, D. B. HINTON AND J. S. W. WONG, *On the strong limit-n classification of linear ordinary differential expressions of order 2n*, Proc. London Math. Soc., 3 (1974), pp. 351–367.

[5] K. FRIEDRICHS, *Criteria for discrete spectra*, Comm. Pure Appl. Math., 3 (1950), pp. 439–449.

[6] C. FULTON, *Two-point boundary value problems with eigenvalue parameter contained in the boundary conditions*, Proc. Roy. Soc. Edinburgh, 77A (1977), pp. 203–308.

[7] ———, *Singular eigenvalue problems with eigenvalue parameter contained in the boundary conditions*, to appear.

[8] C. FULTON AND S. PRUESS, *Numerical methods for a singular eigenvalue problem with eigenparameter in the boundary conditions*, J. Math. Anal. Appl., 71 (1979), pp. 431–462.

[9] D. HINTON, *An expansion theorem for an eigenvalue problem with eigenvalue parameter in the boundary condition*, Quart. J. Math. Oxford, 2 (1979), pp. 33–42.

[10] ———, *Eigenfunction expansions and spectral matrices of singular differential operators*, Proc. Roy. Soc. Edinburgh, 80A (1978), pp. 289–308.

[11] D. HINTON AND R. LEWIS, *Singular differential operators with spectra discrete and bounded below*, Proc. Roy. Soc. Edinburgh, 84A (1979), pp. 117–134.
[12] R. KAUFFMAN, T. READ AND A. ZETTL, *The Deficiency Index-Problem For Powers of Ordinary Differential Expressions*, Lecture Notes in Mathematics, 621, Springer-Verlag, Berlin, 1977.
[13] E. MÜLLER-PFEIFFER, *Spektraleigenschaften singulärer gewöhnlicher Differentialoperatoren*, Teubner-Texte zeur Mathematik, Leipzig, 1977.
[14] M. A. NAIMARK, *Linear Differential Operators*, II, Ungar, New York, 1968.
[15] T. T. READ, *Factorization and discrete spectra for second-order differential expressions*, J. Differential Equations, 35 (1980), pp. 388–406.
[16] J. WALTER, *Regular eigenvalue problems with an eigenvalue parameter in the boundary condition*, Math. Z., 133 (1973), pp. 301–312.

SIAM J. MATH. ANAL.
Vol. 12, No. 2, July 1981

0036-1410/81/1204-0009 $01.00/0

INTEGRABILITY OF RESOLVENTS OF SYSTEMS OF VOLTERRA EQUATIONS*

GUSTAF GRIPENBERG†

Abstract. The integrability of the resolvents of systems of Volterra integral and integrodifferential equations is studied. The matrix kernels in the equations need not be integrable, and some of the conditions used are shown to be both necessary and sufficient.

1. Introduction. The purpose of this paper is to study the integrability of the resolvents of the systems of Volterra equations

$$X(t)+\int_0^t X(t-s)A(s)\,ds=F(t),\qquad t\in\mathbb{R}^+=[0,\infty) \tag{1.1}$$

and

$$Y'(t)+\int_{[0,t]} Y(t-s)\,d\alpha(s)=G(t),\qquad t\in\mathbb{R}^+,\quad Y(0)=Y_0, \tag{1.2}$$

i.e., the solutions of the equations

$$R(t)+\int_0^t R(t-s)A(s)\,ds=A(t),\qquad t\in\mathbb{R}^+ \tag{1.3}$$

and

$$Q'(t)+\int_{[0,t]} Q(t-s)\,d\alpha(s)=0,\qquad t\in\mathbb{R}^+,\quad Q(0)=I. \tag{1.4}$$

Here X, F, Y and G are $\mathbb{C}^n$-valued functions (row vectors), A, R and Q are $n\times n$ matrix-valued functions and α is an $n\times n$ matrix of Borel measures. The reason for studying these resolvents is the fact that the solutions of (1.1) and (1.2) are (under fairly weak assumptions) given by

$$X(t)=F(t)-\int_0^t F(t-s)R(s)\,ds,\qquad t\in\mathbb{R}^+ \tag{1.5}$$

and

$$Y(t)=Y_0Q(t)+\int_0^t G(t-s)Q(s)\,ds,\qquad t\in\mathbb{R}^+. \tag{1.6}$$

The question we will consider here is under what assumptions on A and α we have

$$\int_0^\infty \|R(t)\|\,dt<\infty \quad\text{and}\quad \int_0^\infty \|Q(t)\|\,dt<\infty,$$

where $\|\cdot\|$ is some matrix norm (thus we will not consider weighted spaces); this problem is easily seen to be related to problems concerning the asymptotic behavior of the solutions of (1.1), (1.2) and perturbed forms of these equations. If A is integrable, then it is well known, see [20, p. 60], that R is integrable if and only if $\det(I+\hat{A}(z))\neq 0$, $\operatorname{Im} z\leqq 0$ (here $\hat{}$ denotes the Fourier or Fourier–Stieltjes transform, and functions or

* Received by the editors March 18, 1980, and in revised form November 3, 1980.
† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

measures defined on $\mathbb{R}^+$ are extended as 0 on $(-\infty, 0)$, so that the (j, k)-element in $\hat{A}(z)$ is $\int_0^\infty e^{-izt} a_{jk}(t)\, dt$). A similar result holds for (1.2); see [10] or [18]. For this reason we will consider kernels that are not necessarily integrable, and the approach taken here follows closely that of [3] (where $n = 1$). For earlier results on the resolvents of integral and integrodifferential equations, see, e.g., [1] and [3]–[21].

2. Statement of results. First we consider (1.3) and the resolvent R; the results concerning the resolvent Q will follow from an application of the ones concerning R, (cf. [18]). We will assume that at most one element of each row and each column in the matrix A is not integrable, and we give some examples below that show the difficulties encountered if this assumption is dropped. The nonintegrable functions we consider are, of course, not completely arbitrary, and we will use the following

DEFINITION 1.

$$S_1 = \Big\{ b : \mathbb{R}^+ \to \mathbb{C} \,\Big|\, b(t) = b_0(t) + \sum_{k=0}^{m} \int_{[0,t]} (t-s)^k \, d\beta_k(s), \quad t \in \mathbb{R}^+,\ b_0 \in L^1(\mathbb{R}^+),\ \beta_k \in \mathrm{BM}(\mathbb{R}^+),\ 0 \leqq k \leqq m < \infty \Big\}.$$

Here $\mathrm{BM}(\mathbb{R}^+)$ denotes the set of all finite Borel measures supported on $\mathbb{R}^+$. Since S_1 is the set of all functions that are the sum of an integrable function, a function of bounded variation, a function whose derivative is of bounded variation and so on, we see that, e.g., $b(t) = t^{-p}$, $p \in (0, 1)$, belongs to S_1. But note that for example the function $t^{-p} \cos(t)$ does not belong to S_1 and there is no good reason why such functions should be excluded from consideration. This is the motivation for the hypothesis (2.3) in Theorem 1. Observe also that S_1 is closed under the convolution product, denoted by $*$, since $L^1(\mathbb{R}^+)$, $\mathrm{BM}(\mathbb{R}^+)$ and the space of polynomials on $\mathbb{R}^+$ are closed under this product.

In the second definition we need, we identify locally absolutely continuous measures with their density functions.

DEFINITION 2. $S_2 = \{\beta \,|\, \beta$ is a locally finite Borel measure supported on $\mathbb{R}^+$, $\int_{\mathbb{R}^+} e^{yt} |d\beta(t)| < \infty$, $y < 0$, there exists a function $f \in L^1(\mathbb{R})$, such that $\hat{f}(x) = \lim_{y \to 0^-} 1/\hat{\beta}(x + iy)$, $|x| \leqq x_0$, where $x_0 > 0$ and $\lim_{z \to 0, \mathrm{Im} z < 0} 1/\hat{\beta}(z) = \hat{f}(0)\}$.

Unfortunately, it is in general very difficult to determine whether a given function or measure belongs to S_2; some results that are of use for our purposes are given in Theorems 3 and 4 below. But it is important to realize that (see Theorems 1 and 2 below) this is the central question in the study of the integrability of resolvents of equations with nonintegrable kernels via transformation methods.

Now we can state our first result.

THEOREM 1. *Assume that*

(2.1) $A(t) = (a_{jk}(t))$ *is an* $n \times n$ *matrix function defined on* $\mathbb{R}^+$, $n \geqq 1$;

(2.2) *there exists a permutation* σ *of* $\{1, \cdots, n\}$ *such that* $a_{jk} \in L^1(\mathbb{R}^+)$, $k \neq \sigma(j)$, $1 \leqq j, k \leqq n$;

(2.3) *there exist* N *distinct real numbers* $\omega_1, \cdots, \omega_N$ *such that* $a_{j\sigma(j)}(t) = \sum_{k=1}^{N} e^{i\omega_k t} b_{jk}(t)$, $t \in \mathbb{R}^+$, *where* $b_{jk} \in S_1$, $1 \leqq k \leqq N$, $1 \leqq j \leqq n$;

(2.4) $R(t) = (r_{jk}(t))$ *is the solution of* (1.3).

Then

(2.5) $$r_{jk} \in L^1(\mathbb{R}^+), \qquad 1 \leqq j, k \leqq n$$

if and only if

(2.6) $\inf_{y<0} |\det (I + \hat{A}(x+iy))| > 0,\ x \in \mathbb{R};$

(2.7) $\liminf_{y\to 0-} |\det (I + \hat{A}(\omega_k + iy))| \prod_{j\in J_k} |\hat{b}_{jk}(iy)|^{-1} > 0$, *where* $J_k = \{j | 1 \leqq j \leqq n, \liminf_{y\to 0-} |\hat{b}_{jk}(iy)| = +\infty\},\ 1 \leqq k \leqq N;$

(2.8) $b_{jk} \in S_2,\ j \in J_k,\ b_{jk} \in L^1(\mathbb{R}^+),\ j \notin J_k,\ 1 \leqq j \leqq n,\ 1 \leqq k \leqq N.$

Observe that as a consequence of (2.3) the condition (2.7) is always satisfied if $n = 1$ and is a consequence of (2.6) if the set J_k is empty, but in general (2.7) does not follow from (2.6). We also remark that (2.6) could be replaced by the stronger condition $\inf_{\mathrm{Im} z<0} |\det (I + \hat{A}(z))| > 0$.

Next we state the corresponding result for (1.4); for that purpose we need the following:

DEFINITION 3. $S_1^* = \{\beta | \beta$ is a locally finite Borel measure supported on $\mathbb{R}^+$ such that $\beta([0,t]) = \sum_{k=0}^{m} \int_{[0,t]} (t-s)^k\, d\beta_k(s),\ \beta_k \in \mathrm{BM}(\mathbb{R}^+),\ 0 \leqq k \leqq m < \infty\}$.

This set contains S_1 and is closed under convolution too.

THEOREM 2. *Assume that*

(2.9) $\alpha(E) = (\alpha_{jk}(E))$ *is an* $n \times n$ *matrix of Borel measures supported on* $\mathbb{R}^+,\ n \geqq 1;$

(2.10) *there exists a permutation of* $\{1, \cdots, n\}$ *such that* $\alpha_{jk} \in \mathrm{BM}(\mathbb{R}^+),\ k \neq \sigma(j),\ 1 \leqq j, k \leqq n;$

(2.11) *there exist* N *distinct real numbers* $\omega_1, \cdots, \omega_N$ *such that* $\alpha_{j\sigma(j)}([0,t]) = \sum_{k=1}^{N} \int_{[0,t]} e^{i\omega_k s}\, d\beta_{jk}(s),\ t \in \mathbb{R}^+$, *where* $\beta_{jk} \in S_1^*,\ 1 \leqq k \leqq N,\ 1 \leqq j \leqq n;$

(2.12) $Q(t) = (q_{jk}(t))$ *is the solution of* (1.4).

Then

(2.13) $$q_{jk} \in L^1(\mathbb{R}^+), \qquad 1 \leqq j, k \leqq n$$

if and only if

(2.14) $\inf_{y<0} |\det ((ix - y)I + \hat{\alpha}(x+iy))| > 0,\ x \in \mathbb{R};$

(2.15) $\liminf_{y\to 0-} |\det ((i\omega_k - y)I + \hat{\alpha}(\omega_k + iy))| \prod_{j\in J_k} |\hat{\beta}_{jk}(iy)|^{-1} > 0,$
where $J_k = \{j | 1 \leqq j \leqq n, \liminf_{y\to 0-} |\hat{\beta}_{jk}(iy)| = +\infty,\ 1 \leqq k \leqq N;$

(2.16) $\beta_{jk} \in S_2,\ j \in J_k,\ \beta_{jk} \in \mathrm{BM}(\mathbb{R}^+),\ j \notin J_k,\ 1 \leqq j \leqq n,\ 1 \leqq k \leqq N.$

Clearly these two theorems are not of very much use unless one knows something about the set S_2, and now we proceed to explore this problem. We say that r is the resolvent associated with the locally integrable function a if (1.3) holds with R and A replaced by r and a. The function q is said to be the differential resolvent associated with the locally finite measure α if (1.4) holds with Q and I replaced by q and 1.

First we give some general (and quite obvious) results concerning the set S_2. For the proof, (which will be omitted), one should recall that the Fourier–Stieltjes transform of a finite measure is locally equal to the Fourier transform of an integrable function; see also [2, p. 29].

THEOREM 3. *The following assertions hold*:

(2.17) *If* $\beta \in \mathrm{BM}(\mathbb{R}^+)$ *and* $\hat{\beta}(0) \neq 0$, *then* $\beta \in S_2$.

(2.18) *If* $\beta_1, \beta_2 \in S_2$, *then* $\beta_1 * \beta_2 \in S_2$.

(2.19) *If* $\beta_1 \in S_2$, $\beta_2 \in \mathrm{BM}(\mathbb{R}^+)$ *and* $\lim_{y\to 0^-} \hat{\beta_2}(0)/\hat{\beta_1}(iy) \neq -1$, *then* $\beta_1 + \beta_2 \in S_2$.

Note that the improvement of the results in [21] that is achieved in [14] relies in an essential way on the statement (2.19). The results in Theorem 3 can be combined with the more concrete and useful criteria that are given in

THEOREM 4. *The following assertions hold*:

(2.20) *If* $b(t) = b_1(t) + b_2(t)$, $t \in \mathbb{R}^+$, *where* $b_1 \in L^1_{\mathrm{loc}}(\mathbb{R}^+)$, $b_1 \notin L^1(\mathbb{R}^+)$ *is nonnegative, nonincreasing and convex on* $(0, \infty)$, $\lim_{t\to\infty} b_1(t) = 0$, $b_2 \in \mathrm{AC}_{\mathrm{loc}}((0,\infty))$ *is such that* $\lim_{t\to\infty} b_2(t) = 0$, $b_3, b_4 \in L^1_{\mathrm{loc}}(\mathbb{R}^+)$, $\limsup_{t\to\infty} (\int_0^t b_3(s)\,ds)(\int_0^t b_1(s)\,ds)^{-1} < 2^{-3/2}$ *and* $\limsup_{t\to\infty} (\int_0^t s b_4(s)\,ds)(\int_0^t s b_1(s)\,ds)^{-1} < \infty$, *where* $b_3(t) = \mathrm{var}\,(b_2; [t, \infty))$ *and* $b_4(t) = \int_0^\infty \mathrm{var}\,(b_2'; [s, \infty))\,ds$, $t \in \mathbb{R}^+$, *then* $b \in S_2$ *and* $\lim_{y\to 0^-} 1/\hat{b}(iy) = 0$.

(2.21) *If* $\beta \in S_1^*$ *and* $\beta_m(\mathbb{R}^+) \neq 0$, *then* $\beta \in S_2$ (m is the number appearing in Definition 3).

(2.22) *If* $b(t) = b_1(t) + (b_2 * \beta_2)(t)$, $t \in \mathbb{R}^+$, *where* $e^{yt} b_1(t) \in L^1(\mathbb{R}^+)$, $y < 0$, $\liminf_{y\to 0^-} |\hat{b_1}(iy)| = \infty$, *the resolvent* r_1 *associated with* b_1 *belongs to* $L^1(\mathbb{R}^+)$, $b_2 \in \mathrm{BV}(\mathbb{R}^+)$ *satisfies* $\lim_{t\to\infty} b_2(t) = 0$, $\mathrm{var}\,(b_2; [t, \infty)) \int_t^\infty |r_1(s)|\,ds \in L^1(\mathbb{R}^+)$ *and* $\beta_2 \in \mathrm{BM}(\mathbb{R}^+)$, *then* $b \in S_2$ *and* $\lim_{y\to 0^-} 1/\hat{b}(iy) = 0$.

(2.23) *If* $\beta = \beta_1 + b_2 * \beta_2$, *where* β_1 *is a Borel measure supported on* $\mathbb{R}^+$ *such that* $\int_{\mathbb{R}^+} e^{yt} |\,d\beta_1(t)| < \infty$, $y < 0$, $\liminf_{y\to 0^-} |\hat{\beta_1}(iy)| = \infty$, *the differential resolvent* q_1 *associated with* β_1 *belongs to* $L^1(\mathbb{R}^+)$, $b_2 \in \mathrm{BV}(\mathbb{R}^+)$ *satisfies* $\lim_{t\to\infty} b_2(t) = 0$, $\mathrm{var}\,(b_2; [t, \infty)) \int_t^\infty |q_1(s)|\,ds \in L^1(\mathbb{R}^+)$ *and* $\beta_2 \in \mathrm{BM}(\mathbb{R}^+)$, *then* $\beta \in S_2$ *and* $\lim_{y\to 0^-} 1/\hat{\beta}(iy) = 0$.

Here AC stands for absolute continuity and BV for bounded variation. Observe that Theorem 4 with the hypothesis (2.20) is essentially due to Shea and Wainger [21], and that for real functions b the assumptions in (2.20) can be formulated as follows: $b(t) = b_1(t) - b_2(t)$, where b_1 and b_2 are locally integrable, nonnegative, nonincreasing and convex and b_2 is sufficiently small compared to b_1 for large values of t. In particular, this assumption includes the case $b_2 \equiv 0$ which is the key ingredient of the results in [21]. From (2.21) we see, for example, that if $b(t) = b_1(t) + b_2(t)$, where $b_1 \in L^1(\mathbb{R}^+)$, $b_2 \in \mathrm{BV}(\mathbb{R}^+)$ and $\lim_{t\to\infty} b_2(t) \neq 0$, then $b \in S_2$. Clearly the usefulness of the conditions (2.22) and (2.23) is dependent on the existence of estimates for $\int_t^\infty |r_1(s)|\,ds$ or $\int_t^\infty |q_1(s)|\,ds$ for large t. In the theorem below we collect some known results in this direction.

THEOREM 5. *Let* $b \in L^1_{\mathrm{loc}}(\mathbb{R}^+)$, $b \notin L^1(\mathbb{R}^+)$ *and let* r *be the resolvent associated with* b. *Then the following assertions hold*:

(2.24) *If* b *is positive, nonincreasing and* $\log(b)$ *is convex on* $(0, \infty)$, *then* $\int_t^\infty |r(s)|\,ds \leqq (1 + \int_0^t b(s)\,ds)^{-1}$, $t \in \mathbb{R}^+$.

(2.25) *If* b *is nonnegative, nonincreasing, convex,* $\lim_{t\to\infty} b(t) = 0$ *and* $-b'$ *is convex on* $(0, \infty)$, *then* $\int_t^\infty |r(s)|\,ds = O(t^{-1} \int_0^t s^{-1} \int_0^s (1 + \int_0^u b(v)\,dv)^{-1}\,du\,ds)$ *as* $t \to \infty$.

(2.26) *If* b *satisfies the assumptions of* (2.20) *and* $\inf_{\mathrm{Im}\, z<0} |1 + \hat{b}(z)| > 0$, *then* $\int_t^\infty |r(s)|\,ds = O(t^{-1} \int_0^t s^{-1} \int_0^s b_1(u)\,du \int_0^s (1 + \int_0^u b_1(v)\,dv)^{-2}\,du\,ds)$ *as* $t \to \infty$.

The assertion (2.24) follows from [3, (1.8)–(1.10)], and (2.25) and (2.26) are established in [9].

A special case of (2.25) is established in [7].

Finally we consider some examples that show what can happen if we drop the assumption (2.2).

Example 1. Let $a_{11}(t) = a_{21}(t) = 1$, $a_{12}(t) = a_{22}(t) = 0$, $t \in \mathbb{R}^+$. Then it is easy to see that $r_{11}(t) = r_{21}(t) = e^{-t}$, $r_{12}(t) = r_{22}(t) = 0$, $t \in \mathbb{R}^+$, and note that (2.2) is not satisfied.

Example 2. Let $a_{11}(t) = \sum_{n=1}^{\infty} 2^{-n^2} \exp(-2^{-(n^4+1+(-1)^n)}t)$, $a_{21}(t) = \sum_{n=1}^{\infty} 2^{-n^2} \cdot \exp(-2^{-n^4}t)$, $a_{12}(t) = a_{22}(t) = 0$, $t \in \mathbb{R}^+$. Clearly a_{11} and a_{21} belong to $S_1 \cap S_2$ and a_{12}, $a_{22} \in L^1(\mathbb{R}^+)$. It is straightforward to check that r_{11}, r_{12}, $r_{22} \in L^1(\mathbb{R}^+)$ but $r_{21} \notin L^1(\mathbb{R}^+)$, although $\sup_{\mathrm{Im} z<0} |\hat{r}_{21}(z)| < \infty$, because $(1+2^{-(1+(-1)^n)})^{-1}\hat{r}_{21}(-i2^{-n^4}) \to 2^{-1}$ as $n \to \infty$.

Example 3. Let $a_{11}(t) = 1$, $a_{21}(t) = (t+1)^{-1}\sin(t^p)$, $a_{12}(t) = a_{22}(t) = 0$, $t \in \mathbb{R}^+$, $p \in (0, 1)$. Clearly $a_{11} \in S_1 \cap S_2$, $a_{12}, a_{22} \in L^1(R^+)$ but $a_{21} \in S_1$, $a_{21} \notin L^1(R^+) \cup S_2$ (see [21, p. 340] and the proof of Theorem 1 below), but it is easy to check that $r_{11}(t) = e^{-t}$, $r_{12}(t) = r_{22}(t) = 0$ and $r_{21}(t) = \int_0^t e^{-(t-s)}(ps^{p-1}(s+1)^{-1}\cos(s^p) - (s+1)^{-2}\sin(s^p))\,ds \in L^1(\mathbb{R}^+)$.

3. Proofs of Theorems 1 and 2. Assume that (2.6)–(2.8) hold. In order to prove that (2.5) holds, it is by (1.3) sufficient to show that

$$\hat{R}(z) = (\det(I + \hat{A}(z)))^{-1}\hat{A}(z)\,\mathrm{adj}\,(I + \hat{A}(z)), \qquad \mathrm{Im}\, z < 0 \tag{3.1}$$

is a matrix of $L^1(\mathbb{R}^+)\hat{}$-functions (i.e., Fourier transforms of functions in $L^1(\mathbb{R}^+)$ extended as 0 on $(-\infty, 0)$). From (2.2), (2.3) and (2.6) we see that the function on the right hand side in (3.1) is well defined and continuous in $\mathrm{Im}\, z \leqq 0$, except perhaps at the points $\omega_1, \cdots, \omega_N$ on the real axis.

First we are going to show that $\hat{R}(x)$ is a matrix of $L^1(\mathbb{R})\hat{}$-functions (i.e., Fourier transforms of functions in $L^1(\mathbb{R})$). From (2.3) we see that

$$\hat{a}_{j\sigma(j)}(x) = \sum_{k=1}^{N} \hat{b}_{jk}(x - \omega_k), \qquad x \neq \omega_k, \quad 1 \leqq k \leqq N, \quad 1 \leqq j \leqq n. \tag{3.2}$$

Moreover, there exists a number M (the maximum + 1 of the m's appearing in Definition 1 associated with the b_{jk}'s), such that

$$(ix - i\omega_k)^M (1 + ix - i\omega_k)^{-M} \hat{b}_{jk}(x - \omega_k) \in L^1(\mathbb{R}^+)\hat{}, \qquad 1 \leqq k \leqq N, \quad 1 \leqq j \leqq n. \tag{3.3}$$

In addition we observe that if $f \in L^1(\mathbb{R}^+)$, then

$$(ix - i\omega_k)^M (1 + ix - i\omega_k)^{-M} \hat{f}(x) \in L^1(\mathbb{R}^+)\hat{}, \qquad 1 \leqq k \leqq N \tag{3.4}$$

and

$$\left(\prod_{k=1}^{N} (ix - i\omega_k)^{Mn} - \prod_{k=1}^{N} (1 + ix - i\omega_k)^{Mn}\right) \prod_{k=1}^{N} (1 + ix - i\omega_k)^{-Mn} \in L^1(\mathbb{R}^+)\hat{}. \tag{3.5}$$

Let p_0 be an $L^1(\mathbb{R}^+)\hat{}$-function such that $p \overset{\mathrm{def}}{=} 1 - p_0$ is identically 0 in a neighborhood of the points $\omega_1, \cdots, \omega_N$ and identically 1 outside another neighborhood of these points, (for example the transform of a suitable combination of Fejer kernels). Let j, k, $1 \leqq j, k \leqq n$ be arbitrary and fixed. By (3.1) we can write

$$\begin{aligned} p(x)\hat{r}_{jk}(x) = p(x) &\prod_{m=1}^{N} ((ix - i\omega_m)^{Mn}(1 + ix - i\omega_m)^{-Mn})(\hat{A}(x)\,\mathrm{adj}\,(I + \hat{A}(x)))_{jk} \\ &\cdot \left(1 + \left(\prod_{m=1}^{N} (ix - i\omega_m)^{Mn} \det(I + \hat{A}(x)) \right.\right. \\ &\left.\left. - \prod_{m=1}^{N} (1 + ix - i\omega_m)^{Mn}\right) \prod_{m=1}^{N} (1 + ix - i\omega_m)^{-Mn}\right)^{-1}, \qquad x \in \mathbb{R}. \end{aligned} \tag{3.6}$$

It is a consequence of (2.2) and (3.2)–(3.5) that the right-hand side of (3.6) can be rewritten as $p(x)\,\hat{d_1}(x)(1+\hat{d_2}(x))^{-1}$ where $d_1, d_2 \in L^1(\mathbb{R}^+)$. Since (2.6) implies that $1+\hat{d_2}(x)$ is nonzero in a neighborhood of the set where $p(x)\,\hat{d_1}(x)$ is nonzero, it follows from (3.6) that (see, e.g., [20, pp. 61–63])

$$p(x)\hat{r_{jk}}(x) \in L^1(\mathbb{R})\hat{}. \tag{3.7}$$

Let h, $1 \leqq h \leqq N$, be arbitrary. We want to show that $\hat{r_{jk}}$ is equal to a function in $L^1(\mathbb{R})\hat{}$ in a neighborhood of ω_h. If we can do this, then it follows from (3.7) and an appropriate choice of the function p that (3.9) holds. We have by (3.1)

$$\begin{aligned}\hat{r_{jk}}(x) = \prod_{j\in J_h} (\hat{b_{jk}}(x-\omega_h))^{-1}(\hat{A}(x)\ \mathrm{adj}\ (I+\hat{A}(x)))_{jk} \\ \cdot\Big(\prod_{j\in J_h} (\hat{b_{jh}}(x-\omega_h))^{-1}\ \det\ (I+\hat{A}(x))\Big)^{-1}.\end{aligned} \tag{3.8}$$

We are going to use a localized version of the Wiener–Lévy theorem (see, e.g., [2, p. 29]), and we must make the following observations. By (2.3) and Definition 1, the functions $\hat{b_{de}}$, $1 \leqq d \leqq n$, $1 \leqq e \leqq N$ are equal to some functions in $L^1(\mathbb{R})\hat{}$ on any closed interval not containing 0. From (2.2), (2.3), (2.8) and Definition 2 we therefore conclude that

$$\prod_{j\in J_h} (\hat{b_{jh}}(x-\omega_h))^{-1}(\hat{A}(x)\ \mathrm{adj}\ (I+\hat{A}(x)))_{jk}$$

is equal to a function in $L^1(\mathbb{R})\hat{}$ in a neighborhood of ω_h. By the same reasoning we deduce, if we take (2.7) into account, that

$$\prod_{j\in J_h} (\hat{b_{jh}}(x-\omega_h))^{-1}\ \det\ (I+\hat{A}(x))$$

is equal to an $L^1(\mathbb{R})\hat{}$-function in a neighborhood of ω_h and this function is nonzero at ω_h. But then the desired conclusion follows from (3.8), and we have established

$$\hat{r_{jk}}(x) \in L^1(\mathbb{R})\hat{}. \tag{3.9}$$

If we can show that

$$\hat{r_{jk}}(z) \text{ is bounded and continuous in } \operatorname{Im} z \leqq 0, \tag{3.10}$$

then we can proceed in the same way as in [20, pp. 61–63] to prove that (3.9) and (3.10) imply (2.5) (since j, k are arbitrary). But $\hat{r_{jk}}(z)$ is clearly (by (2.2), (2.3), (2.6) and (3.1)) bounded and continuous everywhere in $\operatorname{Im} z \leqq 0$, except perhaps at the points ω_h, $1 \leqq h \leqq N$. Using (2.2), (2.3), (2.7), (2.8), (3.8) and Definition 2 we can argue in the same manner as above to deduce the continuity and boundedness of $\hat{r_{jk}}(z)$ in $\operatorname{Im} z \leqq 0$ at $z = \omega_h$. This shows that (3.10) holds, and the first part of the proof of Theorem 1 is completed.

Next we assume that (2.5) holds. Equation (3.1) can be rewritten as

$$(I+\hat{A}(z))(I-\hat{R}(z)) = (I-\hat{R}(z))(I+\hat{A}(z)) = I, \qquad \operatorname{Im} z < 0. \tag{3.11}$$

Since $\hat{R}(z)$ is bounded in $\operatorname{Im} z \leqq 0$ by (2.5), it follows from (3.11) that (2.6) holds.

Let h, $1 \leqq h \leqq N$, be arbitrary and define the set J_h^0 by

$$J_h^0 = \Big\{ j \,\Big|\, 1 \leqq j \leqq n,\ \limsup_{z\to 0, \operatorname{Im} z<0} |\hat{b_{jh}}(z)| = \infty \Big\}$$

and the $n \times 2n$ matrix $D(z) = (d_{jk}(z))$ by $d_{jk}(z) = \delta_{jk} + \hat{a_{jk}}(z)$, $1 \leqq j,\ k \leqq n$, $d_{jk}(z) = \delta_{j(k-n)}$, $1 \leqq j$, $(k-n) \leqq n$. Choose any $j \in J_h^0$, divide the jth row in $D(z)$ by $\hat{b_{jh}}(z-\omega_h)$

(when z is sufficiently close to ω_h) and perform row operations on $D(z)$ so that the only nonzero number in the $\sigma(j)$th column lies on the jth row, and finally let $z \to \omega_h$, $\operatorname{Im} z < 0$ so that $|b_{jh}(z-\omega_h)| \to \infty$. If we invoke (3.11) and the definition of the matrix $D(z)$ and recall the Gauss–Jordan method for finding the inverse of a matrix, we then conclude that

$$\hat{r}_{\sigma(j)k}(\omega_h) = \delta_{\sigma(j)k}, \qquad 1 \leqq k \leqq n, \quad j \in J_h^0. \tag{3.12}$$

From (3.11) we obtain

$$\sum_{k=1}^{n} (\delta_{\sigma(j)k} - \hat{r}_{\sigma(j)k})(\delta_{k\sigma(j)} + \hat{a}_{k\sigma(j)}) = 1, \qquad j \in J_h^0. \tag{3.13}$$

Since $\hat{b}_{jk}(z)$ is continuous in $\operatorname{Im} z \leqq 0$ except at 0, it follows from (2.2), (2.3), (3.2), (3.12) and (3.13) that

$$\lim_{z \to 0, Im z < 0} |\hat{b}_{jh}(z)| = \infty, \qquad j \in J_h^0;$$

i.e., we have shown that $J_h^0 = J_h$. Moreover, since each $\hat{b}_{jk}(x)$ is locally equal to a function in $L^1(\mathbb{R})\hat{}$, except near 0, we deduce from (2.2), (2.3), (3.2), (3.12) and (3.13) that $b_{jh} \in S_2$, $j \in J_h$. To see this we have only to note that by (3.13)

$$\begin{aligned}(\hat{b}_{jh}(z-\omega_h))^{-1} &= (\delta_{\sigma(j)j} - \hat{r}_{\sigma(j)j}(z)) \\ &\quad \cdot \Bigg(1 - \sum_{k=1, k \neq j}^{n} (\delta_{\sigma(j)k} - \hat{r}_{\sigma(j)k}(z))(\delta_{k\sigma(j)} + \hat{a}_{k\sigma(j)}(z)) \\ &\quad - (\delta_{\sigma(j)j} - \hat{r}_{\sigma(j)j}(z))(\delta_{j\sigma(j)} + \hat{a}_{j\sigma(j)}(z) - \hat{b}_{jh}(z-\omega_h))\Bigg)^{-1}.\end{aligned}$$

Let $E_1(z)$ be the $(n - |J_h|) \times (n - |J_h|)$ matrix one gets from the matrix $I + \hat{A}(z)$ by deleting the jth row and $\sigma(j)$th column, $j \in J_h$, and let $E_2(z)$ be the matrix of the same size that one gets by deleting the $\sigma(j)$th row and jth column from $I - \hat{R}(z)$, $j \in J_h$. By (3.11) we then have

$$E_1(z)E_2(z) = I + E_3(z), \tag{3.14}$$

where $E_3(z)$ is a matrix of $L^1(\mathbb{R}^+)\hat{}$-functions (by (2.2) and (2.5)) and $E_3(\omega_h) = 0$ (by (3.12)). The fact that $E_2(z)$ is bounded shows by (3.14) that $|\det(E_1(z))|$ must be bounded from below in a neighborhood of ω_h in $\operatorname{Im} z \leqq 0$ and if we recall the definition of $E_1(z)$ we see that (2.7) holds. Since $J_h^0 = J_h$ we know that $\det(E_1(z))$ is bounded in a neighborhood of ω_h in $\operatorname{Im} z \leqq 0$, and hence we can deduce from (3.14) that $\det(E_2(\omega_h)) \neq 0$ or $E_1(z) = (I + E_3(z))E_2(z)^{-1}$ close to ω_h. From this fact we see that if $j \notin J_h$, then $\hat{b}_{jh}(z)$ is continuous at 0 in $\operatorname{Im} z \leqq 0$, and $\hat{b}_{jh}(x)$ is equal to a function in $L^1(\mathbb{R})\hat{}$ at least in a neighborhood of 0, but also, as is easily seen (use (2.3) and Definition 1), on the whole real axis. But then we can conclude that $\hat{b}_{jh}$ is in $L^1(\mathbb{R}^+)\hat{}$ (we use the facts that it is bounded and continuous in $\operatorname{Im} z \leqq 0$); cf. [20, pp. 61–63]. This implies, since h was arbitrary, that we have completed the proof of Theorem 1. □

Now we proceed to the proof of Theorem 2. Let $B(t) = e^{-t}I$. Then we see that Q satisfies the equation

$$Q(t) + \int_0^t Q(t-s)A(s)\,ds = B(t), \qquad t \in \mathbb{R}^+,$$

where $A(t)=\int_{[0,t]}B(t-s)\,d\alpha(s)-B(t)$, $t\in\mathbb{R}^+$. Since now $I+\hat{A}(z)=(1+iz)^{-1}(izI+\hat{\alpha}(z))$ we see that if (2.9)–(2.11) and (2.14)–(2.16) hold, then (2.1)–(2.3) and (2.6)–(2.8) hold too, and we obtain (2.13) from (2.5) since

$$Q(t)=B(t)-\int_0^t B(t-s)R(s)\,ds, \qquad t\in\mathbb{R}^+.$$

The converse can be established in essentially the same way as in the proof of Theorem 1. It turns out to be sufficient to prove (2.6)–(2.8) with A as above and the only difference is that instead of (3.11) we now have

$$(I+\hat{A}(z))\hat{Q}(z)=\hat{Q}(z)(I+\hat{A}(z))=\hat{B}(z),$$

but this fact will not cause any difficulties. This completes the proof of Theorem 2.

4. Proof of Theorem 4. Let the assumptions in (2.20) hold. Since we can conclude from the results below that $\lim_{y\to 0^-}|\hat{b}(iy)|=\infty$, we may by (2.19) add a nonnegative, nonincreasing, convex and integrable function to b_1 and hence we can without loss of generality assume that

$$\int_0^t b_3(s)\,ds\leqq c_1\int_0^t b_1(s)\,ds, \qquad t\in\mathbb{R}^+, \quad c_1<2^{-3/2} \tag{4.1}$$

and

$$\int_0^t sb_4(s)\,ds\leqq c_2\int_0^t sb_1(s)\,ds, \qquad t\in\mathbb{R}^+, \quad c_2<\infty. \tag{4.2}$$

Next we will show that (4.4) holds. From results in [21, p. 320] we have (using the convexity of b_1)

$$\begin{aligned}|\hat{b_1}(x+iy)|&\geqq 2^{-1/2}\int_0^{\pi/|2x|}\cos(xt)\,e^{yt}b_1(t)\,dt\\ &=2^{-1/2}\int_0^{\pi/|2x|}(x\sin(xt)-y\cos(xt))\,e^{yt}\int_0^t b_1(s)\,ds\,dt,\end{aligned} \tag{4.3}$$

$$y\leqq 0, \quad x\neq 0.$$

Applying [21, (1.6)] we obtain after some integrations by parts

$$\begin{aligned}|\hat{b_2}(x+iy)|&\leqq|x|^{-1}\int_0^\infty|1-e^{-ixt}|\,e^{yt}|yb_2(t)+b_2'(t)|\,dt\\ &\leqq -2|x|^{-1}\int_0^{\pi/|2x|}|\sin(xt)|\frac{d}{dt}(e^{yt}b_3(t))\,dt-2|x|^{-1}\int_{\pi/|2x|}^\infty\frac{d}{dt}(e^{yt}b_3(t))\,dt\\ &=2\int_0^{\pi/|2x|}(x\sin(xt)-y\cos(xt))\,e^{yt}\int_0^t b_3(s)\,ds\,dt, \qquad y\leqq 0, \quad x\neq 0.\end{aligned}$$

This inequality combined with (4.1) and (4.3) yields

$$|\hat{b_2}(z)|\leqq c_1 2^{3/2}|b_1(z)|, \qquad \operatorname{Im} z\leqq 0, \quad z\neq 0. \tag{4.4}$$

Since $\operatorname{Re}\hat{b_1}(z)\geqq 0$, $\operatorname{Im} z\leqq 0$, (see [21, p. 320]), we have from (4.4)

$$|1+\hat{b_1}(z)+\hat{b_2}(z)|\geqq(1-c_1 2^{3/2})\max\{1,|\hat{b_1}(z)|\}, \qquad \operatorname{Im} z\leqq 0; \tag{4.5}$$

from the proof of [21, (1.5)] and (4.2) we obtain when we recall the definition of b_4 ($' = d/dx$),

$$|\hat{b_2}(x)'| \leqq 40c_2 \int_0^{1/|x|} tb_1(t)\,dt, \qquad x \neq 0. \tag{4.6}$$

If we use (4.5) and (4.6), then we can deduce in the same way as in [21] that the resolvent associated with b is in $L^1(\mathbb{R}^+)$ and the desired assertion follows from Theorem 1, (2.8), (with $n = 1$). This completes the proof of the statement (2.20).

To establish (2.21) it suffices to observe that $(iz)^m(1+iz)^{-m}\hat{\beta}(z)$ is the Fourier–Stieltjes transform of a finite Borel measure; since $\beta_m(\mathbb{R}^+) \neq 0$ it follows that this transform is $\neq 0$ when $z = 0$, and then it follows easily that $\beta \in S_2$.

Now let the assumptions in (2.22) hold. First we are going to show that (4.8) and (4.9) hold. Define $g(t) = \int_t^\infty |r_1(s)|\,ds$ and $b_3(t) = \operatorname{var}(b_2; [t, \infty))$ and note that since $\liminf_{y\to 0^-} |\hat{b_1}(iy)| = \infty$ it follows that $\hat{r_1}(0) = \int_0^\infty r_1(s)\,ds = 1$. This fact combined with Fubini's theorem yields, for an arbitrary $T > 0$,

$$\begin{aligned}\int_0^T \left| b_2(t) - \int_0^t b_2(t-s)r_1(s)\,ds \right| dt \\ \leqq \int_0^T b_3(t)g(t)\,dt + \int_0^T |r_1(s)| \int_s^T |b_2(t) - b_2(t-s)|\,dt\,ds.\end{aligned} \tag{4.7}$$

Since $|b_2(t) - b_2(t-s)| \leqq b_3(t-s) - b_3(t)$ we deduce with the aid of an integration by parts that

$$\int_0^T |r_1(s)| \int_s^T |b_2(t) - b_2(t-s)|\,dt\,ds \leqq \int_0^T g(t)b_3(t)\,dt,$$

and this inequality combined with (4.7) and the assumption in (2.22) gives

$$b_2 - b_2 * r_1 \in L^1(\mathbb{R}^+). \tag{4.8}$$

Since we clearly have

$$\limsup_{T\to\infty} \left| \int_0^T (b_2(t) - (b_2 * r_1)(t))\,dt \right| \leqq \limsup_{T\to\infty} \int_0^T b_3(T-s)g(s)\,ds = 0,$$

because b_3 and g are nonincreasing and $b_3(t)g(t) \in L^1(\mathbb{R}^+)$, we see that

$$\int_0^\infty (b_2(t) - (b_2 * r_1)(t))\,dt = 0. \tag{4.9}$$

From the definition of r_1 we have $\hat{b_1}(z)(1 - \hat{r_1}(z)) = \hat{r_1}(z)$, so that

$$\begin{aligned}(\hat{b}(z))^{-1} &= (\hat{b_1}(z) + \hat{b_2}(z)\hat{\beta_2}(z))^{-1} \\ &= (1 - \hat{r_1}(z))(\hat{r_1}(z) + \hat{b_2}(z)(1 - \hat{r_1}(z))\hat{\beta_2}(z))^{-1} \qquad \operatorname{Im} z < 0;\end{aligned}$$

since $\beta_2 \in \mathrm{BM}(\mathbb{R}^+)$, (4.8) and (4.9) hold, $r_1 \in L^1(\mathbb{R}^+)$ and $\hat{r_1}(0) = 1$, we see that $b \in S_2$ and that $\lim_{y\to 0^-} 1/\hat{b}(iy) = 0$. This completes the proof of (2.22).

Finally we let the assumptions in (2.23) hold. In the same manner as above we can now show that

$$b_2 * q_1 \in L^1(\mathbb{R}^+) \quad \text{and} \quad \int_0^\infty (b_2 * q_1)(t)\,dt = 0. \tag{4.10}$$

(In this case we use the fact that $\liminf_{y\to 0^-} |\hat{\beta_1}(iy)| = \infty$ implies that $\hat{q_1}(0) = \int_0^\infty q_1(t)\,dt = 0$.) Now we have $\hat{q_1}(z) = (iz + \hat{\beta_1}(z))^{-1}$, so that

$$\begin{aligned}(\hat{\beta}(z))^{-1} &= (\hat{\beta_1}(z) + \hat{b_2}(z)\hat{\beta_2}(z))^{-1} \\ &= \hat{q_1}(z)(1 - iz\hat{q_1}(z) + \hat{b_2}(z)\hat{q_1}(z)\hat{\beta_2}(z))^{-1}, \qquad \operatorname{Im} z < 0,\end{aligned}$$

and it is straightforward to check, using the facts that $\beta_2 \in \mathrm{BM}(\mathbb{R}^+)$, (4.10) holds, $q_1 \in L^1(\mathbb{R}^+)$ and $\hat{q_1}(0) = 0$, that $\beta \in S_2$ and $\lim_{y\to 0^-} 1/\hat{\beta}(iy) = 0$. This completes the proof of Theorem 4. □

REFERENCES

[1] R. W. CARR AND K. B. HANNSGEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979) pp. 961–984.

[2] R. R. GOLDBERG, *Fourier Transforms*, Cambridge University Press, London and New York, 1961.

[3] G. GRIPENBERG, *Integrability Properties of Resolvents of Volterra Equations*, Report-HTKK-MAT-A117, Helsinki University of Technology, Helsinki, 1978.

[4] ———, *On positive, nonincreasing resolvents of Volterra equations*, J. Differential Equations, 30 (1978), pp. 380–390.

[5] ———, *A Volterra equation with nonintegrable resolvent*, Proc. Amer. Math. Soc., 73 (1979), pp. 57–60.

[6] ———, *On rapidly decaying resolvents of Volterra equations*, J. Integral Equations, 1 (1979), pp. 241–247.

[7] ———, *On the asymptotic behavior of resolvents of Volterra equations*, this Journal, 11 (1980), pp. 654–662.

[8] ———, *On the resolvents of Volterra equations with nonincreasing kernels*, J. Math. Anal. Appl., 76 (1980), pp. 134–145.

[9] ———, *Decay estimates for resolvents of Volterra equations*, in preparation.

[10] S. I. GROSSMAN AND R. K. MILLER, *Nonlinear Volterra integrodifferential systems with L^1-kernels*, J. Differential Equations, 13 (1973), pp. 551–566.

[11] K. B. HANNSGEN, *Uniform boundedness in a class of Volterra equations*, this Journal, 6 (1975), pp. 689–697.

[12] ———, *A Wiener–Lévy theorem for quotients, with applications to Volterra equations*, Indiana Univ. Math. J., 29 (1980), pp. 103–120.

[13] G. S. JORDAN, *Asymptotic stability of a class of integrodifferential systems*, J. Differential Equations, 31 (1979), pp. 359–365.

[14] G. S. JORDAN AND R. L. WHEELER, *A generalization of the Wiener–Lévy theorem applicable to some Volterra equations*, Proc. Amer. Math. Soc., 57 (1976), pp. 109–114.

[15] ———, *Structure of resolvents of Volterra integral and integrodifferential systems*, this Journal, 11 (1980), pp. 119–132.

[16] ———, *Rates of decay of resolvents of Volterra equations with certain nonintegrable kernels*, J. Integral Equations, 2 (1980), pp. 103–110.

[17] ———, *Weighted L^1-remainder theorems for resolvents of Volterra equations*, this Journal, 11 (1980), pp. 885–900.

[18] R. K. MILLER, *Asymptotic stability and perturbations for linear Volterra integrodifferential systems*, in Delay and Functional Differential Equations and their Applications, K. Schmitt, ed., Academic Press, New York, 1972.

[19] R. K. MILLER AND A. N. MICHEL, *Stability of linear Volterra integrodifferential equations of order n*, this Journal, 10 (1979), pp. 1089–1091.

[20] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 19, American Mathematical Society, Providence, 1934.

[21] 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.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0010 $01.00/0

ASYMPTOTIC SOLUTIONS OF SOME NONLINEAR VOLTERRA INTEGRAL EQUATIONS*

GUSTAF GRIPENBERG†

Abstract. The asymptotic behavior of solutions of three nonlinear Volterra integral equations of the form

$$u(t)+\int_0^t A(t-s)g(u(s))\,ds=0$$

is studied. These equations arise from certain diffusion problems, in dimensions 1, 2 or 3, with nonlinear boundary conditions.

1. Introduction and statement of results.

The purpose of this paper is to study the asymptotic behavior of the solutions of the equations

$$u(t)+\int_0^t A_n(t-s)g(u(s))\,ds=0, \qquad t\in\mathbb{R}^+=[0,\infty), \quad n=1,2,3, \tag{1.1}$$

where

$$A_n(t)=2\pi^{-2}R^{-1}\int_0^\infty e^{-ts}s^{-1}(J_{n/2}(Rs^{1/2})^2+Y_{n/2}(Rs^{1/2})^2)^{-1}\,ds, \tag{1.2}$$

$$t>0, \quad n=1,2,3.$$

Here $R>0$ and J and Y are the Bessel functions of the first and second kind. We note that of course we can also write $A_1(t)=(\pi t)^{-1/2}$ and $A_3(t)=(\pi t)^{-1/2}-R^{-1}\exp(R^{-2}t)\operatorname{erfc}(R^{-1}t^{1/2})$, $t>0$.

These equations arise from the following diffusion problem. Let $w(t,r)$ be the solution of the heat equation (with a symmetry assumption if $n=2$ or 3),

$$w_t(t,r)=w_{rr}(t,r)+(n-1)r^{-1}w_r(t,r), \qquad r>R, \quad t>0,$$

$$w(0,r)=0, \qquad r>R, \quad \lim_{r\to\infty} w(t,r)=0, \quad t>0,$$

$$w_r(t,R)=g(w(t,R)), \qquad t>0.$$

If the solution of this equation is expressed in terms of $g(w(t,R))$ with the aid of an appropriate Green's function, then one obtains (1.1) with $u(t)=w(t,R)$.

Equation (1.1) with $n=1$ and related equations (i.e., containing forcing terms) have been studied in, e.g., [2], [4]–[6] and [8]–[11], mainly in connection with heat conduction problems but also for other kinds of diffusion processes, (see, e.g., [10]). In most of these papers it is assumed that $g(0)=0$, (thus, this paper does not have a direct bearing on those results); then forcing terms are needed if one would like to have a nontrivial solution. Here it is assumed (as in [10]) that $u(t)\equiv 0$ is not a solution of (1.1), that is, $g(0)\neq 0$; we want to know what value the solution approaches as $t\to\infty$ and, more important, to find the rate of convergence. Observe also that our asymptotic solutions are not derived from formal expansions but we give proofs for all our results. Thus, for example, we obtain the asymptotic estimates (for large t) that were formally derived in [10].

We also remark that the methods used in this paper are, except for certain details concerning the asymptotic estimates, not restricted to the kernels A_n considered here,

* Received by the editors June 11, 1980, and in revised form November 3, 1980.
† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

but can in fact be applied to a much larger class of problems. Hence (1.1) with $n = 1, 2$ or 3 can be considered to be a model problem.

We state our results on the asymptotic behavior of the solutions of (1.1) in the following

THEOREM. *Assume that* $R > 0$, $n = 1, 2$ *or* 3, *and that*

(1.3) *g is a continuous and nondecreasing function on* $(0, U_0]$ *where* $U_0 > 0$ *and* $g(U_0) = 0$;

(1.4) *if* $n = 1$ *or* 2, *then* $g(x) = -\Gamma_0(U_0 - x)^\alpha + O((U_0 - x)^\beta)$ *as* $x \to U_0-$, $\Gamma_0 > 0$, $\alpha \geqq 1$ *and* $\beta > \alpha$;

(1.5) *if* $n = 3$, *then* $g(x) = g(U_R) - \Gamma_R(U_R - x) + O((U_R - x)^\beta)$ *as* $x \to U_R-$, $\Gamma_R \geqq 0$, $\beta > 1$ *and* $U_R \in (0, U_0)$ *satisfies* $Rg(U_R) + U_R = 0$.

Then there exists a unique, nonnegative and continuous solution u of (1.1) *such that* $\int_0^t |g(u(s))|\, ds < \infty$, *u is nondecreasing and*

(1.6) $u(t) = U_0 - (U_0/\Gamma_0)^{1/\alpha}(\pi t)^{-1/(2\alpha)} + O(t^{-\gamma})$ *for any* $\gamma \in (1/(2\alpha), \min\{1/\alpha, (\beta - \alpha + 1)/(2\alpha)\})$ *as* $t \to \infty$ *if* $n = 1$;

(1.7) $u(t) = U_0 - (2U_0)^{1/\alpha}(\Gamma_0 R \ln(t))^{-1/\alpha} + O((\ln(t))^{-\gamma})$ *for any* $\gamma \in (3/(2\alpha), \min\{2/\alpha, (\beta - \alpha + 1)/\alpha\})$ *as* $t \to \infty$ *if* $n = 2$;

(1.8) $u(t) = U_R - U_R R(1 + \Gamma_R R)^{-1}(\pi t)^{-1/2} + O(t^{-\gamma})$, $\gamma = \min\{\beta/2, 3/2\}$ *as* $t \to \infty$ *if* $n = 3$.

Note that if g were defined on $(-\infty, 0)$ and positive there, then there could exist other, negative, solutions of (1.1).

If $\alpha = 1$, then the results in (1.6) and (1.7) can be slightly improved by applying an argument similar to the one used in the proof of (1.8).

Observe also that if g is continuous and nondecreasing on $[U_0, 0)$, $U_0 < 0$, $g(U_0) = 0$, then one can replace u by $-u$ and g by $g^*(x) = -g(-x)$ in (1.1) and apply the theorem.

Finally we remark that we do not assume that g is continuous at 0; see, e.g., [10] where a model for gas absorption in a liquid is considered in which $g(x) = x - x^{-\gamma}$, $\gamma > 0$.

2. Proof of the theorem. First we establish the existence of a solution of (1.1) with the desired qualitative properties, and then we proceed to the quantitative parts of the assertion. This existence result is a very small variation on known theorems; in order to treat all three cases at the same time we state it as a lemma.

LEMMA. *Assume that* $a \in L^1_{\mathrm{loc}}(\mathbb{R}^+)$ *is positive, nonincreasing and* $\ln(a)$ *is convex on* $(0, \infty)$ *and that* (1.3) *holds. Then there exists a unique, continuous solution u of*

$$u(t) + \int_0^t a(t-s)g(u(s))\, ds = 0, \qquad t \in \mathbb{R}^+ \tag{2.1}$$

such that $u(t) \in [0, U_0]$, *u is nondecreasing and* $g(u(t)) \in L^1_{\mathrm{loc}}(\mathbb{R}^+)$. *If* g_0 *is another function satisfying* (1.3) *and* $g_0(x) \geqq g(x)$ *on* $(0, U_0]$, *then the corresponding solution* u_0 *of* (2.1) *satisfies* $u_0(t) \leqq u(t)$, $t \in \mathbb{R}^+$.

Proof. We let $g(x) = 0$, $x > U_0$, and if $\lim_{x \to 0+} g(x)$ is finite, then we let $g(x)$ be that finite value when $x \in (-\infty, 0)$), (otherwise g is left undefined on $(-\infty, 0)$). We define the functions j_λ and g_λ, $\lambda > 0$, by $j_\lambda(x) + \lambda g(j_\lambda(x)) = x$ and $g_\lambda(x) = g(j_\lambda(x))$, $x \in (-\infty, \infty)$,

(g_λ is the Yosida approximation of g). We let b_λ be the solution of the equation

$$b_\lambda(t)+\lambda^{-1}\int_0^t a(t-s)b_\lambda(s)\,ds=\lambda^{-1}a(t),\qquad t\in\mathbb{R}^+. \tag{2.2}$$

It follows from the assumptions on a that b_λ is nonnegative and $\int_0^\infty b_\lambda(t)\,dt\leqq 1$; see, e.g., [7, Chap. IV]. Since j_λ is Lipschitz continuous, there certainly exists a solution of the equation

$$u_\lambda(t)=\int_0^t b_\lambda(t-s)j_\lambda(u_\lambda(s))\,ds,\qquad t\in\mathbb{R}^+ \tag{2.3}$$

for every $\lambda>0$. Since, moreover, $j_\lambda(x)\geqq 0$ if $x\geqq 0$ and j_λ is nondecreasing, a standard argument (using the nonnegativity of b_λ) shows that u_λ is continuous, nonnegative and nondecreasing. Since also $j_\lambda(x)\leqq U_0$ if $x\leqq U_0$, we get $u_\lambda(t)\leqq U_0$ (as $\int_0^\infty b_\lambda(t)\,dt\leqq 1$). If we apply (2.2), then we can rewrite (2.3) as

$$u_\lambda(t)+\int_0^t a(t-s)g_\lambda(u_\lambda(s))\,ds=0,\qquad t\in\mathbb{R}^+, \tag{2.4}$$

and if we invoke the results in [1, Thm. 2], then we conclude that the functions u_λ converge uniformly on compact subsets of $\mathbb{R}^+$ as $\lambda\to 0^+$ to a function u, (that must then also be nonnegative, nondecreasing and bounded by U_0). That this limit function is a solution of (2.1) such that $g(u(t))\in L^1_{\mathrm{loc}}(\mathbb{R}^+)$ is clear once we observe that we have, by (1.3) and (2.4), $\sup_{\lambda\in(0,1)}\int_0^t|g_\lambda(u_\lambda(s))|ds<\infty$, $t\in\mathbb{R}^+$. The uniqueness of this solution is established in the same way as in [1].

If the function $j_{0,\lambda}$ is defined by $j_{0,\lambda}(x)+\lambda g_0(j_{0,\lambda}(x))=x$, then we clearly have $j_{0,\lambda}(x)\leqq j_\lambda(x)$, and hence a standard comparison argument (see [7, § II 6]) applied to (2.3) shows that $u_{0,\lambda}(t)\leqq u(t)$, $t\in\mathbb{R}^+$, ($u_{0,\lambda}$ is the solution of (2.3) when j_λ is replaced by $j_{0,\lambda}$). Since this is true for arbitrary $\lambda>0$, it follows that $u_0(t)\leqq u(t)$, $t\in\mathbb{R}^+$. This completes the proof of the lemma. □

The fact that the kernels A_n satisfy the hypothesis of the lemma is shown in [7, § IV 7].

Now we proceed to prove that (1.6) and (1.7) hold (the proof of (1.8) involves some slightly different ideas). It follows from the last statement in the lemma that we may without loss of generality assume the existence of a positive number c such that

(2.5) the function $x-cg(x)$ is a nondecreasing function of x on $[U_0-\varepsilon, U_0]$ for some $\varepsilon>0$.

(At this point it is essential to assume that $\alpha\geqq 1$.) Let the functions B_1 and B_2 be the solutions of

$$B_n(t)+c^{-1}\int_0^t A_n(t-s)B_n(s)\,ds=c^{-1}A_n(t),\qquad t\in\mathbb{R}^+,\quad n=1,2. \tag{2.6}$$

It is straightforward to check, using the inversion formula for the Stieltjes transform (note that $A_1^\wedge(z)=z^{-1/2}$ and $A_2^\wedge(z)=z^{1/2}K_0(Rz^{1/2})/K_1(Rz^{1/2})$ are the Laplace transforms of A_1 and A_2 where K_0 and K_1 are modified Bessel functions of the second kind), that

$$B_1(t)=(c\pi)^{-1}\int_0^\infty e^{-ts}s^{1/2}(s+c^{-2})^{-1}\,ds,\qquad t>0 \tag{2.7}$$

and

$$B_2(t) = 2(cR)^{-1}\pi^{-2}\int_0^\infty e^{-ts}((s^{1/2}J_1(Rs^{1/2}) + c^{-1}J_0(Rs^{1/2}))^2 + (s^{1/2}Y_1(Rs^{1/2}) + c^{-1}Y_0(Rs^{1/2}))^2)^{-1}\,ds, \quad t>0. \tag{2.8}$$

(For similar results and calculations, see, e.g., [3].) What we really need to know about these two functions is the following:

$$B_n(t) \geqq 0, \quad t>0, \qquad \int_0^\infty B_n(s)\,ds = 1, \quad n = 1, 2, \tag{2.9}$$

$$B_1(t) = c2^{-1}\pi^{-1/2}t^{-3/2} + O(t^{-5/2}) \quad \text{as } t\to\infty \tag{2.10}$$

and

$$B_2(t) = 2cR^{-1}t^{-1}(\ln(t))^{-2} + O(t^{-1}(\ln(t))^{-3}) \quad \text{as } t\to\infty. \tag{2.11}$$

These assertions can be derived from (2.7), (2.8), the properties of the Bessel functions and the fact that $B_n^\wedge(0) = (c(A_n^\wedge(0))^{-1}+1)^{-1} = 1$.

Using (2.6) we can rewrite (1.1) as

$$v(t) = U_0\Big(1 - \int_0^t B_n(s)\,ds\Big) + \int_0^t B_n(t-s)(v(s) + cg(U_0 - v(s)))\,ds, \qquad t\in\mathbb{R}^+, \quad n = 1, 2, \tag{2.12}$$

where

$$v(t) = U_0 - u(t), \qquad t\in\mathbb{R}^+. \tag{2.13}$$

Since $\int_0^\infty A_n(s)\,ds = \infty$ when $n = 1$ or 2 and u is nondecreasing, it follows from (1.1), (1.3) and (1.4) that $\lim_{t\to\infty} u(t) = U_0$ in these cases, and therefore we can by (2.13) choose t_0 so large that

$$0 \leqq v(t) < \varepsilon \quad \text{if } t \geqq t_0. \tag{2.14}$$

Let now $n = 1$ and define the function v_0 by

$$v_0(t) = \begin{cases} v(t), & t\in[0, t_0], \\ \left(\dfrac{U_0}{\Gamma_0}\right)^{1-\alpha}(\pi t)^{-1/(2\alpha)} - t^{-\gamma} - h_1(t), & t > t_0, \end{cases} \tag{2.15}$$

where $\gamma \in ((2\alpha)^{-1}, \min\{\alpha^{-1}, (\beta-\alpha+1)/(2\alpha)\})$ is arbitrary and h_1 is a nonnegative, piecewise continuous function that will be specified later. Suppose that we are able to show that when $n = 1$

$$v_1(t) \overset{\text{def}}{=} U_0\Big(1 - \int_0^t B_n(s)\,ds\Big) + \int_0^t B_n(t-s)(v_0(s) + cg(U_0 - v_0(s)))\,ds \geqq v_0(t), \qquad t\in\mathbb{R}^+. \tag{2.16}$$

Then it would follow from (2.5), (2.9), (2.12)–(2.15) and a standard comparison argument (see [7, § II 6]), that $v(t) \geqq v_0(t)$, $t\in\mathbb{R}^+$.

In order to establish (2.16) with $n=1$ we need the following observations. If $h_1(t)\geqq 0$, then it follows from (1.3) and (1.4) that

$$g\Big(U_0-\Big(\frac{U_0}{\Gamma_0}\Big)^{1/\alpha}(\pi t)^{-1/(2)}+t^{-\gamma}+h_1(t)\Big)$$

$$\geqq -U_0(\pi t)^{-1/2}+U_0^{1-1/\alpha}\Gamma_0^{1/\alpha}\pi^{(1-\alpha)/(2\alpha)}t^{(1-\alpha)/(2\alpha)-\gamma} \tag{2.17}$$

$$+O(t^{-\beta/(2\alpha)}+t^{-1/2+1/\alpha-2\gamma}) \quad \text{as } t\to\infty.$$

Using Laplace transforms one verifies without difficulty that

$$1-\int_0^t B_1(s)\,ds = c\int_0^t B_1(t-s)(\pi s)^{-1/2}\,ds, \qquad t\in\mathbb{R}^+. \tag{2.18}$$

From (2.9), (2.10) and some mechanical calculations one obtains

$$\int_{t_0}^t B_1(t-s)s^{-\delta}\,ds = t^{-\delta}-t^{-\delta}\int_{t-t_0}^{\infty}B_1(s)\,ds+\int_0^{t/2}((t-s)^{-\delta}-t^{-\delta})B_1(s)\,ds$$

$$+\int_{t/2}^{t-t_0}((t-s)^{-\delta}-t^{-\delta})B_1(s)\,ds \tag{2.19}$$

$$=t^{-\delta}+O(t^{-\delta-1/2}) \quad \text{as} \quad t\to\infty, \quad \delta\in(0,1),$$

since $(t-s)^{-\delta}-t^{-\delta}\leqq\delta 2^{\delta}t^{-\delta-1}s$ if $s\in[0,t/2]$. If we combine (2.17)–(2.19) with the definition of v_1, then we see that

$$v_1(t)\geqq\Big(\frac{U_0}{\Gamma_0}\Big)^{1/\alpha}(\pi t)^{-1/(2\alpha)}-t^{-\gamma}-\int_{t_0}^t B_1(t-s)h_1(s)\,ds$$

$$+cU_0\Big(\frac{U_0}{\Gamma_0}\Big)^{-1/\alpha}\pi^{(1-\alpha)/(2\alpha)}t^{(1-\alpha)/(2\alpha)-\gamma} \tag{2.20}$$

$$+O(t^{-\beta/(2\alpha)}+t^{-1/(2\alpha)-1/2}) \text{ as } t\to\infty.$$

Next we are going to choose the function h_1. Let the function k_1 be defined by

$$k_1(t)=\begin{cases}0 & \text{if } t\in[0,t_0) \quad \text{or } t\in(t_1,t_2),\\ c_1 & \text{if } t\in[t_0,t_1],\\ -c_2B_1(t-t_0) & \text{if } t\geqq t_2,\end{cases} \tag{2.21}$$

where t_1, t_2 and c_1 are certain positive constants and $c_2=c_1(t_1-t_0)(1-\int_0^{t_2-t_0}B_1(s)\,ds)^{-1}$. Let h_1 be the solution of the equation

$$h_1(t)-\int_0^t B_1(t-s)h_1(s)\,ds=k_1(t), \qquad t\in\mathbb{R}^+. \tag{2.22}$$

Clearly h_1 is nonnegative when $t\in[t_0,t_2)$. If $t\geqq t_2$, then we have, by (2.6), (2.21), (2.22) and the definition of c_2,

$$h_1(t)=k_1(t)+c^{-1}\int_0^t A_1(t-s)k_1(s)\,ds$$

$$=c^{-1}c_1\int_{t_0}^{t_1}(A_1(t-s)-A_1(t-t_0))\,ds \tag{2.23}$$

$$+c^{-1}c_2\int_{t_0}^{t_2}(A_1(t-s)-A_1(t-t_0))B_1(s-t_0)\,ds, \qquad t\geqq t_2,$$

so that $h_1(t) \geqq 0$ and $h_1(t) = O(t^{-3/2})$ as $t \to \infty$. By choosing c_1, t_1 and t_2 sufficiently large, (observe that $B_1(t)(1 - \int_0^t B_1(s)\,ds)^{-1} = O(t^{-1})$ as $t \to \infty$), we see from (2.20)–(2.22) that (2.16) holds. As we already noted above, this implies that $v(t) \geqq v_0(t)$, $t \in \mathbb{R}^+$; since $h_1(t) = O(t^{-3/2})$ as $t \to \infty$, this implies that

$$v(t) \geqq \left(\frac{U_0}{\Gamma_0}\right)^{1/\alpha} (\pi t)^{-1/(2\alpha)} + O(t^{-\gamma}) \quad \text{as } t \to \infty. \tag{2.24}$$

By almost exactly the same argument as the one used in deriving (2.24) one deduces that

$$v(t) \leqq \left(\frac{U_0}{\Gamma_0}\right)^{1/\alpha} (\pi t)^{-1/(2\alpha)} + O(t^{-\gamma}) \quad \text{as } t \to \infty,$$

and if this result is combined with (2.24), then one obtains (1.6).

Now we proceed to consider the case $n = 2$. The idea of the proof is the same as in the case $n = 1$. We define the function C by

$$C(t) = 2R^{-1}\pi^{-2} \int_0^\infty e^{-ts} s^{-1} (J_0(Rs^{1/2})^2 + Y_0(Rs^{1/2})^2)^{-1}\,ds, \qquad t > 0. \tag{2.25}$$

The reason for introducing this function is that

$$1 - \int_0^t B_2(s)\,ds = c \int_0^t B_2(t-s) C(s), \qquad t \in \mathbb{R}^+. \tag{2.26}$$

To see this, use the fact that $C^\Lambda(z) = z^{-1/2} K_1(Rz^{1/2})/K_0(Rz^{1/2})$ to derive (2.26) and then the inversion formula for the Stieltjes transform to derive (2.25). What we really need to know about the function C is that

$$C(t) = 2R^{-1}(\ln(t))^{-1} + O((\ln(t))^{-2}) \quad \text{as } t \to \infty, \tag{2.27}$$

a fact that is seen to be a consequence of (2.25). Now we define the function v_0 by

$$v_0(t) = \begin{cases} v(t) & \text{if } t \in [0, t_0], \\ \left(\dfrac{U_0}{\Gamma_0}\right)^{1/\alpha} C(t)^{1/\alpha} - (\ln(t))^{-\gamma} - h_2(t) & \text{if } t > t_0, \end{cases} \tag{2.28}$$

where $\gamma \in (3/(2\alpha), \min\{2\alpha^{-1}, (\beta - \alpha + 1)/\alpha\})$ is arbitrary and h_2 is a nonnegative function that will be specified later. Instead of (2.17) we have in this case

$$\begin{aligned} &g\left(U_0 - \left(\frac{U_0}{\Gamma_0}\right)^{1/\alpha} C(t)^{1/\alpha} + (\ln(t))^{-\gamma} + h_2(t)\right) \\ &\qquad \geqq -U_0 C(t) + U_0^{1-1/\alpha} \Gamma_0^{1/\alpha} C(t)^{1-1/\alpha} (\ln(t))^{-\gamma} + O((\ln(t))^{-\beta/\alpha} \\ &\qquad\qquad + (\ln(t))^{-1+2/\alpha-2\gamma}) \quad \text{as } t \to \infty. \end{aligned} \tag{2.29}$$

Using (2.9), (2.11), (2.25) and (2.27) we deduce after some straightforward calculations (cf. (2.29) and assume that $t_0 > 1$) that

$$\int_{t_0}^t B_2(t-s)(\ln(s))^{-\delta}\,ds = (\ln(t))^{-\delta} + O((\ln(t))^{-\delta-1}) \quad \text{as } t \to \infty, \quad \delta > 0 \tag{2.30}$$

and

$$\int_{t_0}^t B_2(t-s) C(s)^{1/\alpha}\,ds = C(t)^{1/\alpha} + O((\ln(t))^{-1/\alpha-1}) \quad \text{as } t \to \infty. \tag{2.31}$$

If we now combine the definition of v_1, (take $n=2$ in (2.16)), with (2.28)–(2.31), then we clearly obtain

$$v_1(t) \geqq \left(\frac{U_0}{\Gamma_0}\right)^{1/\alpha} C(t)^{1/\alpha} - (\ln(t))^{-\gamma} - \int_{t_0}^{t} B_2(t-s)h_2(s)\,ds$$

$$+c\left(\frac{2U_0}{R}\right)^{1-1/\alpha} \Gamma_0^{1/\alpha}(\ln(t))^{-1+1/\alpha-\gamma} \tag{2.32}$$

$$+O((\ln(t))^{-\beta/\alpha} + (\ln(t))^{-1/\alpha-1}) \quad \text{as } t\to\infty.$$

We choose the function h_2 in the same way as h_1; the only difference is that in (2.21)–(2.23) A_1 and B_1 are replaced by A_2 and B_2 respectively, and since $B_2(t)(1-\int_0^t B_2(s)\,ds)^{-1} = O(t^{-1}(\ln(t))^{-1})$ as $t\to\infty$ we can pick the constants c_1, t_1 and t_2 to be such that (2.16) holds. Thus we are able (by the same argument as in the case $n=1$), to conclude that $v(t) \geqq v_0(t)$, $t\in\mathbb{R}^+$, and since by (2.23) $h_2(t) = O(t^{-1})$ as $t\to\infty$ (note that $A_2(t) = O(t^{-1})$ by (1.2)), we obtain from (2.27) and (2.28) that

$$v(t) \geqq (2U_0)^{1-\alpha}(\Gamma_0 R\ln(t))^{-1/\alpha} + O((\ln(t))^{-\gamma}) \quad \text{as } t\to\infty. \tag{2.33}$$

By a similar argument we get

$$v(t) \leqq (2U_0)^{1/\alpha}(\Gamma_0 R\ln(t))^{-1/\alpha} + O((\ln(t))^{-\gamma}) \quad \text{as } t\to\infty,$$

and this result together with (2.33) gives (1.7).

Finally we consider the case $n=3$. Since $A_1^{\wedge}(z) = z^{-1/2}$ and $A_3^{\wedge}(z) = (z^{1/2}+R^{-1})^{-1}$, we see that

$$A_3(t) + R^{-1}\int_0^t A_3(t-s)A_1(s)\,ds = A_1(t), \qquad t\in\mathbb{R}^+,$$

and therefore (1.1) with $n=3$ can be rewritten as

$$u(t) + \int_0^t A_1(t-s)g_1(u(s))\,ds = 0, \qquad t\in\mathbb{R}^+,$$

where $g_1(x) = R^{-1}x + g(x)$. If we apply the first part of the theorem, then we conclude that

$$u(t) = U_R - U_R R(1+\Gamma_R R)^{-1}(\pi t)^{-1/2} + O(t^{-\gamma}), \quad \gamma\in(\tfrac{1}{2}, \min\{1, \beta/2\}) \text{ as } t\to\infty. \tag{2.34}$$

It remains for us to show that we can actually take $\gamma = \min\{3/2, \beta/2\}$. To do this we take $c = R(1+\Gamma_R R)^{-1}$ in (2.6), and we get (2.12) and (2.13) with U_0 replaced by U_R and g by g_1. (Now (2.5) need not hold but that is not a problem.) If we use the definition of g_1, (1.5), (2.9), (2.10), (2.12), (2.13) and (2.34) then we see after some mechanical calculations that (1.9) holds. (We use the fact that if $f_i\in L^1_{\mathrm{loc}}(\mathbb{R}^+)$ satisfy $f_i(t) = O(t^{-\delta_i})$ as $t\to\infty$, $i=1,2$, $0<\delta_1\leqq\delta_2$, $\delta_2>1$, then $\int_0^t f_1(t-s)f_2(s)\,ds = O(t^{-\delta_1})$ as $t\to\infty$.) This completes the proof of the theorem. □

REFERENCES

[1] G. GRIPENBERG, *An abstract nonlinear Volterra equation*, Israel J. Math., 34 (1979), pp. 198–212.

[2] R. A. HANDELSMAN AND W. E. OLMSTEAD, *Asymptotic solution to a class of nonlinear Volterra integral equations*, SIAM J. Appl. Math., 22 (1972), pp. 373–384.

[3] M. E. H. ISMAIL AND C. P. MAY, *Special functions, infinite divisibility and transcendental equations*, Math. Proc. Camb. Phil. Soc., 85 (1979), pp. 453–464.

[4] J. B. KELLER AND W. E. OLMSTEAD, *Temperature of a nonlinearly radiating semi-infinite solid*, Quart. Appl. Math., 29 (1972), pp. 559–566.
[5] N. LEVINSON, *A nonlinear Volterra equation arising in the theory of superfluidity*, J. Math. Anal. Appl., 1 (1960), pp. 1–11.
[6] W. R., MANN AND F. WOLF, *Heat transfer between solids and gases under nonlinear boundary conditions*, Quart. Appl. Math., 9 (1951), pp. 163–184.
[7] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
[8] W. E. OLMSTEAD AND R. A. HANDELSMAN, *Asymptotic solution to a class of nolinear Volterra integral equations*, II, SIAM J. Appl. Math., 30 (1976), pp. 180–189.
[9] ———, *Diffusion in a semi-infinite region with nonlinear surface dissipation*, SIAM Rev., 18 (1976), pp. 275–291.
[10] W. E. OLMSTEAD, *A nonlinear integral equation associated with gas absorption in a liquid*, Z. Angew. Math. Phys., 28 (1977), pp. 513–523.
[11] K. PADMAVALLY, *On a non-linear integral equation*, J. Math. Mech., 7 (1958), pp. 533–555.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0011/$01.00/0

COMPARISON AND STABILITY THEOREMS FOR REACTION-DIFFUSION SYSTEMS*

ROBERT A. GARDNER†

Abstract. This paper extends a comparison technique of Conway and Smoller [Comm. Part. Diff. Eqns., 2 (1977) pp. 679–697] for systems of n reaction-diffusion equations. By altering the definition of the comparison system we obtain 2^n (rather than two) spatially homogeneous comparison vectors. The existence of additional comparison vectors is useful in obtaining a more precise description of the asymptotic behavior of solutions. In particular, we study a few examples in which the above extension enables us to give a description of (1), the domains of attraction of rest points of a system arising in mathematical ecology, and (2), a threshold effect for a system arising in chemical reactor theory.

The second part of this paper relates the (diffusion-independent) domains of attraction (R) of constant rest states (P) which are obtained via the above comparison technique, to the diffusion-dependent stability results obtainable by energy estimates, for the Neumann problem on a bounded domain. In particular, suppose that λ is the measure of the set of values of x for which the initial data lie outside R, and that d is the minimum diffusion rate; if the space average of the initial data lies in R, and if (roughly) $\lambda d^{-4} < K$, where K is a positive constant, then the solution of the reaction-diffusion system must tend uniformly to P as t approaches infinity. Applications to mathematical ecology and mathematical neurophysiology are discussed.

1. Introduction. In this paper we extend a comparison technique for systems of reaction-diffusion equations which was introduced by Conway and Smoller in [6]. This technique consists of finding a maximal and minimal comparison system, the solutions of which provide spatially homogeneous lower and upper bonds, $V^{\pm}(t)$, for the solution $U(x, t)$ of the original reaction-diffusion system; that is,

$$V^{-}(t) \leqq U(x, t) \leqq V^{+}(t).$$

(We say that $A \leqq B$ where A and B are two vectors, if the inequality holds componentwise.) By introducing a simple modification in the definition of the comparison system, we obtain a comparison vector $V(t)$, the components of which satisfy either $v_i(t) \geqq u_i(x, t)$ or $v_i(t) \leqq u_i(x, t)$ depending on the value of i; u_i and v_i are the components of U and V, respectively. In this manner, we obtain 2^n distinct comparison vectors; when the same inequality holds in each component, $1 \leqq i \leqq n$, we recover the vectors $V^{\pm}(t)$ obtained by Conway and Smoller. The existence of additional comparison vectors allows us, in certain cases, to obtain more detailed information about the asymptotic behavior of solutions than would otherwise be possible with only the two vectors, $V^{\pm}(t)$. Applications to mathematical ecology and to chemical reactor theory are discussed.

The above comparison theorem is useful in studying the stability of equilibria and their domains of attraction. Through the application of this theorem, it is sometimes possible to find a (rectangular) region R in the phase plane containing a constant equilibrium P with the property that if the initial data have values in R, then the solution $U(x, t)$ must tend uniformly to P as t approaches infinity. Such results are independent of the diffusion rates, and, for the Neumann problem on a bounded domain, of the volume of the domain in x-space. In a different paper, [5], Conway, Hoff, and Smoller showed for the Neumann problem, by choosing the diffusion rates sufficiently large, (while keeping other parameters fixed), that the solution $U(x, t)$ tends uniformly to its space average, $\bar{U}(t) = |\Omega|^{-1} \int_{\Omega} U(x, t)\, dx$; here, $|\Omega|$ is the volume of Ω. Moreover, with

* Received by the editors December 4, 1979, and in revised form September 16, 1980. This work was supported by a grant from the Science Research Council of Great Britain.

† Department of Mathematics, University of Massachusetts, Amherst, Massachusetts 01003.

notation as above, their results imply that if $\|\nabla_{(x)}U(\cdot\,,0)\|_L\infty$ is small, then $U(x,t)$ tends uniformly to p provided that $\bar{U}(0)\in R$ and that diffusion is sufficiently strong. In this paper, we show that the restriction on the gradient of the initial data can be replaced with a much weaker condition. This condition provides a continuous transition between the uniform, diffusion-independent stability of p with respect to R, (obtained by maximum principle techniques) and the diffusion-dependent stability obtained by energy estimates. More precisely, let d be the minimum diffusion rate and let λ be the measure of the set of values of x for which $U(0,x)$ has values outside of a slightly smaller rectangle, R_ε, where $P\in R_\varepsilon\subsetneq R$. We find a continuous function $\rho(\lambda,d)\geqq 0$ such that if (roughly) $\bar{U}\in R_\varepsilon$ and if $\rho(\lambda,d)<K$, where K is a positive constant independent of d and λ, then $U(x,t)$ approaches P as $t\to\infty$. The function ρ has the property that $\rho(0,d)=\rho(\lambda,\infty)=0$, and is roughly of the form λd^{-4}. We give applications to systems arising in mathematical ecology and mathematical neurophysiology.

Related results have been obtained for the Cauchy problem for a single equation by Aronson and Weinberger [2] and by Chafee [3]. Gardner [11], has proved a related result for the Dirichlet problem on a bounded domain.

The plan for the remainder of the paper is as follows. In § 2 we prove the comparison theorem, and we apply it in § 3 to a few examples; § 4 contains the stability theorem, and finally, § 5 gives a few more applications.

2. The comparison theorem. We shall consider systems of the form

$$U_t = LU + F(U,x,t), \qquad (x,t)\in\Omega\times\mathbb{R}_+, \tag{1}$$

where $\Omega\subseteq\mathbb{R}^m$, $U=(u_1,\cdots,u_n)$, $LU=(L_1u_1,\cdots,L_nu_n)$, with

$$L_ku=\sum_{i,j}a_{ij}^k(x,t)u_{x_ix_j}+\sum_j b_j^k(x,t)u_{x_j}$$

and

$$F(U,x,t)=(f_1(U,x,t),\cdots,f_n(U,x,t)).$$

We shall assume that the coefficients of L are bounded and continuous, and that each L_k is uniformly elliptic. The components of F are assumed to be at least bounded and continuous, with bounded continuous derivatives, on $\Sigma\times\Omega\times\mathbb{R}_+$. The set Ω is either $\mathbb{R}^m$ or a bounded domain in $\mathbb{R}^m$ with C^k boundary, where $k=2[m/4]+2$. If $\Omega=\mathbb{R}^m$ we prescribe continuous Cauchy data

$$U(x,0)=U^0(x), \tag{2a}$$

and if Ω is a bounded domain we suppose in addition to (2a) that

$$\partial U/\partial n = n\cdot\nabla_{(x)}U=0, \qquad (x,t)\in\partial\Omega\times\mathbb{R}_+, \tag{2b}$$

where n is the outward unit normal to $\partial\Omega$.

We shall assume that the interactive term F possesses a bounded invariant set Σ of the form

$$\Sigma=\prod_{i=1}^n[a_i,b_i],$$

where $a_i<b_i$, that is if $U^0(x)\in\Sigma$ for all $x\in\Omega$, then $U(x,t)\in\Sigma$ for all $x\in\Omega$ and $t\geqq 0$ for which $U(x,t)$ is defined. This will be the case if

$$F(U,x,t)\cdot\nu(U)\leqq 0 \tag{3}$$

for all $U \in \partial\Sigma$, where $\nu(U)$ is an outward pointing normal to $\partial\Sigma$. Under these hypotheses it follows that there exists a unique, smooth solution of (1), (2) with values in Σ, defined for all $t \geqq 0$; (cf. [9, Chapt. 5]).

All function space norms are taken with respect to Ω. To simplify notation, we shall, for example, denote $\|\cdot\|_{L^2(\Omega)}$ simply by $\|\cdot\|_{L^2}$.

We shall now define our comparison system. Suppose that $\mathcal{R}_m \cup \mathcal{R}_M = \{1, 2, \cdots, n\}$ and that $\mathcal{R}_m \cap \mathcal{R}_M = \varnothing$. Put

$$f_i^+(U) = \sup\{f_i(\xi_1, \cdots, \xi_{i-1}, u_i, \xi_{i+1}, \cdots, \xi_n, x, t): x \in \Omega, t \geqq 0, a_j \leqq \xi_j \leqq u_j \text{ if } j \in \mathcal{R}_M, j \neq i, \text{ and } u_j \leqq \xi_j \leqq b_j \text{ if } j \in \mathcal{R}_m, j \neq i\}, \tag{4}$$

$$f_i^-(U) = \inf\{f_i(\xi_i, \cdots, \xi_{i-1}, u_i, \xi_{i+1}, \cdots, \xi_n, x, t): x \in \Omega, t \geqq 0, a_j \leqq \xi_j \leqq u_j \text{ if } j \in \mathcal{R}_M, j \neq i, \text{ and } u_j \leqq \xi_j \leqq b_j \text{ if } j \in \mathcal{R}_m, j \neq i\}. \tag{5}$$

Now let $H(V) = (h_1(V), \cdots, h_n(V))$, where $h_i(V) = f_i^+(V)$ if $i \in \mathcal{R}_M$ and $h_i(V) = f_i^-(V)$ if $i \in \mathcal{R}_m$. Our comparison vector $V(x, t)$ is defined to be the solution of

$$V_t = LV + H(V), \qquad V(x, 0) = v_1^0(x), \cdots, v_n^0(x)), \tag{6}$$

where $v_i^0(x) \leqq u_i^0(x)$ if $i \in R_m$ and $v_i^0(x) \geqq u_i^0(x)$ if $i \in \mathcal{R}_M$. For example, let v_i^0 be the supremum or infimum of $u_i^0(x)$; in this case $LV \equiv 0$ so that V satisfies an autonomous ordinary differential equation (o.d.e.).

THEOREM 1. *Let $U(x, t)$ be the solution of* (1), (2) *with values in Σ, and let $V(x, t)$ be the solution of* (6) (*resp.* (6), (2b)) *if Ω is all of $\mathbb{R}^m$* (*resp. a bounded domain*). *Then*

$$v_i(x, t) \geqq u_i(x, t) \quad \text{if } i \in R_M,$$

$$v_i(x, t) \leqq u_i(x, t) \quad \text{if } i \in \mathcal{R}_m,$$

for all $x \in \Omega$ and $t \geqq 0$.

The proof follows, with minor modifications, from that of [6, Thm. 1]. Indeed, set $W = (w_1, \cdots, w_n)$ and $G = (g_1, \cdots, g_n)$, where

$$(w_i, g_i) = \begin{cases} (u_i - v_i, f_i(u, x, t) - h_i(V)) & \text{if } i \in \mathcal{R}_M, \\ (v_i - u_i, h_i(V) - f_i(U, x, t)) & \text{if } i \in \mathcal{R}_m. \end{cases}$$

We then have that

$$W_t = LW + G(W, x, t) \qquad W(x, 0) \leqq 0, \tag{7}$$

and that $\{W \leqq 0\}$ is invariant under the flow of (7). □

3. Examples.

A. Mathematical ecology. Let u and v be the population densities of two competing species. We assume that u and v will satisfy a system of the form

$$\begin{aligned} u_t &= d_1 \Delta u M(u, v), \\ v_t &= d_2 \Delta v + v N(u, v), \end{aligned} \tag{8}$$

together with (2); M and N are the local per-capita growth rates of u and v respectively, where M and N have the qualitative properties shown in Fig. 1 (see [7]); in particular, $M_v < 0$ and $M_u < 0$. The comparison theorem does not yield any useful information if $\mathcal{R}_M = \varnothing$; if $\mathcal{R}_m = \varnothing$ (cf. [7]), it can be seen that lim sup $u(x, t) \leqq b$ and lim sup $v(x, t) \leqq d$. If $\mathcal{R}_M = \{1\}$ then $H = (f_1^-, f_2^+) = (uM, vN)$. Suppose that the data (inf $u(x, 0)$,

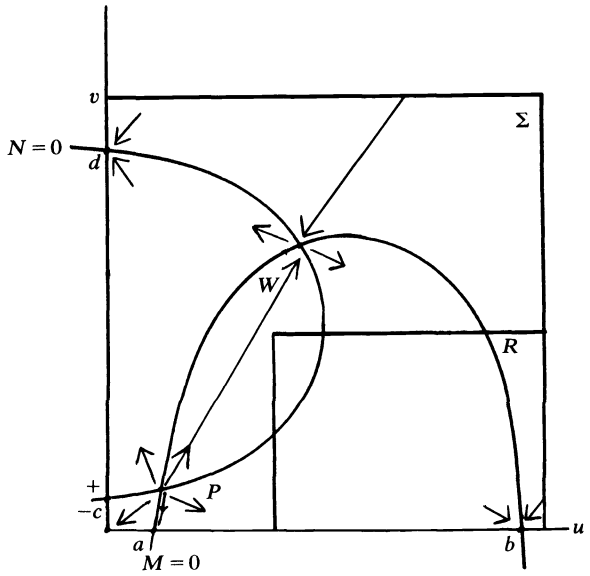


FIG. 1

sup $v(x, 0)$) of the comparison system lie beneath the separatrices connecting $(a, 0)$, P, W, and ∞ for the flow indicated in Fig. 1; that is, the data of (8) have values in the rectangle R indicated in Fig. 1. It follows that $\liminf u(x, t) \geqq b$ and that $\limsup v(x, t) = 0$. Thus the rest point $(b, 0)$ is stable with respect to any data which lie in R; similar remarks apply to the rest point $(0, d)$.

B. Establishment of a chemical process. Gel'fand, in [12, § 16], has considered the following system:

$$
\begin{aligned}
a_t &= D\Delta a - lag + 2kah, & a(0, x) &= a_0(x), \\
g_t &= \varepsilon_1 \Delta g - lag, & g(0, x) &= g_0(x), \\
h_t &= \varepsilon_2 \Delta h + lag - kah, & h(0, x) &\equiv 0,
\end{aligned} \tag{9}
$$

where l, k, and D are positive constants, $\varepsilon_i \geqq 0$, $i = 1, 2$, and $\Omega = \mathbb{R}^m$. Our remarks also apply to the case where Ω is a bounded domain. In this case, the boundary conditions (2b) are now prescribed for the variables which diffuse; that is, if $\varepsilon_1 = \varepsilon_2 = 0$, then (2b) is only appropriate for the variable a. Here, a is the concentration of atomic hydrogen, g is the concentration of oxygen, and h is the concentration of an autocatalyst. Over a certain range of temperatures, g and h diffuse more slowly than does a, so that it is of interest to study (9) when $\varepsilon_1 = \varepsilon_2 = 0$. The catalyst is produced as a product of the combustion of oxygen with atomic hydrogen, and it in turn induces a reaction which liberates more atomic hydrogen.

The following threshold effect can be observed. If $a_0(x)$ is sufficiently small, the reaction damps out; that is, $a(x, \infty) \equiv 0$ whereas $g(x, \infty) > g_0 > 0$ for all $x \in \Omega$. The reaction creates a burned out "crater" in the oxygen. For larger $a_0(x)$, the reaction is induced throughout Ω; in this case the oxygen burns out completely, so that $g(x, \infty) \equiv 0$.

Let the nonlinear terms in (9) be f_1, f_2, and f_3. We shall consider the comparison system with nonlinear terms f_1^+, f_2^-, and f_3^+, defined with respect to the invariant set for (9) found in [4], namely,

$$\Sigma = \{(a, g, h): a \geqq 0, 0 \leqq g \leqq B, 0 \leqq h \leqq C\},$$

where $C > lB/k$. A simple computation shows that

$$f_1^+(a, g, h) = -lag + 2kah,$$

$$f_2^-(a, g, h) = -lag,$$

$$f_3^+(a, g, h) = \begin{cases} laB - kah & \text{if } lB - kh > 0, \\ 0 & \text{if } lb - kh \leqq 0, \end{cases}$$

so that even though Σ is not bounded, it is still possible to define a comparison system in the case $\mathcal{R}_M = \{1, 3\}$, $\mathcal{R}_m = \{2\}$. Let (a_+, g_-, h_+) be the solution of

$$\begin{aligned} a_+' &= f_1^+(a_+, g_-, h_+), & a_+(0) &= a_0 = \sup a_0(x), \\ g_-' &= f_2^-(a_+, g_-, h_+), & g_-(0) &= g_0 = \inf g_0(x), \\ h_+' &= f_3^+(a_+, g_-, h_+), & h_+(0) &= 0. \end{aligned} \tag{10}$$

Since Σ is not bounded (in a), it is not immediately clear that global solutions exist to either (9) or (10). However, g and h, (resp. g_- and h_+) are bounded for all x and $t \leqq T$ for which the solution of (9), (resp. (10)), exists, so that, for example, a satisfies a linear equation of the form

$$a_t = D\Delta a + b(x, t)a,$$

where b is bounded on $\Omega \times [0, T]$. This in turn implies that a is bounded on $\Omega = [0, T]$, where the bound may grow exponentially with T; thus solutions exist for all $t \geqq 0$. (For brevity, we shall not include the details of the above argument.) In the case $\varepsilon_i > 0$, $i = 1, 2$, our comparison theorem may be applied on $\Omega \times [0, T]$, where $T > 0$ is arbitrary; the comparison will therefore be valid on $\Omega \times \mathbb{R}_+$. If $\varepsilon_1 = \varepsilon_2 = 0$, then the solution of (9) on $\Omega \times [0, T]$ may be recovered as the limit of solutions of (9) with $\varepsilon_i > 0$, $i = 1, 2$, as ε_1, ε_2 approach zero. (The continuous dependence of the solution on the ε_i's is proven in [17, Thm. 2.5].)

We shall now give a condition on a_0 and g_0 which implies that $\lim_{t\to\infty} a_+(t) = 0$ and that $\lim_{t\to\infty} g_-(t) > g_1 > 0$, so that the reaction in this case is inhibited. Let $F(u) = g_0(e^{-lu} - 1) + lkBu^2$; the graph of F is indicated in Fig. 2. F has a negative minimum at a value of $u = u_*$ (a simple computation shows that $u_* > g_0/(2kB + lg_0)$). We will show that damping occurs if $a_0 < -F(u_*)$. The technique is in part borrowed from Gel'fand [11].

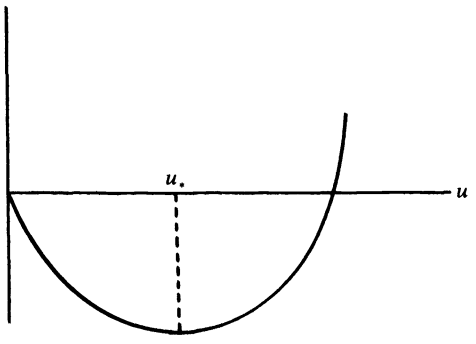


FIG. 2

The second equation in (10) can be explicitly integrated to obtain $g_- = g_0 e^{-lu}$, where $u(t) = \int_0^t a_+(s)\,ds$. The third equation in (10) can also be integrated, to obtain

$$h_+(t) = \int_0^t h'_+(s)\,ds \leqq \int_0^t la_+(s)B\,ds = lBu,$$

which, when substituted into the first equation in (10), yields

$$u'' \leqq -lg_0 u' e^{-lu} + 2klBuu' = (g_0 e^{-lu} + klBu^2)',$$

so that

$$u' \leq a_0 + F(u), \qquad u(0) = 0. \tag{11}$$

From our choice of a_0 we see that there exists a unique u_1, $0 < u_1 < u_*$ such that $a_0 + F(u) \geqq 0$ if $u \leqq u_i$. The differential inequality (11) implies that $u(t) \leqq u_1$ for all $t \geqq 0$, so that $\lim_{t\to\infty} a_+(t) = 0$ and $\lim_{t\to\infty} g_-(t) \geqq g_0 e^{-lu_1}$.

4. Domains of attraction. We shall now study the domains of attraction of a constant equilibrium solution P of (1), (2), where Ω is a bounded domain and homogeneous Neumann boundary conditions (NBC's) are prescribed. In order to simplify the discussion, we shall assume that $L_k = d_k\Delta$, where d_k is a positive constant, and that F depends only on U, although our arguments easily extend to the case where $-L_k$ is a self-adjoint uniformly elliptic operator with coefficients depending only on x, and whose smallest eigenvalue is zero. Suppose that there exists a rectangle $R \subseteq \Sigma$ of the form $R = \prod_{i=1}^n [c_i, d_i]$ such that if $U^0(x) \in R$ for all $x \in \Omega$, then $\lim_{t\to\infty} U(x, t) = P$. (It should be noted that R need not be an invariant set itself; e.g., in Fig. 1, R and $P = (b, 0)$ meet the above requirements. However, if we choose the initial data equal to the upper left hand vertex of R, we see that the solution leaves R for a brief period.) There are many examples of systems which admit such rest points. Examples arising in mathematical ecology and neurophysiology are presented in § 5.

THEOREM 2. *Suppose that $U^0(x) \in \Sigma$ for all $x \in \Omega$ and that if $\bar{U}^0$ is the space average of U^0, then $\bar{U}^0$ is in the interior of R where R and P are as above. Let $\varepsilon = \operatorname{dist}(\bar{U}^0, \partial R)$, and let*

$$\rho(\delta, d) = L(\exp(-\lambda_1 d^{1-1/k}\delta^{1/k}) + d^{(-k+1)/2}\delta^{-1/2})\delta + K(\delta^{1/2k}d^{-1/2k} + \delta^{1/k}d^{-1/k}); \tag{12}$$

here, L (cf. (14)) depends only on Ω and m; K depends only on the supremum of $|F(V)|$ over Σ, m, Ω, and $\partial\Omega$. λ_i is the smallest positive eigenvalue of $-\Delta$ on Ω with NBC's, *$k = 2[m/4]+2$, and $d = \min_i(d_i)$. Suppose that $\|U^0 - \bar{U}^0\|_{L^2} \leqq \delta$, where δ and d are chosen such that*

$$\rho(\delta, d) \leqq \min(\varepsilon/2, 1/2). \tag{13}$$

Then $\lim_{t\to\infty} |U(x, t) - P| = 0$, uniformly for $x \in \Omega$.

Remark. The expression ρ in 13 is $O(d^{-1/k}\delta^{1/2k})$ as δ approaches zero and as d approaches infinity, so that (13) is satisfied by the set of (d, δ) which lie beneath some hyperbola $d^{-2}\delta =$ constant. It is important to note that there is no restriction on ∇u^0 so that it is possible to choose very rough initial data.

Proof. The main idea is to show that diffusion smooths out rough peaks in the data before the nonlinear effects become significant, so that by the time the nonlinearity becomes the dominant term, the solution has values uniformly in R.

To simplify notation we shall omit subscripts; for example, u_i will be denoted by u. The variable u satisfies a scalar equation of the form

$$u_t = d\Delta u + g(x, t), \qquad u(0, x) = u^0(x), \qquad \text{NBC's},$$

where $g(x, t) = f(U(x, t))$. We now let a and b be the solutions of

$$a_t = d\Delta a, \qquad a(x, 0) = u^0(x), \qquad \text{NBC's},$$

$$b_t = d\Delta b + g(x, t), \qquad b(x, 0) \equiv 0, \qquad \text{NBC's}.$$

Let $0 = \lambda_0 < \lambda_1 \leqq \lambda_2 \cdots$ be the eigenvalues of $-\Delta$ on Ω with NBC's, and let $\phi_0, \phi_1, \cdots$ be a corresponding set of orthonormal eigenfunctions (with one exception: we let $\phi_0 = |\Omega|^{-1}$). We may express a as

$$a(x, t) = \sum_{k \geqq 0} e^{-d\lambda_k t} u_k \phi_k,$$

where $u_k = (\phi_k, u^0)_{L^2}$. Thus $\bar{u}^0 = u_0$, and

$$\begin{aligned} \|d^j \Delta^j a(T)\|_{L^2}^2 &= \sum_{k \geqq 1} (d\lambda_k)^{2j} e^{-2d\lambda_k T} u_k^2 \\ &\leqq (\sup_{l \geqq 0} l^{2j} e^{-2lT}) \|u^0 - \bar{u}^0\|_{L^2}^2 \\ &\leqq \left(\frac{j}{T}\right)^{2j} e^{-2j} \delta^2, \end{aligned}$$

so that if $\bar{a}(t) = |\Omega|^{-1} \int_\Omega a(x, t)\, dx$, then $\bar{a}(t) \equiv \bar{u}^0$, and

$$\|\Delta^j (a - \bar{u}^0)(T)\|_{L^2}^2 \leqq \frac{\gamma(j)^2 \delta^2}{(dT)^{2j}},$$

where $\gamma(j) > 0$ is a positive constant. We also have that

$$\|(a - \bar{u}^0)(T)\|_{L^2}^2 \leqq \delta^2 e^{-2\lambda_1 dT},$$

so that if $\|v\|_j^2 = \|v\|_{L^2}^2 + \|\Delta^j v\|_{L^2}^2$ we have that

$$\|(a - \bar{u}^0)(T)\|_j \leqq \left(e^{-2\lambda_1 dT} + \frac{\gamma(j)^2}{(dT)^{2j}}\right)^{1/2} \delta.$$

Now Δ^j is an elliptic operator of order $2j$, so by the estimates of Agmon, Douglis, and Nirenberg [1, Thm. 15.2], there exists a constant M depending only on Ω (where Ω is of class $2j$) and m, such that

$$\|v\|_{H^{2j}(\Omega)} \leqq M(\|v\|_{L^2}^2 + \|\Delta^j v\|_{L^2}^2)^{1/2} = M\|v\|_j.$$

By the Sobolev embedding theorem, there exists a constant $J > 0$, depending only on Ω and m, such that

$$\|v\|_{L^\infty} \leqq J\|v\|_{H^{2j}},$$

provided that $j > m/4$. Thus if $j = [m/4] + 1$, we have that

$$\|(a - \bar{u}^0)(T)\|_{L^\infty} \leqq L(e^{-d\lambda_1 T} + (dT)^{-j})\delta, \tag{14}$$

where $L = MJ \max(1, \gamma(j))$.

We now show that

$$|b(x, T)| \leqq K[(T/d)^{1/2} + T] \tag{15}$$

as T approaches zero; here $K > 0$ depends only on f, Ω, and m. Let $\Gamma(x, t) = (4\pi\, dt)^{n/2} \exp(-|x|^2/4td)$. Then from [10, chapt. 5, (3.5), (3.6), and (3.8)] we have that

$$b(x, t) = \int_0^t \int_{\partial\Omega} \Gamma(x-\xi, t-s)\phi(\xi, s)\, d\xi\, ds + \int_0^t \int_{\Omega} \Gamma(x-\xi, t-s) g(\xi, s)\, d\xi\, ds. \tag{16}$$

The density ϕ is defined on $\partial\Omega \times \mathbb{R}_+$ and is the solution of

$$\phi(x, t) = 2\int_0^t \int_{\partial\Omega} \frac{\partial\Gamma}{\partial n_x}(x-\xi, t-s)\phi(\xi, s)\, d\xi\, ds + 2G(x, t),$$

where

$$G(x, t) = \int_0^t \int_{\Omega} \frac{\partial\Gamma}{\partial n_x}(x-\xi, t-s) g(\xi, s)\, d\xi\, ds, \qquad x \in \partial\Omega$$

and where n_x is the inward unit normal to $\partial\Omega$ at x. If $K_1 = \sup\{|f_i(U)| : U \in \Sigma\}$, we have that

$$\begin{aligned} |G(x, t)| &\leqq K_1 \int_0^t \int_{\Omega} |x-\xi| (2sd)^{-1} (4\pi sd)^{m/2} \exp\left(\frac{-|x-\xi|}{4sd}\right) d\xi\, ds \\ &\leqq K_1 \pi^{m/2} d^{-1/2} \int_0^t \int_{\mathbb{R}^m} s^{-1/2} |\eta| e^{-|\eta|^2}\, d\eta \\ &\leqq K_2 \left(\frac{t}{d}\right)^{1/2}, \end{aligned}$$

when K_2 depends only on K_1 and m, and where $\eta = (x-\xi)/2(sd)^{1/2}$. We claim that ϕ satisfies a similar inequality. Suppose that $\partial\Omega$ is at least C^2, so that there exists a constant K_3 depending only on Ω such that

$$|n_x \cdot (x-\xi)| < K_3 |x-\xi|^2$$

for all $x, \xi \in \partial\Omega$. Let

$$\Phi(t) = \sup\{|\phi(x, s)| : x \in \Omega, 0 \leqq s \leqq t\};$$

(by [10, Chapt. 5, Thm. 2], ϕ is bounded and continuous). Then, if $G_{t,x} = (x - \partial\Omega)/(4\, dt)^{1/2}$, we have that

$$\begin{aligned} |\phi(x, t)| &\leqq \Phi(t) K_3 \int_0^t \int_{\partial\Omega} |x-\xi|^2 (2sd)^{-1} \Gamma(x-\xi, s)\, d\sigma_{(\xi)}\, ds + 2|G(x, t)| \\ &\leqq \Phi(t) K_3 \int_0^t \int_{G_{s,x}} |\eta|^2 e^{-|\eta|^2} (4\pi sd)^{-1/2}\, d\sigma_{(\eta)} ds + 2K_2 \left(\frac{t}{d}\right)^{1/2}, \end{aligned}$$

where $d\sigma_{(\xi)}$ and $d\sigma_{(\eta)}$ are surface measure on $\partial\Omega$ and $G_{t,x}$, respectively. It is easily seen that there exists a constant $K_4 > 0$ depending only on m and $\partial\Omega$ such that

$$\left| \int_{G_{s,x}} |\eta| e^{-|\eta|^2}\, d\sigma_{(\eta)} \right| < K_4$$

for all $s > 0$ and for all $x \in \partial\Omega$. For example, if we fix x, we may select a system of local

coordinates for $\partial\Omega$ such that x is contained in exactly one coordinate patch P_0. If the remaining patches are $P_1, \cdots, P_r$, and if $\bar{P}_{i,s}$ is the image of P_i under the transformation $\xi \to \eta$, we have that

$$\lim_{s\to 0} \int_{\bar{P}_{i,s}} |\eta|^2 e^{-|\eta|^2} d\sigma_{(\eta)} = \begin{cases} K\int_{\mathbb{R}^{m-1}} |\eta|^2 e^{-|\eta|^2} d\eta_1 \cdots d\eta_{n-1}, & i=0, \\ 0, & i>0, \end{cases}$$

where K depends only on $\partial\Omega$. We therefore have that

$$|\phi(x,t)| \leqq \Phi(t) K_5 \left(\frac{t}{d}\right)^{1/2} + 2K_2 \left(\frac{t}{d}\right)^{1/2},$$

where K_5 depends only on m and Ω. Since the right-hand side of this inequality is increasing in t, we have that

$$\Phi(t) \leqq \Phi(t) K_5 \left(\frac{t}{d}\right)^{1/2} + 2K \left(\frac{t}{d}\right)^{1/2}.$$

Now suppose that

$$K_5 \left(\frac{t}{d}\right)^{1/2} \leqq \frac{1}{2} \tag{17}$$

(this will be the case if $t \leqq (\delta/d)^{1/k}$ and if (13) holds). We then have that

$$\Phi(t) \leqq 4K_2 \left(\frac{t}{d}\right)^{1/2}. \tag{18}$$

By an argument similar to the above and by (17), we have that

$$\int_0^t \int_{\partial\Omega} \Gamma(x-\xi, t-s)\, d\xi\, ds \leqq K_6, \tag{19}$$

where K_6 depends only on m and Ω.

Finally, we note that

$$\int_0^t \int_\Omega \Gamma(x-\xi, t-s) g(\xi, s)\, d\xi\, ds \leqq K\pi^{-m/2} \int_0^t \int_{\mathbb{R}^m} e^{-|\eta|^2} d\eta\, ds$$

$$\leqq K_1 t,$$

so that from (16), (18), (19), and the above, we obtain (15).

We now evaluate (14) and (15) at $T_0 = (\delta/d)^{1/k}$, where $k = 2j$. Since $u = a + b = (a - \bar{u}^0) + \bar{u}^0 + b$, by the definition (12) of $\rho(\delta, d)$, by hypothesis (13), and by (14) and (15), we have that

$$\|(u - \bar{u}_0)(T_0)\|_{L^\infty} \leqq \|(a - \bar{u}_0)(T_0)\|_{L^\infty} + \|b(T_0)\|_{L^\infty}$$

$$\leqq \rho(\delta, d) \leqq \frac{\varepsilon}{2};$$

hence $U(x, T_0) \in R$ for all $x \in \Omega$. □

The hypothesis of Theorem 2 which requires $\|U_0 - \bar{U}_0\|_{L^2}$ to be small is quite restrictive. This hypothesis can be substantially weakened; to do so, we make use of the comparison functions discussed earlier. To begin with, we assume that some vertex Q of the rectangle R coincides with a vertex of Σ, so that the components q_i of Q are equal

to either a_i or b_i. Let $\mathscr{R}_m$ be the set of indices such that $q_i = b_i$, and let $\mathscr{R}_M = \{1, \cdots, n\} \backslash \mathscr{R}_m$. Let R_ε be the rectangle $\prod_{i=1}^n [e_i, f_i]$, where $[e_i, f_i] = [c_i + \varepsilon, b_i]$ if $i \in \mathscr{R}_m$ and $[e_i, f_i] = [a_i, d_i - \varepsilon]$ if $i \in \mathscr{R}_M$ (see Fig. 3). As before, we assume that R contains a rest point P with the property that if $U^0(x) \in R$ then the solution of (1), (2) tends to P.

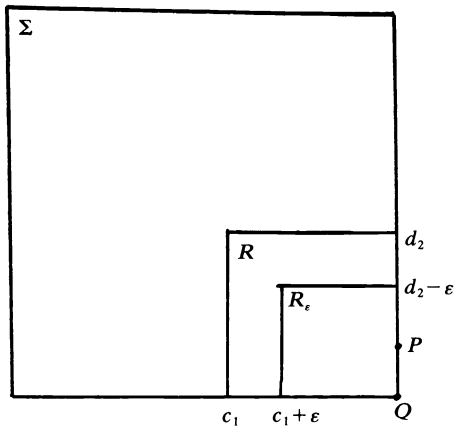


FIG. 3

THEOREM 3. *Suppose that P and R are as above and that* $U^0(x) \in \Sigma$ *for all* $x \in \Omega$. *Let* λ *be the measure of the set of values of* $x \in \Omega$ *for which* $U^0(x) \notin R_{\frac{1}{4}}$. *Suppose that*

$$\lambda \leqq \frac{\varepsilon |\Omega|}{2\kappa}, \tag{20}$$

where $\kappa = \max(\max\{c_i - a_i + \varepsilon : i \in \mathscr{R}_m\}, \max\{b_i - d_i + \varepsilon : i \in \mathscr{R}_M\})$, *and that* $\delta = [(\kappa^2 + \varepsilon\kappa/2\lambda]^{1/2}$. *If in addition, d and* λ *are chosen such that*

$$\rho(\delta, d) < \min\left(\frac{\varepsilon}{4}, \frac{1}{2}\right), \tag{21}$$

then $\lim_{t\to\infty} U(x, t) = P$, *uniformly for* $x \in \Omega$.

Remarks. (i) If, for example, $i \in \mathscr{R}_M$, (20) implies that

$$\bar{u}_i^0 - \left(d_i - \frac{\varepsilon}{2}\right) \leqq \frac{\lambda\kappa}{|\Omega|} < \varepsilon;$$

hence this condition implies that $\bar{U} \in R_{\varepsilon/2}$. (ii) Condition (21) holds for all (λ, d) which lie beneath some hyperbola, $\lambda d^{-4} = \text{constant}$.

Proof. Let $H(V)$ be the vector field constructed in § 2 with the partition of indices $\mathscr{R}_m$, $\mathscr{R}_M$ as in Theorem 3. The solution $V(x, t)$ of our comparison system will therefore satisfy

$$V_t = D\Delta V + H(V), \qquad V(x, 0) = Z(x), \qquad \text{NBC's}, \tag{22}$$

where $D = \text{diag}(d_1, \cdots, d_n)$, H is as in (6), and $z_i(x) \geqq u_i^0(x)$ if $i \in \mathscr{R}_M$, $z_i(x) \leqq u_i^0(x)$ if $i \in \mathscr{R}_m$. Suppose that

$$\omega_i = \begin{cases} \{x \in \Omega : u_i^0(x) \leqq c_i + \varepsilon\}, & i \in \mathscr{R}_m, \\ \{x \in \Omega : u_i^0(x) \geqq d_i - \varepsilon\}, & i \in \mathscr{R}_M, \end{cases}$$

and define

$$z_i(x) = \begin{cases} c_i + \varepsilon & \text{if } x \in \Omega\backslash\omega_i \quad \text{and } i \in \mathcal{R}_m, \\ d_i - \varepsilon & \text{if } x \in \Omega\backslash\omega_i \quad \text{and } i \in \mathcal{R}_M, \\ u_i^0(x) & \text{if } x \in \omega_i. \end{cases}$$

Note that Z is Lipschitz continuous and has values in Σ; this is sufficient for obtaining the existence of a solution of (22). We have that λ is the measure of $\cup_{i=1}^n \omega_i$. We will show that by choosing λ sufficiently small, the data $Z(x)$ can be made to satisfy the hypotheses of Theorem 2.

If $\bar{z}_i$ is the space average of z_i, we have that

$$\begin{aligned} d_i > \bar{z}_i > -|\Omega|^{-1}\kappa\lambda + d_i + \varepsilon, \qquad & i \in \mathcal{R}_m, \\ c_i < \bar{z}_i < |\Omega|^{-1}\kappa\lambda + c_i - \varepsilon, \qquad & i \in \mathcal{R}_M, \end{aligned} \tag{23}$$

so that if (20) holds, we see that $\bar{Z} \in R_{\varepsilon/2}$, and in fact, dist $(\bar{Z}, \partial R \backslash \partial\Sigma) > \varepsilon/2$. Moreover, for $x \in \Omega\backslash\omega_i$, we have that $|z_i(x) - \bar{z}_i(x)| < |\Omega|^{-1}\kappa\lambda$, so that

$$\begin{aligned} \|z_i - \bar{z}_i\|_{L^2}^2 &= \int_{\omega_i} (z_i - \bar{z}_i)^2 \, dx + \int_{\Omega\backslash\omega_i} (z - \bar{z}_i)^2 \, dx \\ &\leqq \kappa^2\lambda + |\Omega|^{-1}\kappa^2\lambda^2 \\ &\leqq \left(\kappa^2 + \frac{\varepsilon\kappa}{2}\right)\lambda = \delta^2. \end{aligned}$$

Thus if λ is chosen so that (21) holds we may apply Theorem 2 to the system (22); this yields a time $T_0 = (\delta/d)^{1/k}$ with the property that

$$c_i + \frac{\varepsilon}{4} < v_i(x, T_0), \qquad i \in \mathcal{R}_m,$$

$$v_i(x, T_0) < d_i - \frac{\varepsilon}{4}, \qquad i \in \mathcal{R}_M,$$

so that $V(x, T_0) \in R$ for all $x \in \Omega$. Since Σ is an invariant set, we have that

$$\begin{aligned} v_i(x, T_0) \leqq u_i(x, T_0) \leqq b_i, \qquad & i \in \mathcal{R}_m, \\ a_i \leqq u_i(x, T_0) \leqq v_i(x, T_0), \qquad & i \in \mathcal{R}_M; \end{aligned}$$

thus $U(x, T_0) \in R$ for all x. □

COROLLARY 4. *Suppose that R is contained in the interior of Σ, and that $R_\varepsilon = \prod_{i=1}^n [c_i + \varepsilon, d_i - \varepsilon]$. With notation and hypotheses as in Theorem 3, we have $\lim_{t\to\infty} U(x, t) = P$, uniformly for $x \in \Omega$.*

Proof. Let $\mathcal{R}_m = \{1, \cdots, n\}$, $\mathcal{R}_M = \varnothing$, and let $\mathcal{R}'_m = \varnothing$, $\mathcal{R}'_M = \mathcal{R}_m$. If V^- and V^+ are the solutions of the corresponding comparison systems with data constructed as in the proof of Theorem 3, we have that

$$V^-(x, t) \leqq U(x, t) \leqq V^+(x, t)$$

for all $x \in \Omega$, and that at time $T_0 = (\delta/d)^{1/k}$, $V^-(x, T_0)$ and $V^+(x, T_0)$ both lie in R for all $x \in \Omega$. □

We remark that Theorem 3 is applicable if P lies on the boundary of R, whereas Corollary 4 is applicable if P lies in the interior of R. In the former case, it is important that the vertex Q be chosen so that P lies on a face of R which meets Q, since the bulk of the data lies in R_ε. If Q is not chosen in this manner, then P will not lie in R_ε, and our theorem would not apply to data which was chosen uniformly close to P. In such cases, it may be preferable to use one of the additional comparison systems obtained in § 2 rather than V^+ or V^-.

5. Examples.

A. Mathematical ecology. We have already observed in § 3A in the case of competing species that the rest points $(0, 0)$, $(b, 0)$, and $(0, d)$ are stable with respect to data which lie in certain rectangles (see Fig. 1), so that Theorem 3 can be applied. In the case of the origin, we choose $Q = (0, 0)$, so that the nonlinearity of the comparison system of Theorem 3 is (uM^+, vN^+), whereas for the rest point $(b, 0)$ we choose $Q = (b, 0)$ where $B > b$, so that the appropriate comparison system has (uM^-, vN^+) as its nonlinear term.

Other ecological interactions can be described by systems such as (8). For example, the qualitative properties of the field (uM, vN) for the interactions of predation and symbiosis are given in Figs. 4a and 4b respectively. (See [6] and [7] for a discussion of

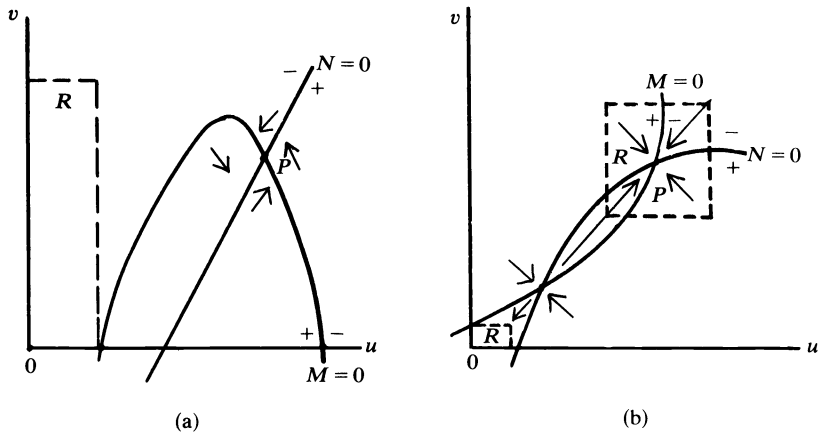


FIG. 4

these interactions.) The stability of O with respect to R in the first case and the stability of O and P with respect to the appropriate R in the second is proved in [8] and [7] respectively. Hence, Theorem 3 is applicable and shows that these rest points are stable with respect to data which on the average have values in the appropriate R.

B. Neurophysiology. A mathematical model for the propagation of a nerve pulse along an axon was proposed by Hodgkin and Huxley [13]. Fitzhugh and Nagumo [12] have proposed a simpler model whose solutions exhibit behavior which is qualitatively similar to the Hodgkin–Huxley model, namely

$$\begin{aligned} &v_t = v_{xx} + f(v) - u, \qquad v(x, 0) = v_0(x), \qquad \text{NBC's}, \\ &u_t = \varepsilon u_{xx} + \sigma v - \gamma u, \qquad u(x, 0) = u_0(x); \end{aligned} \tag{24}$$

here σ and γ are positive constants and $\varepsilon \geqq 0$. If $\varepsilon > 0$, then we prescribe NBC's for u as well. The function $f(v) = -v(a-v)(b-v)$; the phase plane is as in Fig. 5. If $-f'(0) > \sigma/\gamma$, Rauch and Smoller [16] have proved the existence of arbitrarily large and arbitrarily small invariant sets containing the origin, Σ and R respectively. Moreover,

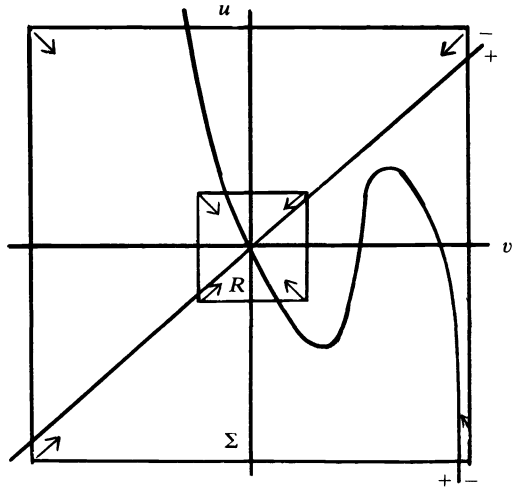


FIG. 5

they have proved the stability of the origin with respect to data which lie in R. (They have also proved stability of the origin with respect to data with small H_s-norm with $s > \frac{1}{2}$. However, such functions are also uniformly small, so this is still a local result.) If $\varepsilon > 0$, we may clearly apply Corollary 4 to (24) to obtain the stability of 0 with respect to more general data.

It is actually correct, however, to let $\varepsilon = 0$. In this case, Corollary 4 is not directly applicable; however, we can still salvage something. In particular, suppose that $R = (a, b) \times (c, d)$, and that the data for u satisfy

$$\frac{c}{2} < u_0(x) < \frac{d}{2}$$

for all $x \in \Omega$. If the data $(v_0(x), u_0(x)) \in \Sigma$ for all $x \in \Omega$, there exists $T_1 > 0$ depending only on c, d, and

$$\max \{|\sigma v - \gamma u| \colon (v, u) \in \Sigma\},$$

such that $u(x, t) \in (c, d)$ for all $t \in [0, T_1]$.

We construct comparison functions (v^-, u^-), (v^+, u^+), with data as in the proof of Theorem 3. We may now apply the techniques of the proofs of Theorems 2 and 3 to the first components v^- and v^+ of these systems. At time $T_0 = \sqrt{\delta}$ where δ is chosen in accordance with (20) and (21), we will therefore have that

$$a \leqq v^-(x, T_0) \leqq v(x, T_0) \leqq v^+(x, T_0) \leqq b$$

for all $x \in \Omega$. If δ is also chosen such that $T_0 \leqq T_1$, the solution (v, u) will have values uniformly in R at $t = T_0$. Thus in this case we may perturb the data for the first component to conclude that the origin is stable with respect to certain data with values in the horizontal strip containing R.

A similar argument can be applied to the "fast-slow" system arising in chemical reactor theory, considered in § 3 with the following minor modification. If we perturb the data for a above the value a_0 found in § 3b, we no longer have time independent bounds for a; however, we do know that $a \leqq c\, e^{kT}$ for some c, $k > 0$. Thus, the estimate (15) will now be of the form

$$|b(x, T)| \leqq K e^{kT}\left(\frac{T}{d}\right)^{1/2} + T;$$

hence, we must replace $\rho(\delta, d)$ with $e^{k(\delta/d)^{1/2}}\rho(\delta, d)$.

Acknowledgments. It is a pleasure to thank Joel Smoller and Edward Conway, who have strongly influenced me; indeed, the topic treated in § 4 was suggested in informal conversations with the former. I would also like to thank the referees for many valuable suggestions.

REFERENCES

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] D. G. ARONSON AND H. F. WEINBERGER, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[3] N. CHAFEE, *A stability analysis for a semilinear parabolic partial differential equation*, J. Diff. Equations 15 (1974), pp. 522–540.

[4] K. N. CHUEH, C. C. CONLEY AND J. A. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–391.

[5] E. CONWAY, D. HOFF, AND J. A. SMOLLER, *Large time behavior of solutions of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.

[6] E. CONWAY AND J. A. SMOLLER, *A comparison theorem for systems of reaction-diffusion equations*, Comm. Part. Diff. Eqns., 2 (1977), pp. 679–697.

[7] ———, *Diffusion and the classical ecological interactions*, Proc. Conference on Nonlinear Diffusion, Pitman, London, 1977, pp. 53–69.

[8] ———, *Diffusion and the predator-prey interaction*, SIAM J. Appl. Math., 33 (1977), pp. 673–686.

[9] P. FIFE, *Mathematical Aspects of Reacting and Diffusing Systems*, Springer, New York, 1979.

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

[11] R. A. GARDNER, *Global stability of stationary solutions of reaction-diffusion systems*, J. Diff. Equations, 37 (1980), pp. 60–69.

[12] I. M. GEL'FAND, *Some problems in the theory of quasilinear equations*, Usp. Mat. Nauk., 14 (1959), pp. 87–158; = Amer. Math. Soc. Trans. Ser. 2, 29 (1963), pp. 295–381.

[13] S. HASTINGS, *Some mathematical problems from neurobiology*, AMS Monthly, 82 (1975), pp. 881–895.

[14] A. L. HODGKIN AND A. F. HUXLEY, *A quantitative description of membrane current and its application to conduction and excitation in nerves*, J. Physiol., 117 (1952), pp. 500–544.

[15] H. J. KUIPER, *Existence and comparison theorems for nonlinear diffusion systems*, J. Math. Anal. Appl., 60 (1977), pp. 166–181.

[16] ———, *Invariant sets for nonlinear elliptic and parabolic systems*, Tech. Rep., Univ. of Wisconsin Math. Research Center, 1978.

[17] J. RAUCH AND J. A. SMOLLER, *Qualitative theory of the Fitzhugh-Nagumo equations*, Adv. Math., 27 (1978), pp. 12–44.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0012/$01.00/0

KILLING TENSORS AND NONORTHOGONAL VARIABLE SEPARATION FOR HAMILTON-JACOBI EQUATIONS*

E. G. KALNINS† AND WILLARD MILLER, JR.‡

Abstract. Every separable coordinate system for the Hamilton–Jacobi equation on a Riemannian manifold V_n corresponds to a family of $n-1$ Killing tensors in involution, but the converse is false. For general n we find a practical characterization of those involutive families of Killing tensors that correspond to variable separation, orthogonal or not.

1. Introduction. We study the separation of variables problem for the Hamilton–Jacobi equation

$$g^{ij}\partial_{x^i}W\partial_{x^j}W = E, \qquad g^{ij} = g^{ji}, \qquad 1 \leqq i, j \leqq n \tag{1.1}$$

($n \geqq 2$) and the relation between variable separation and second order Killing tensors on the (local) manifold V_n with metric tensor $\{g_{ij}\}$ in the local coordinates $\{x^i\}$. (We allow all coordinates and tensors to be complex and adopt the tensor notation in Eisenhart's book [1].)

In this paper we treat the general separation problem for (1.1), with emphasis on nonorthogonal separable coordinates. An analogous study for the more restricted orthogonal separation problem was presented in [2], and we assume familiarity with the basic definitions and results of that paper. Since every (multiplicative) separable system for the Helmholtz equation

$$\frac{1}{\sqrt{g}}\partial_{x^i}(\sqrt{g}g^{ij}\partial_{x^j}\Psi) = E\Psi, \qquad g = \det(g_{ij}) \tag{1.2}$$

is an (additively) separable system for (1.1), our treatment has direct applicability to the Helmholtz equation and the important families of special functions that arise as the separable solutions of this equation. (See [2] for a discussion of the relationship between these two equations together with additional references. The passage from (1.1) to (1.2) is closely analogous to the passage from classical mechanics to quantum mechanics.)

It is easily verified that to every separable coordinate system for (1.1), orthogonal or not, there corresponds a family of $n-1$ Killing tensors in involution. (The precise correspondence can be found in § 2.) However, not every such involutive family is associated with variable separation. In this paper we provide a solution to one of the fundamental problems in the theory of variable separation. We develop a decision procedure to determine precisely which families of Killing tensors are associated with separation, and for Killing tensors so associated we show how to construct the separable coordinates. Our procedure involves the determination of the eigenvalues and eigenforms of the Killing tensors, and is easy to implement for $n = 3$, though less so for $n \geqq 4$.

It is important for many reasons to be able to compute separable coordinates directly from Killing tensors. Indeed, for flat spaces and spaces of constant curvature all second order Killing tensors can be expressed as second order polynomials in the Killing vectors, so for such spaces the possible involutive families of Killing tensors can be constructed explicitly through the use of Lie algebra techniques and then tested for

* Received by the editors May 2, 1980.

† 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.

variable separation. Furthermore, in the Lie theory treatment of special functions which arise through separation of variables in the Helmholtz equation [3] it is the symmetry operators, not the separable coordinates, that are fundamental.

Nonorthogonal separable coordinates, though considered from the earliest days in the classical literature (see, for example [4]), have received relatively little attention in comparison with orthogonal coordinates. However, nonorthogonal separable coordinates are of very frequent occurrence for the equations of mathematical physics, in particular for the real Klein–Gordon, wave, heat and time-dependent Schrödinger equations and their Hamilton–Jacobi counterparts. The special definition of nonorthogonal separation given in § 2 is due to the authors [5], [6] and clearly exhibits the nature of the separation. (Levi-Civita's classical definition in its original form [4] is, though intuitively appealing, very inconvenient for a detailed analysis of separable coordinate types.) Independently, Benenti [7] has arrived at our same classification of coordinates, which he calls "normal separable coordinates". He proves, roughly speaking, that all separable coordinates in the sense of Levi-Civita are equivalent to normal separable coordinates. (See [7], [8] for a more detailed discussion of the classical literature.)

In § 2 we discuss our definition of variable separation for the Hamilton–Jacobi equation in some detail, and show how to construct the involutive family of Killing tensors associated with a given separable system. In § 3 we show how to check if a given coordinate system $\{x^j\}$ permits variable separation in (1.1). Our results extend the well-known test for Stäckel form in the special case of orthogonal coordinates [1]. In § 4 we present our principal result: necessary and sufficient conditions that a given involutive family of Killing tensors determines a separable coordinate system. Our Theorem 4 is much stronger than earlier such results which have appeared in the literature [1], [8], because we have explicitly proved, rather than assumed, that the basis of differential forms which appears naturally in this problem is normalizable. (Hainzl [9] has studied variable separation for linear partial differential equations of arbitrary order through use of the Stäckel method and has obtained interesting partial analogues of our Theorems 2 and 3. However, when specialized to the Helmholtz equation his definition of separability omits the possibility of type 2 and nonorthogonal ignorable coordinates.) In § 5 we present a nontrivial example of the application of our Theorem 4 to three-dimensional Minkowski space.

2. Nonorthogonal separation. Our definition of separation of variables for the H–J equation (1.1) is identical with that presented in [5], [6], [10] and is based on a division of the separable coordinates into three classes: ignorable, essential of type 1 and essential of type 2. Let $\{x^1, \cdots, x^n\}$ be a coordinate system on the manifold with metric (g^{ij}) such that the n_1 coordinates x^a, $1 \leqq a \leqq n_1$, are essential of type 1, the n_2 coordinates x^r, $n_1+1 \leqq r \leqq n_1+n_2$, are essential of type 2, and the n_3 coordinates x^α, $n_1+n_2+1 \leqq \alpha \leqq n_1+n_2+n_3 = n$, are ignorable. (In the following, indices a, b, c range from 1 to n_1, 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 the metric (g^{ij}), expressed in terms of coordinates $\{x^k\}$, is independent of the x^α, and that the separation equations take the form

$$W_a^2 + \sum_{\alpha,\beta=n_1+n_2+1}^{n} A_a^{\alpha,\beta}(x^a) W_\alpha W_\beta = \Phi_a(x^a; \lambda_1, \cdots, \lambda_{n_1+n_2}), \tag{2.1}$$

$$\sum_{\alpha=n_1+n_2+1}^{n} 2B_r^\alpha(x^r) W_r W_\alpha + \sum_{\alpha,\beta=n_1+n_2+1}^{n} C_r^{\alpha,\beta}(x^r) W_\alpha W_\beta = \Phi_r(x^r; \lambda_1, \cdots, \lambda_{n_1+n_2}), \qquad n_1+1 \leqq r \leqq n_1+n_2, \tag{2.2}$$

$$W_\alpha = \lambda_\alpha, \qquad n_1+n_2+1 \leqq \alpha \leqq n. \tag{2.3}$$

Here $A_a^{\alpha,\beta}(=A_a^{\beta,\alpha})$, B_r^α, $C_r^{\alpha,\beta}(=C_r^{\beta,\alpha})$ and Φ_i are defined and analytic in a neighborhood $N \times S \subseteq C^{n_1+n_2} \times C^{n_1+n_2}$, where N is a neighborhood of $(x_0^1, \cdots, x_0^{n_1+n_2})$ and S is a neighborhood of $(0, \cdots, 0)$ in the Euclidean space with coordinates $\lambda_1, \cdots, \lambda_{n_1+n_2}$. The parameters λ_α are arbitrary. Furthermore, the complex parameters $\lambda_1, \cdots, \lambda_{n_1+n_2}$ are independent; i.e. the Jacobian

$$\varphi(x^i, \lambda_1, \cdots, \lambda_{n_1+n_2}) = \det\left(\frac{\partial\Phi_a}{\partial\lambda_i}, \frac{\partial\Phi_r}{\partial\lambda_i}\right) \tag{2.4}$$

is nonzero in $N \times S$.

We say that the coordinates $\{x^i\}$ are *separable* for the H–J equation 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 the H–J equation

$$\sum g^{ij}\partial_i W \partial_j W = E \tag{2.5}$$

can be written in the form

$$\sum_a U_a(x^i)\Phi_a + \sum_r V_r(x^i)\Phi_r = E \tag{2.6}$$

(identically in the parameters $\lambda_1 = E, \lambda_2, \cdots, \lambda_n$), where $W = \sum_{j=1}^n W^{(j)}(x^j)$, $W_i = \partial_i W = \partial_i W^{(i)}$.

Comparison of (2.5) and (2.6) determines the functions U_a, V_r uniquely. Furthermore, differentiating (2.6) with respect to λ_l, we have

$$\sum_a U_a \frac{\partial\Phi_a}{\partial\lambda_l} + \sum_r V_r \frac{\partial\Phi_r}{\partial\lambda_l} = \delta_{1l},$$

and this leads to the usual Stäckel form

$$U_a(x^i) = \frac{\varphi^{a1}}{\varphi}, \qquad V_r(x^i) = \frac{\varphi^{r1}}{\varphi}, \tag{2.7}$$

where φ^{lm} is the (lm)-cofactor of the matrix (2.4). The nonzero components of the contravariant matric tensor are thus

$$g^{ab} = \left(\frac{\varphi^{a1}}{\varphi}\right)\delta^{ab}, \qquad g^{r\alpha} = g^{\alpha r} = \left(\frac{\varphi^{r1}}{\varphi}\right)B_r^\alpha(x^r),$$

$$\tfrac{1}{2}g^{\alpha\beta} = \sum_a A_a^{\alpha,\beta}(x^a)\frac{\varphi^{a1}}{\varphi} + \sum_r C_r^{\alpha,\beta}(x^r)\frac{\varphi^{r1}}{\varphi}, \qquad \alpha \neq \beta, \tag{2.8}$$

$$g^{\alpha\alpha} = \sum_a A_a^{\alpha,\alpha}\frac{\varphi^{a1}}{\varphi} + \sum_r C_r^{\alpha,\alpha}\frac{\varphi^{r1}}{\varphi}.$$

The generality of the functions Φ_l is illusory, due to the restrictive conditions (2.7) which require that the functions φ^{l1}/φ are independent of $\lambda_1, \cdots, \lambda_{n_1+n_2}$. Indeed, setting $\theta_{lm}(x^l) = \partial\Phi_l(x^l, \mathbf{0})/\partial\lambda_m$, $1 \leqq l,\ m \leqq n_1+n_2$ and $\theta(x^i) = \varphi(x^i, \mathbf{0})$, where $\mathbf{0} = (0, \cdots, 0) \in S$, we have

$$U_a = \frac{\theta^{a1}}{\theta} \not\equiv 0, \qquad V_r = \frac{\theta^{r1}}{\theta} \not\equiv 0. \tag{2.9}$$

Furthermore, since $\theta \not\equiv 0$ in N there exist functions $G_l(\mathbf{x}, \boldsymbol{\lambda})$, analytic in $N+S$, such that

$$\Phi_p(x^p, \boldsymbol{\lambda}) = \sum_{m=1}^{n_1+n_2} G_m(\mathbf{x}, \boldsymbol{\lambda})\theta_{pm}(x^p), \qquad 1 \leqq p \leqq n_1+n_2. \tag{2.10}$$

Substituting (2.9) and (2.10) in (2.6) we find $G_1(\mathbf{x}, \boldsymbol{\lambda}) \equiv E = \lambda_1$. Furthermore, $\partial_{\lambda_m} G_l(\mathbf{x}, \mathbf{0}) = \delta_{ml}$ and, from the fact that the minors θ^{a1}, θ^{r1} are nonzero in a neighborhood of $\mathbf{x}_0$, $\partial_{x_j} G_l(\mathbf{x}, \boldsymbol{\lambda}) \equiv 0$ for $j = 1, \cdots, n_1+n_2$, $l = 2, \cdots, n_1+n_2$. Thus, $G_l(\mathbf{x}, \boldsymbol{\lambda}) \equiv G_l(\boldsymbol{\lambda})$ and, in terms of the new parameters $E_l = G_l(\boldsymbol{\lambda})$, $l = 1, \cdots, n_1+n_2$, $E_\alpha = \lambda_\alpha$, $\alpha = n_1+n_2+1, \cdots, n$, the functions Φ_p assume the standard form

$$\Phi_p(x^p, \boldsymbol{\lambda}) \equiv \Phi_p(x^p, \mathbf{E}) = \sum_{l=1}^{n_1+n_2} E_l \theta_{pl}(x^p). \tag{2.11}$$

The separation equations (2.1)–(2.3) become

$$W_a^2 + \sum_{\alpha,\beta=n_1+n_2+1}^{n} A_a^{\alpha,\beta}(x^a) E_\alpha E_\beta = \sum_{l=1}^{n_1+n_2} E_l \theta_{al}(x^a), \tag{2.12}$$

$$\sum_{\alpha=n_1+n_2+1}^{n} 2B_r^\alpha(x^r) E_\alpha W_r + \sum_{\alpha,\beta=n_1+n_2+1}^{n} C_r^{\alpha,\beta}(x^r) E_\alpha E_\beta = \sum_{l=1}^{n_1+n_2} E_l \theta_{rl}(x^r), \tag{2.13}$$

$$W_\alpha = E_\alpha. \tag{2.14}$$

These expressions are the master equations for separation of variables in the Hamilton–Jacobi equation (2.5).

Remarks.

1) Since the metric tensor (g^{ij}) is nonsingular, $n_3 \geqq n_2$.

2) From (2.11) we have

$$\sum_{l=1}^{n_1+n_2} \frac{\theta^{lm}}{\theta} \Phi_l = E_m, \qquad m = 1, \cdots, n_1+n_2.$$

Thus,

$$\begin{aligned} &A_m(\mathbf{x}, \mathbf{p}) = E_m, \qquad m = 1, \cdots, n_1+n_2, \\ &L_\alpha(\mathbf{x}, \mathbf{p}) = E_\alpha, \qquad \alpha = n_1+n_2+1, \cdots, n, \end{aligned} \tag{2.15}$$

where

$$\begin{aligned} &A_m(\mathbf{x}, \mathbf{p}) = \sum_{i,j=1}^{n} a_{(m)}^{i,j} p_i p_j, \qquad L_\alpha(\mathbf{x}, \mathbf{p}) = p_\alpha, \\ &p_i = \partial_{x^i} W \end{aligned} \tag{2.16}$$

and the nonzero terms of the symmetric quadratic form $(a_{(m)}^{ij})$ are given by

$$\begin{aligned} &a_{(m)}^{ab} = \left(\frac{\theta^{am}}{\theta}\right) \delta^{ab}, \qquad a_{(m)}^{r\alpha} = \left(\frac{\theta^{rm}}{\theta}\right) B_r^\alpha, \\ &\tfrac{1}{2} a_{(m)}^{\alpha\beta} = \sum_c A_c^{\alpha,\beta} \frac{\theta^{cm}}{\theta} \sum_r C_r^{\alpha,\beta} \frac{\theta^{rm}}{\theta}, \qquad \alpha \neq \beta, \\ &a_{(m)}^{\alpha\alpha} = \sum_c A_c^{\alpha,\alpha} \frac{\theta^{cm}}{\theta} + \sum_r C_r^{\alpha,\alpha} \frac{\theta^{rm}}{\theta}. \end{aligned} \tag{2.17}$$

(Note that $A_1 = E_1$ is the original Hamilton–Jacobi equation.)

3) By definition, the quadratic form $H = \sum_{l=1}^{n_1+n_2} H_l^{-2} p_l^2$ is in *Stäckel form* if $H_l^2 = \theta/\theta^{l1}$, where

$$\Theta = (\theta_{lm}(x^m))$$

is a Stäckel matrix, $\theta = \det \Theta$ and θ^{l1} is the $(l, 1)$ minor of Θ. It is well known [1] that

necessary and sufficient conditions that H be in Stäckel form are

$$\partial^2_{x^j x^k} \ln H_i^2 - \partial_{x^j} \ln H_i^2 \, \partial_{x^k} \ln H_i^2 + \partial_{x^j} \ln H_i^2 \, \partial_{x^k} \ln H_j^2 + \partial_{x^k} \ln H_i^2 \, \partial_{x^j} \ln H_k^2 = 0, \qquad j \neq k \tag{2.18}$$

4) If H is in Stäckel form as in 3), the expressions $\theta^{lm}/\theta \equiv \rho_l^{(m)}\theta^{l1}/\theta = \rho_l^{(m)} H_l^{-2}$ are characterized by the equations

$$\partial_{x^k}\rho_l = (\rho_k - \rho_l)\, \partial_{x^k}(\ln H_l^{-2}), \qquad k \neq l,$$
$$\partial_{x^l}\rho_l = 0; \tag{2.19}$$

see [1]. In particular, (2.18) constitute the integrability conditions for the system (2.19), and this system admits an (n_1+n_2)-dimensional space of vector-valued solutions $(\rho_1, \cdots, \rho_{n_1+n_2})$. To any basis of solutions $(\rho_j^{(m)})$ with $\rho_j^{(1)} \equiv 1$ there corresponds a Stäckel matrix Θ with $\theta^{lm}/\theta = \rho_l^{(m)} H_l^{-2}$.

5) To understand the significance of the quadratic forms A_m and linear forms L_α (2.16), we use the natural symplectic structure on the cotangent bundle $\tilde{V}_n$ of the Riemannian manifold V_n. Corresponding to local coordinates $\{x^j\}$ on V_n we have coordinates $\{x^j, p_j\}$ on $\tilde{V}_n$. If $\{\hat{x}^k(x^j)\}$ is another local coordinate system on V_n then it corresponds to $\{\hat{x}^k, \hat{p}_k\}$ 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 the function

$$[F, G] = \partial_{x^i} F \partial_{p_i} G - \partial_{p_i} F \partial_{x^i} G. \tag{2.20}$$

(We are employing the summation convention for variables that range from 1 to n.)

It is straightforward, though tedious, to verify the relations

$$[A_l, A_m] = 0, \qquad [L_\alpha, A_l] = 0, \qquad [L_\alpha, L_\beta] = 0. \tag{2.21}$$

(For $n \leqq 4$ these relations were already noted in [5] and [6]. We will give an explicit proof for general n in § 3.) Thus, the A_m for $m \geqq 2$ are second order Killing tensors and the L_α are Killing vectors (first order Killing tensors) for the manifold V_n. Moreover, the family of $n-1$ Killing tensors $\{A_m (m \geqq 2), L_\alpha\}$ is in involution.

The relations (2.21) associating separable coordinates on V_n with an involutive family of Killing tensors are not difficult to prove. Much more difficult is the characterization of exactly those involutive families of Killing tensors that define variable separation and the development of a constructive procedure to determine the coordinates from a knowledge of the Killing tensors. For orthogonal separable coordinate systems this problem was given an elegant solution in [2]. For the more general case in which the coordinates may not be orthogonal, we provide a (less elegant) solution in the following two sections.

3. Generalized Stäckel form. Here, we are given a Riemannian manifold V_n and the contravariant metric tensor g^{ij} on V_n, expressed in terms of the local coordinates $x^1, \cdots, x^n$. We wish to determine necessary and sufficient conditions on the g^{ij} in order that the Hamilton–Jacobi equation (1.1) permit separation in these local coordinates.

If $g^{ij} = H_i^{-2}\delta^{ij}$, i.e., if the coordinates $\{x^k\}$ are orthogonal, then the necessary and sufficient condition for separation is that $H = g^{ij}p_i p_j$ be in Stäckel form [1, App. 13]. In other words, the relations (2.18) must be satisfied.

For nonorthogonal coordinates the conditions are somewhat more complicated. To derive these conditions we need some preliminary lemmas related to Stäckel form. Let $ds^2 = h_i^2 (dx^i)^2 = g_{ij}\, dy^i\, dy^j$ be a metric that is in Stäckel form with respect to the

local coordinates $y^1, \cdots, y^N$; i.e., there exists an $N \times N$ Stäckel matrix Θ such that $h_i^2 = \theta/\theta^{i1}$, where $\theta = \det \Theta$ and θ^{i1} is the $(i1)$ minor of Θ. A scalar valued function $f(\mathbf{y})$ is a *Stäckel multiplier* (*for* ds^2) if the metric $d\hat{s}^2 = f\,ds^2 = fh_i^2(dy^i)^2$ is also in Stäckel form.

LEMMA 1. *f is a Stäckel multiplier for ds^2 if and only if it satisfies the relations*

$$\partial_{y^i y^k} f + \partial_{y^i} f \partial_{y^k} \ln h_j^2 + \partial_{y^k} f \partial_{y^i} \ln h_k^2 = 0, \qquad j \neq k. \tag{3.1}$$

Proof. These relations follow directly from the fact that (2.18) must hold for $H_i^2 = h_i^2$ and also for $H_i^2 = fh_i^2$ if f is a Stäckel multiplier. □

LEMMA 2. *f is a Stäckel multiplier for ds^2 if and only if there exist local analytic functions $\varphi_l = \varphi_l(y^l)$ such that*

$$f(\mathbf{y}) = \sum_{l=1}^{N} \varphi_l(y^l) h_l^{-2}. \tag{3.2}$$

Proof. Suppose f is a Stäckel multiplier for ds^2. Then there exists a Stäckel matrix $\hat{\Theta}$ such that $fh_l^2 = \hat{\theta}/\hat{\theta}^{l1}$. But $h_l^2 = \theta/\theta^{l1}$, so $f\theta/\hat{\theta} = \theta^{l1}/\hat{\theta}^{l1}$, a function independent of y^l. Since the preceding relation holds for all l we have $f\theta/\hat{\theta} = \theta^{l1}/\hat{\theta}^{l1} = \mathcal{K} \in C$ and, without loss of generality, we can renormalize $\hat{\Theta}$ so that $\mathcal{K} = 1$. Thus,

$$f = \frac{\hat{\theta}}{\theta} = \sum_l \hat{\theta}_{l1} \frac{\hat{\theta}^{l1}}{\theta} = \sum_l \hat{\theta}_{l1} \frac{\theta^{l1}}{\theta}$$

and we obtain (3.2) with $\varphi_l = \hat{\theta}_{l1}$.

Conversely, if f can be expressed in the form (3.2), where $h_l^2 = \theta/\theta^{l1}$ is in Stäckel form, then it follows directly from (2.18) with $H_l^2 = h_l^2$ that relations (3.1) are satisfied. Hence, f is a Stäckel multiplier. □

Note that (3.2) is the general solution of (3.1).

Let (g^{ij}) be a given contravariant metric in the coordinates $x^1, \cdots x^n$. We wish to determine if these coordinates permit separation for the Hamilton–Jacobi equation. It is convenient to reorder the coordinates in a standard form. Let n_3 be the number of ignorable variables x^α (recall that x^α is ignorable if $\partial_{x^\alpha} g^{ij} = 0$ for all i, j). 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 relabel the variables so that $1 \leqq a \leqq n_1$, $n_1 + 1 \leqq r \leqq n_1 + n_2$, and $n_1 + n_2 + 1 \leqq \alpha \leqq n_1 + n_2 + n_3 = n$.

THEOREM 1. *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*

$$(g^{ij}) = \begin{array}{c} \begin{matrix} \quad n_1 \quad & \quad n_2 \quad & \quad n_3 \quad \end{matrix} \\ \left[\begin{array}{c|c|c} \delta^{ab} H_a^{-2} & 0 & 0 \\ \hline 0 & 0 & H_r^{-2} B_r^\alpha \\ \hline 0 & H_r^{-2} B_r^\alpha & g^{\alpha\beta} \end{array}\right] \end{array} \begin{matrix} n_1 \\ n_2 \\ n_3 \end{matrix}$$

where $B_r^\alpha = B_r^\alpha(x^r)$.

2) *The metric*

$$d\hat{s}^2 = \sum_{a=1}^{n_1} H_a^2 (dx^a)^2 + \sum_{r=n_1+1}^{n_1+n_2} H_r^2 (dx^r)^2$$

is in Stäckel form; i.e., relations (2.18) *hold for $1 \leqq i, j, k \leqq n_1 + n_2$.*

3) *Each $g^{\alpha\beta}(\mathbf{x})$ is a Stäckel multiplier for the metric $d\hat{s}^2$.*

Proof. The theorem follows immediately from expressions (2.8) and Lemmas 1 and 2.

Note that Theorem 1 reduces the problem of determining whether the Hamilton–Jacobi equation is separable in given coordinates to the verification of two systems of partial differential equations. If the coordinates are orthogonal, then $n_2 = 0$ and the separation requirement is simply that the metric be in Stäckel form.

Let $A = a^{ij}(\mathbf{x})p_ip_j$, $B = b^{ij}(\mathbf{x})p_ip_j$ be symmetric quadratic functions on $\tilde{V}_n$. It follows from (2.20) that these functions are in involution with respect to the Poisson bracket if and only if

$$a^{[i,j}\partial_j b^{\overline{k,l}]} = b^{[i,j}\partial_j a^{\overline{k,l}]}, \qquad 1 \leqq i, k, l \leqq n, \tag{3.3}$$

where

$$a^{[i,j}\partial_j b^{\overline{k,l}]} = a^{ij}\partial_j b^{kl} + a^{lj}\partial_j b^{ik} + a^{kj}\partial_j b^{li}.$$

A scalar-valued function $\rho(\mathbf{x})$ is a *root* of the form $a^{ij}(x)$ if

$$\det\,(a^{ij}(\mathbf{x}) - \rho(\mathbf{x})g^{ij}(\mathbf{x})) = 0 \tag{3.4}$$

in a coordinate neighborhood, where (g^{ij}) is the metric on V_n. A form $\psi = \lambda_j(\mathbf{x})\,dx^j$ such that

$$(a^{ij} - \rho g^{ij})\lambda_j = 0, \qquad \psi \neq 0$$

in the same coordinate neighborhood is an *eigenform* corresponding to the root ρ.

THEOREM 2. *Let (g^{ij}) be the contravariant metric tensor on V_n in the coordinates $\{x^i\}$. If the Hamilton–Jacobi equation is separable in these coordinates, then there exists a Q-dimensional vector space $\mathscr{A}$ of second order Killing tensors on V_n such that*

(1) $[A, B] = 0$ *for each* $A, B \in \mathscr{A}$.

(2) *For each of the n_1 essential coordinates of type 2, x^a, the form dx^a is a simultaneous eigenform for every $A \in \mathscr{A}$, with root ρ_a^A.*

(3.5) (3) *For each of the n_2 essential coordinates of type 1, x^r, the form dx^r is a simultaneous eigenform for every $A \in \mathscr{A}$, with root ρ_r^A. The root ρ_r^A has multiplicity 2 but corresponds to only one eigenform.*

(4) $\partial_i(a^{\alpha\beta}) - \rho_i^A\partial_i g^{\alpha\beta} = 0$, $i = 1, \cdots, n_1 + n_2$ *for all* $A \in \mathscr{A}$, *and all n_3 ignorable variables* $\alpha, \beta = n_1 + n_2 + 1, \cdots, n$.

(5) $[A, L_\alpha] = 0$ *for each* $A \in \mathscr{A}$ *and* $L_\alpha = p_\alpha$, $\alpha = n_1 + n_2 + 1, \cdots, n$.

(6) $Q = n + n_3(n_3 - 1)/2$.

This theorem is easily obtained from the proof of the following deeper result. Let $\{x^i\}$ be a coordinate system on V_n with coordinates divided into three classes, containing n_1, n_2, and n_3 variables respectively ($n = n_1 + n_2 + n_3$). (We will call them essential variables of type 1, essential variables of type 2 and ignorable variables, respectively, even though at this point they have nothing to do with separation.) Let $H = g^{ij}p_ip_j$.

THEOREM 3. *Suppose there exists a Q-dimensional vector space $\mathscr{A}$ of second order Killing tensors on V_n such that $H \in \mathscr{A}$ and conditions* (1)–(6) *in* (3.5), *are satisfied. Furthermore, suppose $g^{ab} = 0$ if $1 \leqq a < b \leqq n_1$ and $g^{ar} = g^{a\alpha} = g^{rs} = 0$ for $1 \leqq a \leqq n_1$, $n_1 + 1 \leqq r, s \leqq n_1 + n_2$, $n_1 + n_2 + 1 \leqq \alpha \leqq n$. Then the Hamilton–Jacobi equation* (1.1) *is separable in the coordinates $\{x^i\}$. The Killing tensors A_m, $m = 1, \cdots, n_1 + n_2$,* (2.16), *and $L_\alpha L_\beta = p_\alpha p_\beta$, $n_1 + n_2 + 1 \leqq \alpha \leqq \beta \leqq n$, form a basis for $\mathscr{A}$.*

Proof. From conditions (2), (3) and our assumptions on the vanishing of certain matrix elements of (g^{ij}), we see that the matrix corresponding to any $A \in \mathscr{A}$ takes the

form

$$(a^{ij}) = \begin{array}{c} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \left[\begin{array}{c|c|c} \delta^{ab}\rho_a H_a^{-2} & 0 & 0 \\ \hline 0 & 0 & \rho_a g^{a\alpha} \\ \hline 0 & \rho_a g^{a\alpha} & a^{\alpha\beta} \end{array}\right] \end{array} \begin{matrix} \\ n_1 \\ n_2 \\ n_3 \end{matrix} \tag{3.6}$$

If $(\rho_i^A) = (\rho_i^B)$ for $A, B \in \mathscr{A}$, it follows from (3.6) and condition (4) that $A - B$ is a linear combination of the $n_3(n_3+1)/2$ Killing tensors $L_\alpha L_\beta = p_\alpha p_\beta$, $\alpha \geqq \beta$. It follows that for each $\mathbf{x}$ the set of (n_1+n_2)-tuples $\{(\rho_i^A(\mathbf{x})), A \in \mathscr{A}\}$ spans $C^{n_1+n_2}$.

The relation $[H, A] = 0$ is equivalent to

$$g^{[i,j}\partial_j a^{\overline{k,l}]} = a^{[i,j}\partial_j g^{\overline{k,l}]}. \tag{3.7}$$

Setting $(i, k, l) = (a, b, c)$ in (3.7) and utilizing (3.6) we obtain

$$\partial_a \rho_b = (\rho_a - \rho_b)\partial_a(\ln H_b^{-2}), \qquad \partial_a \rho_a = 0. \tag{3.8}$$

Setting $(i, k, l) = (a, r, \alpha)$ in (3.7) we find

$$\partial_a \rho_r = (\rho_a - \rho_r)\partial_a \ln g^{r\alpha} \quad \text{if } g^{r\alpha} \not\equiv 0. \tag{3.9}$$

For $(i, j, k) = (r, \alpha, \beta)$ we obtain

$$g^{\beta s}g^{\alpha r}\partial_s \rho_r + g^{\beta r}g^{\alpha s}\partial_s \rho_r = (\rho_s - \rho_r)g^{\beta s}\partial_s g^{\alpha r} + (\rho_s - \rho_r)g^{\alpha s}\partial_s g^{\beta r} \qquad \text{(sum on } s\text{)}. \tag{3.10}$$

The case $(i, j, k) = (a, a, \alpha)$ leads to

$$\partial_r \rho_a = (\rho_r - \rho_a)\partial_r \ln H_a^{-2}. \tag{3.11}$$

The cases $(i, j, k) = (a, \alpha, \beta)$, (α, β, γ) are satisfied as a consequence of condition (4), and all remaining cases are satisfied identically.

Multiplying both sides of (3.10) by $g_{R\alpha}g_{S\beta}$, $(n_1+1 \leqq R, S \leqq n_1+n_2)$, and summing on α and β we find

$$\delta_R^r \partial_S \rho_R + \delta_S^r \partial_R \rho_S = (\rho_S - \rho_r)g_{R\alpha}\partial_S g^{\alpha r} + (\rho_R - \rho_r)g_{S\beta}\partial_R g^{\beta r}. \tag{3.12}$$

Setting $R = r$, $S = s$ in (3.12), solving for $\partial_s \rho_R$, substituting this result in (3.10) and equating coefficients of ρ_s, $s \neq r$, we find after some manipulation

$$\partial_r(\ln g^{\gamma s}) = \partial_r(\ln g^{\alpha s}) = \sum_\beta g_{s\beta}\partial_r(g^{\beta s}), \qquad r \neq s \tag{3.13}$$

for all α, γ such that $g^{\gamma s}$, $g^{\alpha s} \not\equiv 0$.

Since (g^{ij}) is nonsingular, for each s, $n_1+1 \leqq s \leqq n_1+n_2$, there is at least one $\alpha = \alpha(s)$ such that $g^{\alpha s} \not\equiv 0$. We define $H_s^{-2} = g^{\alpha(s)s}$. It follows from (3.9) and (3.13) that there exist functions $B_r^\gamma(x^r)$ such that

$$g^{\gamma r} = g^{r\gamma} = H_r^{-2}(\mathbf{x})B_r^\gamma(x^r), \quad n_1+1 \leqq r \leqq n_1+n_2, \quad n_1+n_2+1 \leqq \gamma \leqq n.$$

Thus, expressions (3.8), (3.9), (3.11) and (3.12) reduce to

$$\partial_i \rho_j = (\rho_i - \rho_j)\partial_j(\ln H_j^{-2}), \qquad 1 \leqq i, j \leqq n_1+n_2. \tag{3.14}$$

The integrability conditions for the system (3.14) are precisely (2.18); i.e., the metric $d\hat{s}^2 = \sum_{i=1}^{n_1+n_2} H_i^2 (dx^i)^2$ must be in Stäckel form. Similarly, the integrability requirements $\partial_i \partial_j a^{\alpha\beta} = \partial_j \partial_i a^{\alpha\beta}$ for condition (4) are (through use of (3.14)) simply that each $g^{\alpha\beta}$ be a Stäckel multiplier for the metric $d\hat{s}^2$. Thus the contravariant metric (g^{ij}) takes the form (2.8); hence the Hamilton–Jacobi equation separates in the coordinates $\mathbf{x}$. The

stated relation between the $A \in \mathscr{A}$ and the quadratic forms A_m of § 2 is provided by (2.17) and (2.19). In particular, expressions (2.17) for the $a^{\alpha\beta}$ satisfy conditions (4) and are determined by these conditions to within additive constants. □

The role of condition (4) needs clarification. It is not difficult to construct examples of Killing tensors that satisfy conditions (2), (3) and (5) but violate condition (4). However, we have

COROLLARY 1. *Let (g^{ij}) be the metric for V_n in the separable coordinates $\{x^i\}$, the coordinates ordered as in Theorems* 2 *and* 3, *and let $\mathscr{A}$ be the space of second order Killing tensors described in Theorem* 2. *Suppose C is a second order Killing tensor satisfying conditions* (2), (3) *and* (5) *of Theorem* 2 *and such that $[C, A] = 0$ for all $A \in \mathscr{A}$. Then $C \in \mathscr{A}$; i.e., C satisfies condition* (4).

Proof. Let (ρ_i^C) be the roots of C. Then there exists a $B \in \mathscr{A}$ such that $(\rho_i^C) \equiv (\rho_i^B)$. Thus, the Killing tensor $F = C - B$ has roots $\rho_i^F \equiv 0$ and takes the form $F = f^{\alpha\beta} p_\alpha p_\beta$. The condition $[F, A] = 0$ for all $A \in \mathscr{A}$ becomes

$$\rho_a^A g^{\alpha r} \partial_r (f^{\beta\gamma}) + \rho_r^A g^{\gamma r} \partial_r (f^{\alpha\beta}) + \rho_r^A g^{\beta r} \partial_r (f^{\gamma\alpha}) = 0, \qquad \partial_a f^{\alpha\beta} = 0. \tag{3.15}$$

The coefficient of ρ_r^A in (3.15) must vanish, so we have

$$g^{\alpha r} \partial_r f^{\beta\gamma} + g^{\gamma r} \partial_r f^{\alpha\beta} + g^{\beta r} \partial_r f^{\gamma\alpha} = 0 \qquad \text{(no sum on } r\text{)}. \tag{3.16}$$

(Recall that for fixed r there is at least one γ such that $g^{\gamma r} \not\equiv 0$.)

Suppose $g^{\alpha r} \not\equiv 0$. Setting $(\alpha, \beta, \gamma) = (\alpha, \alpha, \alpha)$ in (3.16) we find $g^{\alpha r} \partial_r f^{\alpha\alpha} = 0$, so that $\partial_r f^{\alpha\alpha} \equiv 0$. On the other hand, if $g^{\alpha r} \equiv 0$ but $g^{\gamma r} \not\equiv 0$, then setting (α, β, γ) in (3.16) $g^{\gamma r} \partial_r f^{\alpha\alpha} = 0$. Thus in all cases $\partial_r f^{\alpha\alpha} \equiv 0$.

If $g^{\alpha r} \not\equiv 0$, then setting $(\alpha, \beta, \gamma) = (\alpha, \beta, \alpha)$ in (3.16) we find $g^{\alpha r} \partial_r f^{\alpha\beta} = 0$, so $\partial_r f^{\alpha\beta} \equiv 0$. However, if $g^{\alpha r} \equiv 0$ but $g^{\gamma r} \not\equiv 0$, then, since $\partial_r f^{\beta\gamma} = \partial_r f^{\gamma\alpha} \equiv 0$, (3.16) becomes $g^{\gamma r} \partial_r f^{\alpha\beta} = 0$. Thus in all cases $\partial_r f^{\alpha\beta} \equiv 0$.

We have shown that $f^{\alpha\beta}$ is a constant, hence that $F = f^{\alpha\beta} p_\alpha p_\beta = f^{\alpha\beta} L_\alpha L_\beta \in \mathscr{A}$. □

Remark. It is sufficient to require that condition (4) of Theorem 2 be valid for $i = n_1 + 1, \cdots, n_1 + n_2$, since the requirement $[H, A] = 0$ for $(i, j, k) = (a, \alpha, \beta)$ yields this condition for $i = 1, \cdots, n_1$.

4. The main result. We come now to the fundamental question: given an involutive family of $n - 1$ Killing tensors, how do we determine if this family corresponds to a separable coordinate system for the Hamilton–Jacobi equation?

Let $\{x^j\}$ be a local coordinate system on the Riemannian manifold V_n and let $\theta_{(j)} = \lambda_{i(j)}\, dx^i$, $1 \leqq j \leqq n$, be a local basis of one-forms on V_n. The dual basis of vector fields is $X^{(h)} = \Lambda^{i(h)} \partial_{x^i}$, $1 \leqq h \leqq n$, where $\Lambda^{i(h)} \lambda_{i(j)} = \delta_{(j)}^{(h)}$. 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). (Equivalently, $X^{(h)} = g_{(h)}^{-1} \partial_{y^h}$.) It is classical that the forms are normalizable if and only if the coefficient of $X^{(l)}$ is zero in the expansion of $[X^{(h)}, X^{(k)}]$ in terms of the $\{X^{(j)}\}$ basis whenever $h, k \neq l$; see [1, § 35].

LEMMA 3. *The one-forms $\{\theta_{(j)}\}$ are normalizable if and only if*

$$(\partial_{x^j} \lambda_{i(l)} - \partial_{x^i} \lambda_{j(l)}) \Lambda^{i(h)} \Lambda^{j(k)} = 0, \qquad h, k \neq l. \tag{4.1}$$

This condition can also be expressed in terms of the inner products

$$G_{(h,l)} = \lambda_{(h)}^{\ i} \lambda_{i(l)}. \tag{4.2}$$

We have $\lambda_{j(l)} = G_{(h,l)} \Lambda_j^{(h)}$ or $\Lambda^{j(h)} = \lambda_{(l)}^{\ j} G^{(l,h)}$ where $G^{(h,l)} G_{(l,j)} = \delta_{(j)}^{(h)}$. Thus condition (4.1) can be written in the form

$$G^{(h,h')} G^{(k,k')} (\gamma_{(lh'k')} - \gamma_{(lk'h')}) = 0, \qquad h, k \neq l, \tag{4.3}$$

where

$$\gamma_{(lhk)} = \lambda_{i(l),j}\lambda^{i}_{(h)}\lambda^{j}_{(k)} \tag{4.4}$$

and $\lambda_{i(l),j}$ is the jth covariant derivative of $\lambda_{i(l)}$ [1]. Let $H = g^{ij}p_ip_j$.

THEOREM 4. *Suppose there exists a Q-dimensional vector space $\mathcal{A}$ of second order Killing tensors on V_n such that $H \in \mathcal{A}$ and:*

(1) $[A, B] = 0$ *for each* $A, B \in \mathcal{A}$.

(2) *There is a basis of one forms $\theta_{(h)} = \lambda_{i(h)}\,dx^i$, $1 \leqq h \leqq n$ such that*

(a) *the n_1 forms $\theta_{(a)}$, $1 \leqq a \leqq n_1$ are simultaneous eigenforms for every $A \in \mathcal{A}$ with root ρ_a^A:*

$$(a^{ij} - \rho_a^A g^{ij})\lambda_{j(a)} = 0,$$

(b) *the n_2-forms $\theta_{(r)}$, $n_1+1 \leqq r \leqq n_1+n_2$, are simultaneous eigenforms for every $A \in \mathcal{A}$ with root ρ_r^A:*

$$(a^{ij} - \rho_r^A g^{ij})\lambda_{j(r)} = 0. \tag{4.5}$$

The root ρ_r^A has multiplicity 2 but corresponds to only one eigenform.

(3) $X^{(h)}(\lambda_{i(\alpha)}a^{ij}\lambda_{j(\beta)}) = \rho_h^A X^{(h)}(\lambda_{i(\alpha)}g^{ij}\lambda_{j(\beta)})$, $h = 1, \cdots, n_1+n_2$ *for all* $A \in \mathcal{A}$ *and all* $\alpha, \beta = n_1+n_2+1, \cdots, n$.

(4) $[L_\alpha, L_\beta] = 0$ *where* $L_\alpha = \Lambda^{i(\alpha)}p_i$.

(5) $[A, L_\alpha] = 0$ *for each* $A \in \mathcal{A}$.

(6) $Q = \frac{1}{2}(2n + n_3^2 - n_3)$, *where* $n_3 = n - n_1 - n_2$.

(7) $G_{(ab)} = 0$ *if* $1 \leqq a < b \leqq n_1$, *and* $G_{(ar)} = G_{(a\alpha)} = G_{(rs)} = 0$ *for* $1 \leqq a \leqq n_1$, $n_1+1 \leqq r, s \leqq n_1+n_2$, $n_1+n_2+1 \leqq \alpha \leqq n$.

Then there exist local coordinates $\{y^i\}$ for V_n such that $\theta_{(j)} = f^{(j)}(\mathbf{y})\,dy^j$ for suitably chosen functions $f^{(j)}$, and the Hamilton–Jacobi equation is separable in these coordinates. Conversely, to every separable coordinate system $\{y^i\}$ for the Hamilton–Jacobi equation there corresponds a family $\mathcal{A}$ of second order Killing tensors on V_n with properties (1)–(7).

Proof. It is enough to show that conditions (1)–(7) imply that the one-forms $\theta_{(j)}$ are normalizable; the remainder of the proof follows immediately from Theorems 2 and 3.

From conditions (4) and (5) it follows that there exists a coordinate system $\{x^i\}$ on V_n such that $\theta_{(\alpha)} = dx^\alpha + \sum_{h=1}^{n_1+n_2}\lambda_{(\alpha)h}(x^a, x^r)\,dx^h$ and $X^{(\alpha)} = \partial_{x^\alpha}$. Clearly, conditions (4.1) hold for $h = \alpha$ and any values of k, l.

Some other conditions (4.1) follow directly from [2, proof of Theorem 5]. It follows from that proof that conditions (1), (2) and (7) imply $\gamma_{(lhk)} = \gamma_{(lkh)} = 0$ for pairwise distinct numbers l, h, k such that $1 \leqq l, h, k \leqq n_1+n_2$. Thus, (4.1) holds for $l = 1, \cdots, n_1+n_2$ and $1 \leqq h, k \leqq n_1$, $h, k \neq l$.

The remainder of the proof is essentially a systematic exploitation of condition (1) for $A, B \in \mathcal{A}$. Writing this condition in the form (3.3), multiplying by $\lambda_{(m_1)i}\lambda_{(m_2)j}\lambda_{(m_3)k}$ and summing for $i, j, k = 1, \cdots, n$, we obtain an identity $E^{A,B}_{m_1,m_2,m_3}$. This identity can be simplified through use of conditions (2) and (3). In particular, condition (2) leads to

$$\partial_u a^{jv} = -a^{jw}\Lambda^{v(z)}\partial_u\lambda_{w(z)} + \partial_u(\rho^{(z)}g^{jw})\Lambda^{v(z)}\lambda_{w(z)} + \rho^{(z)}g^{jw}\Lambda^{v(z)}\partial_u\lambda_{w(z)}, \tag{4.6}$$

where in this and the following expressions u, v, w, z range from 1 to n_1+n_2 and $\partial_i = \partial_{x^i}$. Condition (3) leads to

$$\partial_u(\lambda_{i(\alpha)}a^{ij}\lambda_{j(\beta)}) = \lambda_{u(z)}\Lambda^{v(z)}\rho^{(z)}\partial_v(\lambda_{i(\alpha)}g^{ij}\lambda_{j(\beta)}). \tag{4.7}$$

Furthermore, convariant differentiation of (4.2) leads to the relation

$$\gamma_{(hlk)} + \gamma_{(lhk)} = \lambda^{j}_{(k)} G_{(hl),j}. \tag{4.8}$$

Through use of these relations we can express the identities $E^{A,B}_{m_1,m_2,m_3}$ in terms of $\lambda_{i(u)}$, $\Lambda^{i(u)}$, $\rho^{(u)}$ and g^{ij} alone.

Let $A, B \in \mathscr{A}$ have roots $\{\rho_i\}$, $\{\mu_i\}$, respectively. Equating coefficients of $\rho_r\mu_s$ in $E^{A,B}_{c,\alpha,\beta}$ we find

$$G^{(rh')}G^{(s,k')}(\gamma_{(ch'k')} - \gamma_{(ck'h')}) = 0; \tag{4.9}$$

i.e., (4.1) holds for $(h, k, l) = (r, s, c)$. Equating coefficients of $\rho_a\mu_r$ in $E^{A,B}_{a,b,\alpha}$, $a \neq b$, we find similarly that (4.1) holds for $(h, k, l) = (r, a, c)$. Thus the forms $\theta_{(c)}$ are normalizable.

Equating coefficients of $\rho_r\mu_s$ in $E^{A,B}_{t,\alpha,\beta}$, r, s, t pairwise distinct, we verify that (4.1) holds for $(h, k, l) = (r, s, t)$. Finally, equating coefficients of ρ_r in $E^{A,H}_{a,t,\alpha}$, $r \neq t$, we verify that (4.1) holds for $(h, k, l) = (r, a, t)$. This shows that the forms $\theta_{(t)}$ are normalizable.

We see at this point that, by renormalization of $\theta_{(a)}$, $\theta_{(r)}$ if necessary, we can find local coordinates $\{y^j\}$ such that

$$\begin{gathered}\theta_{(a)} = dy^a, \quad a = 1, \cdots, n_1, \qquad \theta_{(r)} = dy^r, \quad r = n_1+1, \cdots, n_1+n_2, \\ \theta_{(\alpha)} = dy^\alpha + \sum_{h=1}^{n_1+n_2} \lambda_{h(\alpha)}\, dy^h.\end{gathered} \tag{4.10}$$

Replacing $\theta_{(\alpha)}$ by $\hat{\theta}_{(\alpha)} = \theta_{(\alpha)} - \sum_{r=n_1+1}^{n_1+n_2} \lambda_{r(\alpha)}\theta_{(r)}$, we see that the new forms $\theta_{(\alpha)}$ (dropping the hat) satisfy conditions (1)–(7), since $G_{(ar)} = 0$, and further that $\lambda_{r(\alpha)} = 0$ for the new forms. Equating coefficients of ρ_b in $E^{A,H}_{b,c,\alpha}$, $b \neq c$, we find $\partial_b\lambda_{c(\alpha)} = \partial_c\lambda_{b(\alpha)}$, and equating coefficients of $\rho_c\mu_r$ in $E^{A,B}_{c,\alpha,\beta}$ we obtain $\partial_r\lambda_{(\beta)c} = 0$. Thus $\theta_{(\alpha)} = dy^\alpha + df^\alpha$, where $df^\alpha = \sum_{c=1}^{n_1} \lambda_{c(\alpha)}\, dy^c$. Setting

$$z^h = y^h, \quad h = 1, \cdots, n_1+n_2, \qquad z^\alpha = y^\alpha + f^\alpha, \tag{4.11}$$

we have $\theta_{(j)} = dz^j$, $X^{(j)} = \partial_{z^j}$, $j = 1, \cdots, n$ and our one-forms are normalizable. □

Remark. It is sufficient to require that condition (3) of Theorem 4 be valid for $h_{n1}+1, \cdots, n_1+n_2$ since the identity $E^{A,H}_{a,\alpha,\beta}$ yields this condition for $h = 1, \cdots, n_1$. Thus condition (3) is unnecessary when $n_2 = 0$.

5. An example. To show how Theorem 4 can be employed in practice we treat a single example in some detail. The real Hamilton–Jacobi equation

$$W_t^2 - W_x^2 - W_y^2 = E \tag{5.1}$$

admits the pseudo-Euclidean algebra $e(2, 1)$ as its symmetry algebra of Killing vectors. A basis for the symmetry algebra is

$$\begin{gathered}K_1 = xp_t + tp_x, \quad K_2 = yp_t + tp_y, \quad L_3 = yp_x - xp_y \\ P_0 = p_t, \quad P_1 = p_x, \quad P_2 = p_y.\end{gathered} \tag{5.2}$$

As is well known (e.g., [2]), the space of second order Killing tensors for the pseudo-Riemannian manifold with (5.1) as its associated equation is spanned by products of Killing vectors (5.2). Thus, it is easy to display the second order Killing tensors for this manifold.

Recall that two separable coordinate systems for a Hamilton–Jacobi equation are considered as equivalent if the defining symmetry operators for the two systems are equivalent under the adjoint action of the local Lie symmetry group of the equation [5], [6], [10]. Thus, if we are looking for all separable coordinate systems with one ignorable

variable we can limit our search to those cases where the Killing vector X_α corresponding to this variable is an explicitly chosen representative of one of the conjugacy classes of one-dimensional subalgebras of $e(2, 1)$. We consider the particularly interesting case where $L_\alpha = P_0 + P_2$. (As shown in [5], all nonorthogonal separable coordinates for (5.1) correspond to this case. Moreover, it is easily shown that any coordinate system with $P_0 + P_2$ as a generator for an ignorable coordinate must necessarily be nonorthogonal [10].) For such a system the Killing tensor A must commute with L_α. Thus A can be chosen from the real vector space of homogeneous second order polynomials in the symmetries (5.2). Furthermore, we can identify two Killing tensors that lie on the same orbit under the adjoint action of the normalizer for $P_0 + P_2$. The normalizer has basis $\{K_2, L_3 - K_1, P_0 + P_2, P_1, P_0 - P_2\}$. (See [11] for a more detailed discussion of this problem.) One family of orbit representatives is

$$(5.3)\quad (L_3 - K_1)^2 + 4P_1^2 + a(L_3 - K_1)(P_0 - P_2) + b(P_0 - P_2)^2 + c(P_0 + P_2)^2 + dP_1(P_0 - P_2).$$

(That is, two such representatives lie on the same orbit if and only if they are identical. We could, of course, easily compute all possible families of orbit representatives and apply the following considerations to each such family.) Group theory can take us no further than this point. We still have to determine which, if any, of the Killing tensors (5.3) actually correspond to separable coordinates.

In the following it is convenient to choose new coordinates $\{x, \tau, w\}$ such that $\tau = \frac{1}{2}(y + t)$, $w = \frac{1}{2}(y - t)$, so $p_\tau = p_y + p_t$, $p_w = p_y - p_t$. In terms of these coordinates,

$$(5.4)\quad A - \rho H = \begin{bmatrix} 4w^2 + 4 + \rho & -2xw & -aw - \dfrac{d}{2} \\ -2xw & x^2 + c & \dfrac{ax + \rho}{2} \\ -aw - \dfrac{d}{2} & \dfrac{ax + \rho}{2} & b \end{bmatrix}.$$

Since ρ_2 is a double root, we must have $f'(\rho_2) = 0$. Also, $f(\rho) = \frac{1}{4}(\rho - \rho_2)^2(\rho - \rho_1)$. It is orthogonal, we must have $n_1 = n_2 = 1$ for any separable coordinates. Thus A must have a single root ρ_1 and a distinct root ρ_2 of multiplicity 2 which has only one eigenform. The characteristic equation $f(\rho) = \det(A - \rho H) = 0$ reads

$$(5.5)\quad \begin{aligned} &\frac{\rho^3}{4} + \rho^2\left(w^2 + 1 + \frac{ax}{2}\right) - \rho\left(bx^2 - \frac{a^2}{4}x^2 - 2ax + cb - axw^2 + dxw\right) \\ &\quad + \left(dacw - c(4b - a^2)w^2 + \left(-4b + a^2 + \frac{d^2}{4}\right)x^2 + \frac{d^2c}{4} - 4cb\right) = 0. \end{aligned}$$

Since ρ_2 is a double root, we must have $f'(\rho_2) = 0$. Also, $f(\rho) = \frac{1}{4}(\rho - \rho_2)^2(\rho - \rho_1)$. It is straightforward, though tedious, to verify that these conditions on ρ_1, ρ_2 are inconsistent unless $a = b = c = d = 0$, in which case

$$(5.6)\quad \rho_1 = -4(w^2 + 1), \qquad \rho_2 = 0.$$

Thus,

$$(5.7)\quad A = \begin{bmatrix} 4w^2 + 4 & -2xw & 0 \\ -2xw & x^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and $\theta_{(1)} = (2w^2+2)\,dx - 2xw\,dw$, $\theta_{(2)} = dw$. To satisfy conditions (5) and (7) of Theorem 4 we must require $\theta_{(3)} = xw(w^2+1)^{-1}\,dx + d\tau + f\,dw$. We choose f such that $\theta_{(3)}$ is a perfect differential and obtain

$$(5.8)\quad (\lambda_{j(k)}) = \begin{bmatrix} 2w^2+2 & 0 & \dfrac{xw}{(w^2+1)} \\ 0 & 0 & 1 \\ -2xw & 1 & \dfrac{x^2(1-w^2)}{2(1+w^2)^2} \end{bmatrix}, \qquad (\Lambda^{(l)j}) = \begin{bmatrix} \dfrac{1}{2w^2+2} & \dfrac{-xw}{2(w^2+1)^2} & 0 \\ \dfrac{xw}{w^2+1} & \dfrac{-x^2}{2(1+w^2)} & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Condition (3) can be verified directly.

Finally, $Q = 3$ and $\mathscr{A}$ has the basis $\{A, H, X^{(3)}\}$.

We conclude that among the operators (5.3) only $A = (L_3 - K_1)^2 + 4P_1^2$ corresponds to a separable coordinate system. Furthermore, in this case it is now straightforward to derive the separable coordinates. They are $\{x^1, x^2, x^3\}$, where

$$(5.9)\qquad x = x^1[1+(x^2)^2]^{1/2}, \qquad \tau = \frac{[x^3 - (x^1)^2 x^2]}{2}, \qquad w = x^2.$$

Indeed,

$$(5.10)\qquad X^{(1)} = \tfrac{1}{2}(1+w^2)^{-3/2}\partial_1, \qquad X^{(2)} = \partial_2, \qquad X^{(3)} = 2\partial_1.$$

REFERENCES

[1] L. P. EISENHART, *Riemannian Geometry*, Princeton University Press, Princeton, NJ, (2nd printing), 1949.

[2] E. G. KALNINS AND W. MILLER, Jr., *Killing tensors and variable separation for Hamilton–Jacobi and Helmoholtz equations*, this Journal, 11 (1980), pp. 1011–1026.

[3] W. MILLER, Jr., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[4] T. LEVI-CIVITA, *Sulla integrazione della equazione di Hamilton–Jacobi per separazione di variabili*, Math. Ann., 59 (1904), pp. 383–397.

[5] E. G. KALNINS AND W. MILLER, Jr., *Separable coordinates for three-dimensional complex Riemannian spaces*, J. Differential Geometry, 14 (1979), pp. 221–236.

[6] C. P. BOYER, E. G. KALNINS AND W. MILLER, Jr., *Separable coordinates for four-dimensional Riemannian spaces*, Commun. Math. Phys., 59 (1978), pp. 255–302.

[7] S. BENENTI, *Separability structures on Riemannian manifolds*, Proceedings of Conference on Differential Geometrical Methods in Mathematical Physics, Salamanca 1979, Lecture Notes in Mathematics, Springer-Verlag, to appear.

[8] S. BENENTI AND M. FRANCAVIGLIA, *The theory of separability of the Hamilton–Jacobi equation and its application to general relativity*, in General Relativity and Gravitation, Vol. 1, A. Held, ed., Plenum, New York, 1980.

[9] J. HAINZL, *Separation of variables and commuting operators*, Math. Meth. Appl. Sci., 1 (1979), pp. 468–479.

[10] E. G. KALNINS AND W. MILLER, Jr., *Nonorthogonal separable coordinate systems for the flat 4-space Helmholtz equation*, J. Phys. A. (GB), 12 (1979), pp. 1129–1147.

[11] W. MILLER, Jr., J. PATERA AND P. WINTERNITZ, *Subgroups of Lie groups and separation of variables*, J. Math. Phys. to appear.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0013 $01.00/0

INVARIANT CURVES, HOMOCLINIC POINTS, AND ERGODICITY IN AREA-PRESERVING MAPPINGS*

MARTIN BRAUN†

Abstract. In this paper we study an area-preserving transformation which arises in the study of the motion of a charged particle in the earth's magnetic field. It is shown that this mapping possesses invariant curves in one region, and is a horseshoe map in a second region. Numerical experiments are presented which indicate that the mapping is transitive in a global neighborhood of its homoclinic points.

Introduction. In this paper we present a theoretical and numerical discussion of an area-preserving mapping M of the plane into itself. This mapping arose as an "approximate" model of the global flow for the Störmer problem [2], [3], [4], which is the study of the motion of a charged particle in a magnetic dipole field, like that of the earth. Our mapping is also similar to the "linked twist" mappings studied by Bowen [1], Devaney [5] and Thurston [11])

The mapping M. Let ϕ_1 be the mapping which rotates each point (x, y) by an angle $\alpha(r)$, $r = (x^2+y^2)^{1/2}$:

$$\phi_1 : (r, \theta) \to (r, \theta + \alpha(r)).$$

The "twist" $\alpha(r)$ is a C^1 function satisfying

$$-r\alpha'(r) \to \infty \quad \text{as } r \to 0. \tag{1}$$

Let ϕ_2 be the corresponding mapping which rotates each point (x, y) about the point $(\varepsilon, 0)$; i.e., if r^+ denotes the distance of (x, y) from $(\varepsilon, 0)$ and θ^+ is the corresponding polar angle (see Fig. 1), then

$$\phi_2 : (r^+, \theta^+) \to (r^+, \theta^+ + \alpha(r^+)).$$

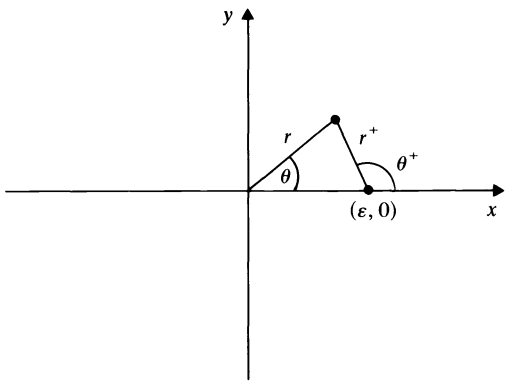


FIG. 1

* Received by the editors January 30, 1979, and in revised form October 17, 1980. This research was partially supported by the National Science Foundation under grant MCS77-03999.

† Department of Mathematics, Queens College, Flushing, New York 11367.

Our mapping M is defined to be the composition of ϕ_1 with ϕ_2, that is, $M = \phi_2 \circ \phi_1$. Clearly, M is area preserving, since it is the composition of two area-preserving transformations. (We will sometimes denote M by ϕ.)

As the distance r gets larger and larger, the radii r and r^+ approach each other. If $\alpha(r)$ is sufficiently smooth, then M has infinitely many closed invariant curves surrounding the origin. This is a direct consequence of the Moser twist theorem [8]. However, for sufficiently small r, the mapping M is very erratic. There are infinitely many unstable periodic orbits, as well as infinitely many homoclinic and heteroclinic points. This is a direct consequence of the following theorem.

THEOREM. *Given any positive integer N, there exists an $\varepsilon(N)$ such that for $0 < \varepsilon < \varepsilon(N)$ the mapping M possesses the subshift on N^2 symbols as a subsystem.*

We will prove the above theorem by constructing a suitable "rectangle" in which ϕ resembles a "horseshoe map" (Smale [10]). Our proof is based on the proof of J. Moser (personal communication) for a similar mapping (Braun [4]).

We begin by choosing two constants c_1, c_2 such that

$$\frac{1}{2} < c_1 < c_2 < \frac{3}{2}, \tag{2}$$

and consider the annuli

$$B = \{(r, \theta) | c_1\varepsilon \leqq r \leqq c_2\varepsilon\}, \qquad B^+ = \{(r, \theta) | c_1\varepsilon \leqq r^+ \leqq c_2\varepsilon\}.$$

The restriction (2) is imposed so that the annuli B and B^+ intersect in two components, and the boundaries of these components intersect under a nonzero angle (see Fig. 2).

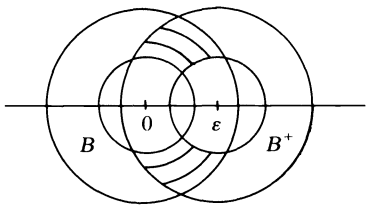


FIG. 2

Our next task is to find an annulus $A \subset B$, where the points on the inner boundary of A are rotated about the origin N times more than the points on the outer boundary of A. To this end, form the annulus

$$A_m = \{(r, \theta) | (2m-1)\pi \leqq \alpha(r) \leqq (2m+1)\pi\}$$

and set

$$A = \bigcup_{s=0}^{N-1} A_{n+s}.$$

Here, n is a large positive integer, soon to be determined. Points on the outer boundary of A are rotated by an angle $(2n-1)\pi$, while points on the inner boundary of A are rotated by an angle $(2n+2N-1)\pi$. The condition that $A \subset B$ is thus

(i) $(2n+2N-1)\pi < \alpha(c_1\varepsilon)$ and

(ii) $(2n-1)\pi > \alpha(c_2\varepsilon)$.

The inequalities (i) and (ii) can be combined into the simpler inequality

$$\alpha(c_2\varepsilon) < (2n-1)\pi < \alpha(c_1\varepsilon) - 2N\pi. \tag{3}$$

To satisfy (3) we first choose ε so small so that

$$\alpha(c_1\varepsilon) - \alpha(c_2\varepsilon) > (2N+2)\pi. \tag{4}$$

That we can choose ε sufficiently small so as to satisfy (4) is a direct consequence of (1). Then we choose an integer n large enough so that (3) holds.

Next, let A^+ be the corresponding annulus A centered about the point $(\varepsilon, 0)$. The annuli A and A^+ intersect in two "rhombus-like" regions. Let Q be the region lying in the upper-half plane. This region Q will play the role of the rectangle in the horseshoe map. It is clear from the construction of A that the image $\phi_1(Q)$ spirals N times around in A, and intersects Q in N components $\hat{U}_1, \hat{U}_2, \cdots, \hat{U}_N$ (see Fig. 3). Similarly, the

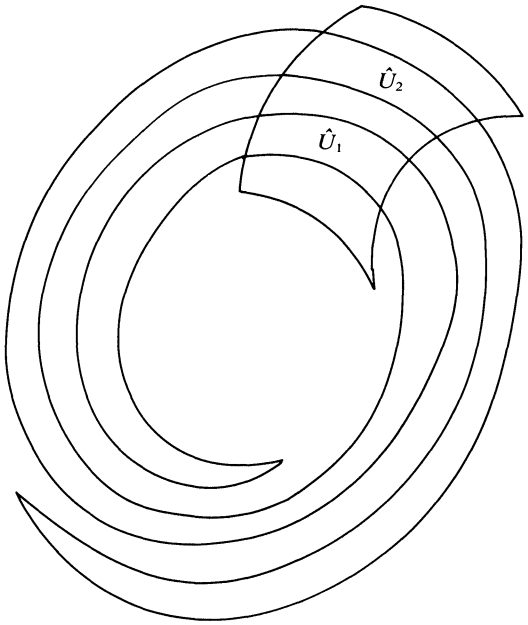


FIG. 3. *Graph of* $Q \cap \phi_1(Q)$.

image $\phi_2(\hat{U}_j)$ spirals N times around in A, and intersects Q in N components (see Fig. 4). Thus $\phi(Q)$ intersects Q in N^2 components $U_1, U_2, \cdots, U_{N^2}$ so labeled that U_1 is the outer one and U_{N^2} the inner one. Similarly, $\phi^{-1}(Q)$ intersects Q in N^2 components $V_1, V_2, \cdots, V_{N^2}$, which are so labeled that V_1 is the outer one and V_{N^2} the inner one. We then verify that

$$\phi(V_j) = U_j, \qquad j = 1, \cdots, N^2.$$

This relation follows immediately, up to the labeling, from the fact that

$$\phi\Big(\bigcup_{j=1}^{N^2} V_j\Big) = \phi(\phi^{-1}(Q) \cap Q) = Q \cap \phi(Q) = \bigcup_{j=1}^{N^2} U_j.$$

We also verify that the corresponding boundaries of V_j go into the corresponding boundaries of U_j, as required in [9].

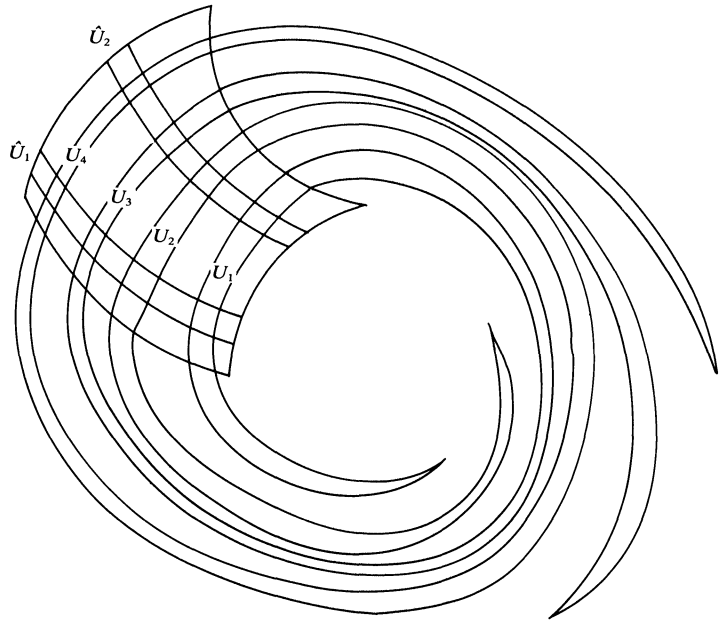


FIG. 4. $(U_1 \cup U_3 = \phi_2(\hat{U}_1) \cap Q;\ U_2 \cup U_4 = \phi_2(\hat{U}_2) \cap Q.$

To complete the proof of our theorem, we must show that both $d\phi$ and $d\phi^{-1}$ contract a bundle of sectors (Moser [9]). This is the content of the following lemma.

LEMMA. *Introduce coordinates r and r^+ in Q, so that Q appears as a square contained in the square $c_1\varepsilon \leqq r, r^+ \leqq c_2\varepsilon$. Choose a number $\lambda > 1$ and introduce the sectors*

$$S^+: |\delta r^+| < \lambda^{-1}|\delta r|, \qquad S^-: |\delta r| < \lambda^{-1}|\delta r^+|, \tag{5}$$

as well as the larger sectors

$$\Sigma^+: |\delta r^+| < \lambda |\delta r|, \qquad \Sigma^-: |\delta r| < \lambda |\delta r^+|.$$

Then for ε sufficiently small, and $d\phi : (\delta r, \delta r^+) \to (\delta r_1, \delta r_1{}^+)$,

$$d\phi(\Sigma^+) \subset S^+ \quad \text{and} \quad |\delta r_1| < \lambda^2 |\delta r| \quad \text{for } (\delta r, \delta r^+) \in S^+, \tag{5}$$

and

$$(d\phi)^{-1}(\Sigma^-) \subset S^- \quad \text{and} \quad |\delta r| < \lambda^2 |\delta r_1| \quad \text{for } (\delta r_1, \delta r_1^+) \in S^- \tag{6}$$

Proof of lemma. We will verify the first part of (6); the second part follows in exactly the same manner. To this end, we stretch the variables by setting

$$\rho = \varepsilon^{-1} r, \qquad \rho^+ = \varepsilon^{-1} r^+.$$

Then

$$\rho^+ = f(\rho, \theta) = [1 + \rho^2 - 2\rho \cos\theta]^{1/2}$$

is independent of ε. The domain Q is now described by $c_1 \leqq \rho, \rho^+ \leqq c_2$. Next choose $c > 1$ such that

$$|f_1|, |f_2|, \text{ and } |f_2^{-1}| \leqq c \quad \text{in } Q,$$

where $f_1 = \partial f/\partial\rho$ and $f_2 = \partial f/\partial\theta$. Note that c only depends on c_1 and c_2.

(i) We express the condition $|\delta\rho^+| \leqq \lambda|\delta\rho|$ in terms of $\delta\theta$ and $\delta\rho$. To this end observe that

$$\delta\theta = f_2^{-1}[\delta\rho^+ - f_1\delta\rho].$$

Hence,

$$\begin{aligned}|\delta\theta| &\leqq |f_2^{-1}|[|\delta\rho^+| + |f_1||\delta\rho|] \\ &\leqq \varepsilon[\lambda|\delta\rho| + c|\delta\rho|] = k|\delta\rho|,\end{aligned}$$

where $k = c(\lambda + c)$.

(ii) Now we apply $\phi_1 : (\delta\theta, \delta\rho) \to (\delta\theta_*, \delta\rho_*)$. Since $\delta\rho_* = \delta\rho$ and

$$\delta\theta_* = \delta\theta + \varepsilon\alpha'(\varepsilon\rho)\delta\rho_*$$

we see from (7) that

$$|\delta\theta_*| > (|\varepsilon\alpha'(\varepsilon\rho)| - k)|\delta\rho_*|.$$

If we choose ε sufficiently small so that $|\varepsilon\alpha'(\varepsilon\rho)| > 2k$, we see that

$$|\delta\theta_*| > k|\delta\rho_*|. \tag{8}$$

(iii) Next, we express (8) in terms of $\delta\rho_*$ and $\delta\rho_*^+$. From the relation $\delta\theta_* = f_2^{-1}(\delta(\rho_*^+ - f_1\delta\rho_*)$, we see that

$$|\delta\theta_*| \leqq c(|\delta\rho_*^+| + c|\delta\rho_*|).$$

Hence, from (8),

$$c(|\delta\rho_*^+| + c|\delta\rho_*|) \geqq k|\delta\rho_*|)$$

or

$$(k - c^2)|\delta\rho_*| \leqq c|\delta\rho_*^+|.$$

Since $k - c^2 = \lambda c$, we see that

$$|\delta\rho_*| \leqq \lambda^{-1}|\delta\rho_*^+| < \lambda|\delta\rho_*^+|. \tag{9}$$

(iv) Next, we interpret (9) in terms of $\delta\theta_*^+$ and $\delta\rho_*^+$. From the relation $\rho^* = f(\rho_*^+, \pi - \theta_*^+)$, we see that

$$\delta\rho_* = f_1\delta\rho_*^+ - f_2\delta\theta_*^+.$$

Hence,

$$|\delta\theta_*^+| \leqq c[c|\delta\rho_*^+| + \lambda|\delta\rho_*^+|] = k|\delta\rho_*^+|. \tag{10}$$

(v) Next, applying ϕ_2: $(\delta\theta_*^+, \delta\rho_*^+) \to (\delta\theta_1^+, \delta\rho_1^+)$ we get, as before, that

$$|\delta\theta_1^+| > k|\delta\rho_1^+|. \tag{11}$$

(vi) Finally, we interpret (11) in terms of $|\delta\rho_1|$ and $|\delta\rho_1^+|$. From the relation $\delta\rho_1 = f_1\delta\rho_1^+ - f_2\delta\theta_1^+$ we see that

$$|\delta\theta_1^+| \leqq c[c|\delta\rho_1^+| + |\delta\rho_1|],$$

and thus from (11) we conclude that

$$c^2|\delta\rho_1^+| + c|\delta\rho_1| > k|\delta\rho_1^+|. \tag{12}$$

It follows immediately from (12) that

$$|\delta\rho_1^+| < \lambda^{-1}|\delta\rho_1|, \tag{13}$$

or, equivalently, $|\delta r_1^+| < \lambda^{-1}|\delta r_1|$. Thus $d\phi : \Sigma^+ \to S^+$.

It remains to show that $|\delta r_1| > \lambda^2|\delta r|$.

From (13),

$$|\delta\rho_1| > \lambda|\delta\rho_1^+| = \lambda|\delta\rho_*^+|.$$

Since $\rho_*^+ = f(\rho_*, \theta_*)$, we have that $\delta\theta_* = f_2^{-1}(\delta\rho_*^+ - f_1\delta\rho_*)$, so that $|\delta\theta_*| \leqq c(|\delta\rho_*^+| + c|\delta\rho|)$, or

$$|\delta\rho_*^+| \geqq c^{-1}|\delta\theta_*| - c|\delta\rho| \geqq c^{-1}k|\delta\rho| - c|\delta\rho|,$$

Thus,

$$|\delta\rho_1| > \lambda(c^{-1}k - c)|\delta\rho| = \lambda^2|\delta\rho|. \qquad \square$$

Remarks. A slight modification of this proof shows that the map $M = \tau \circ \phi_1$, where

$$\tau : (x, y) \to (x + \varepsilon, y)$$

also possesses the shift on N symbols as a subsystem. It has been verified by Dragt [6] that the map $\tau \circ \phi_1$, is the correct model of the global flow for the Störmer problem.

Numerical results. We now present some numerical studies of the mapping M. We took $\alpha(r) = r^{-4/3}$, $\varepsilon = .02$, and placed the two centers at $(-.01, 0)$ and $(.01, 0)$. Each of Figs. 5–15 represents 2,500 iterates of the given initial point, and they are scaled so that the maximum radius of any given point is three. It is helpful in analyzing these pictures to observe that the mapping M has many fixed points on the y-axis. These correspond to points which are fixed under both ϕ_1 and ϕ_2, and to points $(0, y)$ which are mapped by ϕ_1 into $(0, -y)$. Some of these fixed points are elliptic, with their islands of invariant curves surrounding them, and some are hyperbolic. It can be verified numerically that the stable and unstable manifolds of some of these hyperbolic points intersect homoclinically.

Fig. 5 is an invariant curve of M. It is somewhat surprising that we can find invariant curves this close to the origin, since Moser's invariant curve theorem would require r to be exceedingly large. Figs. 6 and 7 illustrate how a family of closed invariant curves surrounding an elliptic fixed point, disintegrates. The intersection of the "curves" in Fig. 7 is a hyperbolic fixed point (the largest fixed point of M on the y-axis) and the stable and unstable manifolds of this point intersect homoclinically. (The elliptic fixed point is centered in the "smile" at the bottom of Fig. 11.)

The orbit in Fig. 8 appears to define an invariant curve, while the orbit in Fig. 9 is periodic, of very high order. (These orbits have been magnified to nearly 15 times their original size.) It is quite surprising that we should find such "stable" behavior so close to the origin. Notice, though, in Figs. 10 and 11 how quickly we change from stability, or integrability, to instability and statistical behavior. Fig. 11 strongly suggests that M is ergodic in a suitable region, and would also seem to verify Easton's [7], conjecture on the ergodicity of linked twist mappings.

The "empty spaces" in Fig. 11 are invariant regions which are enclosed by invariant curves belonging to elliptic periodic points of M. An orbit starting outside these regions can never enter into them. Conversely, an orbit starting in these regions remains inside them, both forwards and backwards. Fig. 12 shows 3 closed curves

belonging to an elliptic periodic point of period 3. As we slowly vary the initial point of the orbit, these closed curves begin to disintegrate (Figs. 13, 14), and then we immediately come back (Fig. 15) to the exact same statistical behavior observed in Fig 11. This reinforces our notion that M is indeed ergodic in a suitable region.

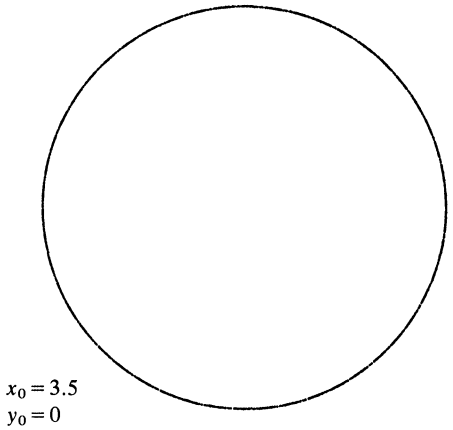


FIG. 5

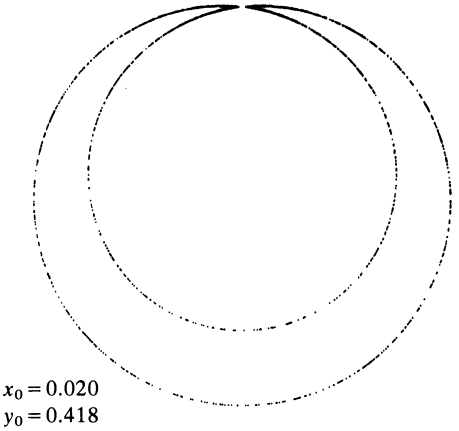


FIG. 6

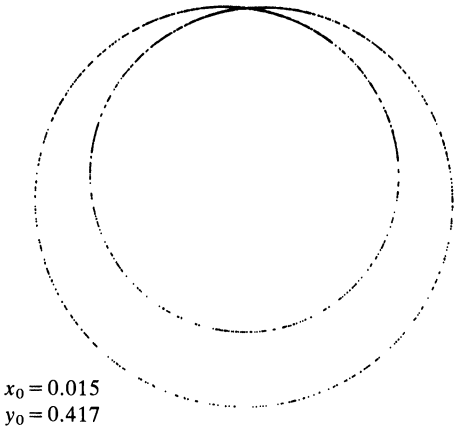


FIG. 7

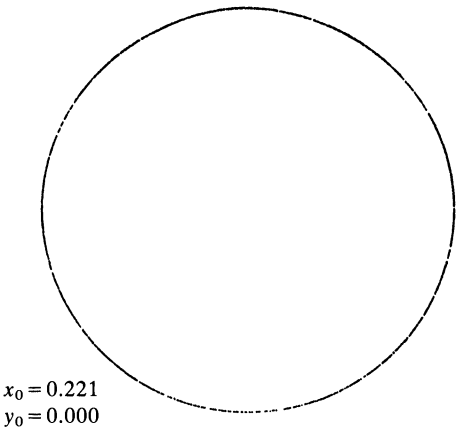
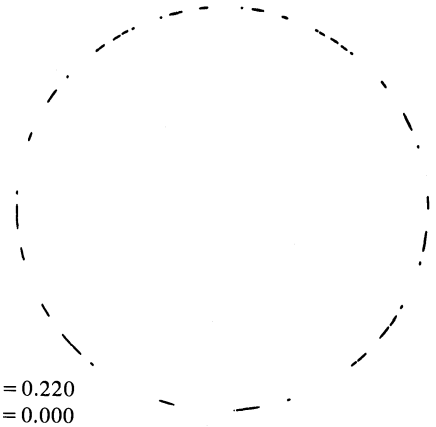
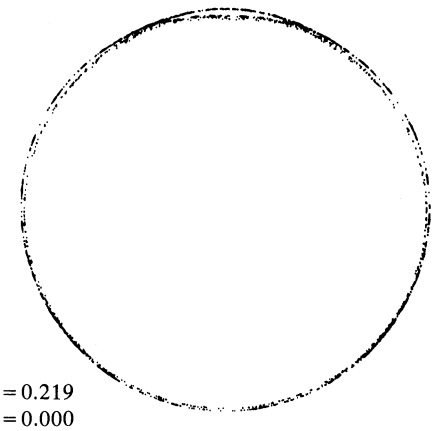


FIG. 8



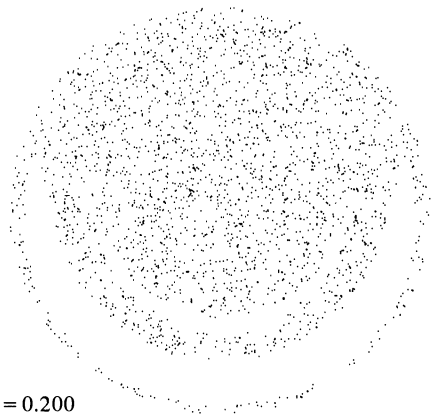
$x_0 = 0.220$
$y_0 = 0.000$

FIG. 9



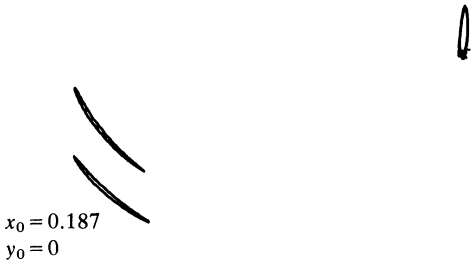
$x_0 = 0.219$
$y_0 = 0.000$

FIG. 10



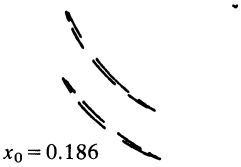
$x_0 = 0.200$
$y_0 = 0.000$

FIG. 11



$x_0 = 0.187$
$y_0 = 0$

FIG. 12



$x_0 = 0.186$
$y_0 = 0.000$

FIG. 13

$x_0 = 0.185$
$y_0 = 0$

FIG. 14

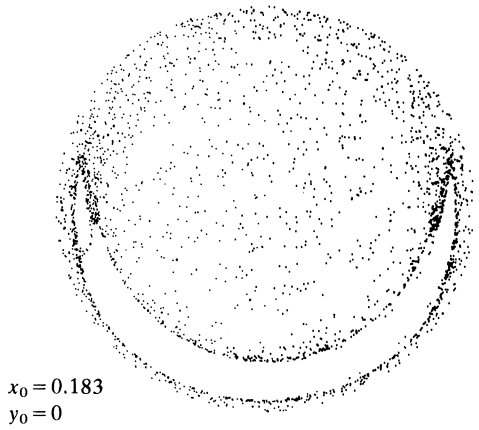


FIG. 15

REFERENCES

[1] R. BOWEN, *On Axiom A Diffeomorphisms*, CBMS Regional Conference Series in Applied Mathematics 35, American Mathematical Society, Providence, RI, 1978.
[2] M. BRAUN, *Particle motions in a magnetic field*, J. Differential Equations, 8 (1979), pp. 294–332.
[3] ———, *Structural stability and the Störmer problem*, Indiana Univ. Math. J. 20 (1970), pp. 469–497.
[4] ———, *Mathematical remarks on the Van Allen radiation belt*, SIAM Rev., 23 (1981), pp. 61–94.
[5] R. L. DEVANEY, *Subshifts of finite type in linked twist mappings*, Proc. Amer. Math. Soc., 71 (1978), pp. 334–338.
[6] A. J. DRAGT AND JOHN M. FINN, *Insolubility of trapped particle motion in a magnetic dipole field*, Tech. Rep. 75–098, U. of Maryland Center for Theoretical Physics, July, 1975.
[7] R. EASTON, *Proc. NSF Conference on Attractors* (Fargo, ND, June, 1977), Springer-Verlag, to appear.
[8] MOSER, J. K., *On invariant curves of area-preserving mappings of an annulus*, Nachr. Abad. Wiss., Göttingen: Math. Phys., kl IIa 1962, pp. 1–20.
[9] J. K. MOSER, *Stable and Random Motions in Dynamical Systems*, Annals of Mathematical Studies 77, Princeton University Press, Princeton, NJ 1973.
[10] S. SMALE, *Differentiable dynamical systems*, Bull. AMS. 73 (1967), pp. 747–817.
[11] W. THURSTON, *On the geometry and dynamics of surfaces*, to appear.

SIAM J. MATH. ANAL.
Vol. 12, No. 4, July 1981

0036-1410/81/1204-0014 $01.00/0

WIDDER'S INVERSION THEOREM AND THE INITIAL DISTRIBUTION PROBLEM*

D. G. ARONSON†

Abstract. Widder's inversion theorem gives the precise form of the trace on $t=0$ of a positive temperature defined in $\mathbb{R}\times(0, T]$ for some $T>0$. In this paper, Widder's theorem is extended to positive temperatures defined in $\mathbb{R}^d\times(0, T]$ for $d\geqq 1$, and the result is applied to derive a necessary and sufficient condition for a given Borel measure on $\mathbb{R}^d$ to be the initial trace of a positive temperature. Similar results are also obtained for positive weak solutions of divergence structure parabolic equations.

Introduction. In 1944, D. V. Widder published his celebrated representation theorem for positive temperatures in an infinite rod [5]. Specifically, if $u=u(x, t)$ is a nonnegative solution of the equation of heat conduction

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2} \tag{1}$$

in $\mathbb{R}\times(0, T)$ for some $T>0$, then there exists a nonnegative Borel measure ρ such that $u=g*\rho$. Here

$$g(x, t)=(4\pi t)^{-1/2}\, e^{-x^2/4t}$$

is the fundamental solution of (1) and

$$(g*\rho)(x, t)\equiv\int_{\mathbb{R}} g(x-\xi, t)\rho(d\xi).$$

If ρ has a density ψ then the convolution $u=g*\rho$ can be written

$$u(x, t)=\int_{\mathbb{R}} g(x-\xi, t)\psi(\xi)\, d\xi, \tag{2}$$

and it is well known that

$$\lim_{t\downarrow 0} u(x, t)=\psi(x)$$

at all points of continuity of ψ. Thus, if ρ has a continuous density ψ, then (2) represents a solution of (1) in $\mathbb{R}\times(0, T]$ with initial datum $u(\cdot, 0)=\psi$. By analogy, in the general case one expects $u=g*\rho$ to represent a solution of (1) with the measure ρ as initial datum in some appropriate sense. Widder's inversion theorem [6] makes this expectation precise. It states that, if $u=g*\rho$ is a solution of (1) then

$$\lim_{t\downarrow 0}\int_a^b u(x, t)\, dx=\rho(a, b)+\tfrac{1}{2}\rho\{a\}+\tfrac{1}{2}\rho\{b\} \tag{3}$$

for all a and b which satisfy $-\infty<a<b<+\infty$.

Recently, C. H. Wilcox [7] has considered the converse problem. Let μ be a nonnegative Borel measure. A nonnegative solution u of (1) is said to have initial distribution μ if

$$\lim_{t\downarrow 0}\int_a^b u(x, t)\, dx=\mu(a, b)$$

* Received by the editors August 15, 1980. This work was supported in part by the National Science Foundation under grant MCS78-02158.

† School of Mathematics, University of Minnesota, Minneapolis, MN 55455.

for all a and b such that $-\infty < a < b < +\infty$ and $\mu\{a\} = \mu\{b\} = 0$. If $u = g * \rho$, then it follows from the Widder inversion theorem (3) that u has initial distribution ρ. On the other hand, Wilcox shows that if u has initial distribution μ then $u = g * \mu$.

In this note I shall extend Widder's inversion theorem and Wilcox's result on the initial distribution problem, first to the equation of heat conduction in $\mathbb{R}^d \times (0, T)$ for any $d \geqq 1$, and then to a broad class of linear parabolic equations with divergence structure. Some of the material needed for these extensions is already at hand, since the appropriate generalization of the Widder representation theorem has been proved in [1]. The principal new result to be given here is the extension of the Widder inversion theorem to $\mathbb{R}^d$ for $d > 1$. A weak convergence version of the inversion theorem in $\mathbb{R}^d$ was proved in [2]. The results for the equation of heat conduction are stated precisely in § 1. The corresponding results for divergence structure equations are given in § 4.

1. Results for the equation of heat conduction. Let $T \in \mathbb{R}^+$ be fixed and consider the equation of heat conduction

$$\frac{\partial u}{\partial t} = \Delta u \tag{1.1}$$

in $S_T \equiv \mathbb{R}^d \times (0, T)$ for $d \geqq 1$, where

$$\Delta = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}$$

is the Laplace operator on $\mathbb{R}^d$. The fundamental solution of (1.1) is given by

$$g(x, t) = (4\pi t)^{-d/2} e^{-|x|^2/4t},$$

where

$$|x|^2 = \sum_{j=1}^{d} x_j^2.$$

Let $C^{2,1}(S_T)$ denote the class of functions $u: S_T \to \mathbb{R}$ such that u, $\partial u/\partial x_i$, $\partial u/\partial t$ and $\partial^2 u/\partial x_j\, \partial x_i$ are all continuous on S_T, and define

$$H^+ \equiv H^+(S_T) \equiv \left\{ u \in C^{2,1}(S_T) \colon u \geqq 0 \text{ and } \frac{\partial u}{\partial t} = \Delta u \text{ in } S_T \right\}.$$

The Widder representation theorem establishes a relationship between H^+ and the class M^+ of nonnegative Borel measures on $\mathbb{R}^d$. The following version of the representation theorem is proved in [1].

THEOREM A. *If $u \in H^+$, then there exists a unique $\rho \in M^+$ such that $u = g * \rho$.*

If $u = g * \rho \in H^+(S_T)$, then ρ cannot have too much mass near $|\xi| = +\infty$. Specifically, if $u = g * \rho \in H^+(S_T)$ then $u(0, T) < +\infty$, which implies that

$$\int_{\mathbb{R}^d} e^{-|\xi|^2/4T} \rho(d\xi) = (4\pi T)^{d/2} u(0, T) < +\infty. \tag{1.2}$$

It follows that ρ is regular and therefore σ-finite. As is shown in [2], the converse is also true. That is, if $\rho \in M^+$ satisfies

$$\int_{\mathbb{R}^d} e^{-|\xi|^2/4T} \rho(d\xi) < +\infty,$$

then $g * \rho \in H^+(S_T)$.

Let $\mathcal{B}$ denote the collection of Borel subsets of $\mathbb{R}^d$. For each $A \in \mathcal{B}$, define

$$\gamma_A(\xi, t) \equiv \int_A g(x-\xi, t)\, dx.$$

Since

$$\int_{\mathbb{R}^d} g(x-\xi, t)\, dx = 1,$$

it follows that $\gamma_A(\xi, t)$ has its values in $[0, 1]$. Therefore

$$\gamma_A^+(\xi) \equiv \limsup_{t \downarrow 0} \gamma_A(\xi, t) \quad \text{and} \quad \gamma_A^-(\xi) \equiv \liminf_{t \downarrow 0} \gamma_A(\xi, t)$$

both exist in $[0, 1]$ for all $\xi \in \mathbb{R}^d$. Let

$$A^* \equiv \{\xi \in \mathbb{R}^d : \gamma_A^+(\xi) \neq \gamma_A^-(\xi)\},$$

and define $\gamma_A : \mathbb{R}^d \backslash A^* \to [0, 1]$ by $\gamma_A(\xi) = \gamma_A^{\pm}(\xi)$. It is easy to verify using the explicit formula for g that

$$\gamma_A^{\pm}(\xi) = \begin{cases} 1 & \text{for } \xi \in \mathring{A}, \\ 0 & \text{for } \xi \in \mathbb{R}^d \backslash \bar{A}. \end{cases} \tag{1.3}$$

Thus $A^* \subset \partial A \equiv \bar{A} \backslash \mathring{A}$. In general, $A^* \neq \varnothing$. In the Appendix I give an example of a bounded open set $A \subset \mathbb{R}^d$ for $d \geqq 2$ such that $A^* \neq \varnothing$.

THEOREM B. *Let $u \in H^+$ be given by $u = g * \rho$ for $\rho \in M^+$. If $A \in \mathcal{B}$ is bounded and $\rho(A^*) = 0$, then*

$$\lim_{t \downarrow 0} \int_A u(x, t)\, dx = \rho(\mathring{A}) + \int_{\partial A} \gamma_A(\xi) \rho(d\xi). \tag{1.4}$$

Theorem B contains Widder's inversion theorem for $d = 1$. Indeed, if $A = (a, b)$ then $\partial A = \{a\} \cup \{b\}$. Since $\gamma_A(\xi) = \frac{1}{2}$ for $\xi = a$ or b, it follows that $A^* = \varnothing$, and (1.4) reduces to (3). Moreover, Theorem B is sharp. Let $A \in \mathcal{B}$ be such that $A^* \neq \varnothing$. Fix $\xi_0 \in A^*$ and let $\rho = \delta_{\xi_0}$, where δ_{ξ_0} denotes the Dirac measure concentrated at ξ_0. Then

$$u(x, t) = \int_A g(x-\xi, t) \delta_{\xi_0}(d\xi) = g(x-\xi_0, t)$$

and

$$\int_A u(x, t)\, dx = \gamma_A(\xi_0, t).$$

Thus, in view of the definition of A^*,

$$\lim_{t \downarrow 0} \int_A u(x, t)\, dx$$

does not exist in this case.

For arbitrary a and b in $\mathbb{R}^d$, define

$$(a, b) \equiv \{x \in \mathbb{R}^d : a_i < x < b_i,\ i = 1, 2, \cdots, d\},$$

$$(a, b] \equiv \{x \in \mathbb{R}^d : a_i < x \leqq b_i,\ i = 1, 2, \cdots, d\},$$

$$[a, b] \equiv \{x \in \mathbb{R}^d : a_i \leqq x \leqq b_i,\ i = 1, 2, \cdots, d\}.$$

These sets will be referred to as intervals. The boundary of any nonempty interval $(a, b]$ in $\mathbb{R}^d$ can be written in the form

$$\partial(a, b] = \bigcup_{j=0}^{d-1} L_j,$$

where for $j \geqq 1$ each L_j is a finite union of open subsets of j-dimensional hyperplanes and L_0 consists of the 2^d vertices of $(a, b]$. Moreover, it is not difficult to verify that

$$\gamma_{(a,b]}(\xi) = 2^{j-d} \quad \text{for } \xi \in L_j.$$

Thus, in particular, $(a, b]^* = \varnothing$.

If $Q \subset \mathbb{R}$ then Q^d will denote the d-fold Cartesian product of Q with itself. Define

$$I_Q \equiv \{(a, b]: a \in Q^d \text{ and } b \in Q^d\}.$$

A function $u \in H^+$ will be said to have *initial distribution* $\mu \in M^+$ if μ is σ-finite and there exists a countable dense subset $Q \subset \mathbb{R}$ such that

$$\lim_{t \downarrow 0} \int_{(a,b]} u(x, t)\, dx = \mu(a, b)$$

and

$$\mu(\partial(a, b]) = 0$$

for all $(a, b] \in I_Q$. Note that this differs from Wilcox's definition discussed in the Introduction since it involves only a countable number of conditions.

THEOREM C. *$u \in H^+$ has initial distribution $\mu \in M^+$ if and only if $u = g * \mu$.*

Theorems B and C are proved in the next two sections.

2. Proof of Theorem B. By Tonelli's theorem,

$$\int_A u(x, t)\, dx = \int_{\mathbb{R}^d} \gamma_A(\xi, t) \rho(d\xi).$$

In view of the definitions of A^* and $\gamma_A(\cdot)$,

$$\lim_{t \downarrow 0} \gamma_A(\xi, t) = \gamma_A(\xi) \quad \text{for all } \xi \in \mathbb{R}^d \backslash A^*.$$

Since by hypothesis $\rho(A^*) = 0$, this convergence is ρ-almost everywhere. As will be shown below, there exists a constant $K \in \mathbb{R}^+$ such that

$$0 \leqq \gamma_A(\xi, t) \leqq K e^{-|\xi|^2/4T} \quad \text{for all } (\xi, t) \in \mathbb{R}^d \times (0, T/2]. \tag{2.1}$$

According to (1.2), the function $K e^{-|\xi|^2/4T}$ is ρ-integrable. Therefore, by the dominated convergence theorem,

$$\lim_{t \downarrow 0} \int_A u(x, t)\, dx = \int_{\mathbb{R}^d} \gamma_A(\xi) \rho(d\xi),$$

and (1.4) follows from $\mathbb{R}^d = \mathring{A} \cup \partial A \cup (\mathbb{R}^d \backslash \bar{A})$ and (1.3).

Since A is bounded, there exists a constant $R > 0$ such that $x \in A$ implies $|x| \leqq R$. Set $\delta = 1/2T$. Then

$$\gamma_A(\xi, t) = \int_A g(x - \xi, t)\, dx \leqq e^{\delta R^2} (4\pi t)^{-d/2} \int_{\mathbb{R}^d} \exp\left(-\frac{|x - \xi|^2}{4t} - \delta |x|^2 \right) dx.$$

By the standard device of completing squares one finds that

$$\frac{|x-\xi|^2}{4t}+\delta|x|^2=\frac{|\zeta|^2}{4t}+\frac{\delta|\xi|^2}{1+4\delta t},$$

where

$$\zeta=(1+4\delta t)^{1/2}x-(1+4\delta t)^{-1/2}\xi.$$

Thus

$$\gamma_A(\xi,t)\leqq e^{\delta R^2}(4\pi t)^{-d/2}\,e^{-\delta|\xi|^2/(1+4\delta t)}\int_{\mathbb{R}^d}e^{-|\zeta|^2/4t}\,dx$$

or, with the change of variables $\nu=\zeta/\sqrt{4t}$,

$$\gamma_A(\xi,t)\leqq e^{\delta R^2}\{\pi(1+4\delta t)\}^{-d/2}\left(\int_{\mathbb{R}^d}e^{-|\nu|^2}\,d\nu\right)e^{-\delta|\xi|^2/(1+4\delta t)}.$$

Since

$$\{\pi(1+4\delta t)\}^{-d/2}\int_{\mathbb{R}^d}e^{-|\nu|^2}\,d\nu=(1+4\delta t)^{-d/2}\leqq 1$$

and

$$\frac{\delta}{1+4\delta t}\geqq\frac{\delta}{1+2\delta T}=\frac{1}{4T}$$

for $t\in[0,\,T/2]$, it follows that (2.1) holds with $K=e^{R^2/2T}$.

3. Proof of Theorem C. Suppose that $u=g*\mu\in H^+(S_T)$. Since $(a,\,b]^{\#}=\varnothing$ for every interval $(a,\,b]$, it follows from Theorem B that

$$\lim_{t\downarrow 0}\int_{(a,b]}u(x,t)\,dx=\mu(a,b)$$

for all intervals $(a,\,b]$ such that $\mu(\partial(a,\,b])=0$.

Define the measure

$$\nu(A)\equiv\int_A e^{-|\xi|^2/4T}\,\mu(d\xi)$$

for $A\in\mathcal{B}$. In view of (1.2) and $g*\mu\in H^+(S_T)$,

$$\nu(\mathbb{R}^d)=\int_{\mathbb{R}^d}e^{-|\xi|^2/4T}\,\mu(d\xi)<+\infty. \tag{3.1}$$

For each $l\in\{1,\,2,\,\cdots,\,d\}$ and $s\in\mathbb{R}$ let

$$H_s^l=\{x\in\mathbb{R}^d:x_l=s\};$$

that is, let H_s^l denote the hyperplane $x_l=s$. Clearly $\mathbb{R}^d=\bigcup_{s\in\mathbb{R}}H_s^l$ for each fixed l. Let

$$P^l=\{s\in\mathbb{R}:\,\nu(H_s^l)>0\}.$$

Then

$$P^l=\bigcup_{n=1}^{\infty}P_n^l, \tag{3.2}$$

where

$$P_n^l = \left\{ s \in \mathbb{R}: \frac{1}{n} < \nu(H_s^l) \leqq \frac{1}{n-1} \right\}.$$

I claim that for each $l \in \{1, 2, \cdots, d\}$ the set P^l is countable.

Suppose for contradiction that P^l is not countable. Then, in view of (3.2), there exists an integer $n \geqq 1$ such that P_n^l is not countable. Let $\{s_j\}$ be a sequence of distinct elements of P_n^l. For all positive integers N,

$$\bigcup_{j=1}^{N} H_{s_j}^l \subset \mathbb{R}^D,$$

so that

$$\frac{N}{n} \leqq \sum_{j=1}^{N} \nu(H_{s_j}^l) \leqq \nu(\mathbb{R}^d),$$

However, since N is arbitrary and $N/n \to +\infty$ as $N \to +\infty$, this contradicts (3.1).

Let $E \equiv \bigcup_{l=1}^{d} P^l$. Then E is countable and $s \in \mathbb{R} \backslash E$ implies

$$\nu(H_t^l) = \int_{H_s^l} e^{-|\xi|^2/4T} \mu(d\xi) = 0$$

for every $l \in \{1, 2, \cdots, d\}$. In particular, it follows that

$$\mu(S) = 0, \tag{3.3}$$

where S is any bounded subset of any hyperplane H_s^l with $l \in \{1, 2, \cdots, d\}$ and $s \in \mathbb{R} \backslash E$.

Since E is countable, $\mathbb{R} \backslash E$ is everywhere dense in $\mathbb{R}$, and it is an easy matter to construct a countable dense subset Q of $\mathbb{R} \backslash E$. For example, fix $q_0 \in \mathbb{R} \backslash E$, and for each positive integer n choose a sequence $\{q_{nj}: j = 0, \pm 1, \pm 2, \cdots\}$ such that $q_{n0} = q_0$ and

$$q_{nj} \in \left(q_0 + \frac{j}{2^{n-1}} - \frac{1}{2^{2+n}}, q_0 + \frac{j}{2^{n-1}} + \frac{1}{2^{2+n}} \right) \cap (\mathbb{R} \backslash E).$$

Then, as is easily verified,

$$Q = \bigcup_{n=1}^{\infty} \{q_{nj}: j = 0, \pm 1, \pm 2, \cdots\}$$

is a countable dense subset of $\mathbb{R} \backslash E$.

If $(a, b] \in I_Q$ then $\partial(a, b]$ is a finite union of bounded subsets of hyperplanes H_s^l for $l \in \{1, \cdots, d\}$ and $s \in Q \subset \mathbb{R} \backslash E$. According to (3.3), $\mu(\partial(a, b]) = 0$. Therefore u has initial distribution μ.

Now assume that $u \in H^+$ has initial distribution μ. By definition, this means that there exists a countable dense subset $Q \subset \mathbb{R}$ such that

$$\mu(\partial(a, b]) = 0$$

and

$$\lim_{t \downarrow 0} \int_{(a,b]} u(x, t)\, dx = \mu(a, b) \tag{3.4}$$

for all $(a, b] \in I_Q$. On the other hand, since $u \in H^+$ it follows from Theorem A that there exists a unique $\rho \in M^+$ such that $u = g * \rho$. Moreover, by Theorem B,

$$\lim_{t \downarrow 0} \int_{(a,b]} u(x, t)\, dx = \rho(a, b) + \int_{\partial(a,b]} \gamma_A(\xi) \rho(d\xi). \tag{3.5}$$

I claim that

$$\rho(a, b] = \mu(a, b] \quad \text{for all } (a, b] \in I_Q. \tag{3.6}$$

In view of (3.4) and (3.5), in order to prove (3.6) it suffices to prove that

$$\rho(\partial(a, b]) = 0 \quad \text{for all } (a, b] \in I_Q.$$

Fix $(a, b] \in I_Q$. Since Q is dense in $\mathbb{R}$ one can choose strictly monotone sequences $\{a_{nj}^{\pm}\}$ and $\{b_{nj}^{\pm}\}$ in Q such that

$$a_{nj}^{+} \downarrow a_j, \quad a_{nj}^{-} \uparrow a_j, \qquad b_{nj}^{+} \downarrow b_j, \quad b_{nj}^{-} \uparrow b_j$$

as $n \to +\infty$. Let $a_n^{\pm} = (a_{n1}^{\pm}, \cdots, a_{nd}^{\pm})$ and $b_n^{\pm} = (b_{n1}^{\pm}, \cdots, b_{nd}^{\pm})$. Then $(a_n^{\pm}, b_n^{\mp}] \in I_Q$ with

$$(a_n^{-}, b_n^{+}] \downarrow (a, b] \quad \text{and} \quad (a_n^{+}, b_n^{-}] \uparrow (a, b]$$

as $n \to +\infty$. Moreover,

$$S_n \equiv (a_n^{-}, b_n^{+}] \setminus (a_n^{+}, b_n^{-}] \downarrow \partial(a, b].$$

Observe that

$$\mu(S_n) = \mu(a_n^{-}, b_n^{+}] - \mu(a_n^{+}, b_n^{-}] = \lim_{t \downarrow 0} \int_{(a_n^{-}, b_n^{+}]} u \, dx - \lim_{t \downarrow 0} \int_{(a_n^{+}, b_n^{-}]} u \, dx$$

$$= \lim_{t \downarrow 0} \int_{S_n} u \, dx \geqq \rho(\mathring{S}_n) \geqq \rho(\partial(a, b]).$$

Therefore, $S_n \downarrow \partial(a, b]$ implies that

$$0 = \mu(\partial(a, b]) \geqq \rho(\partial(a, b]).$$

If $(a, b]$ and $(a', b']$ are arbitrary intervals, then

$$(a, b] \cap (a', b'] = (a \vee a', b \wedge b']$$

where

$$a \vee a' = (a_1 \vee a_1, \cdots, a_d \vee a_d') \quad \text{and} \quad b \wedge b' = (b_1 \wedge b_1', \cdots, b_d \wedge b_d').$$

In particular, if $(a, b]$ and $(a', b']$ both belong to I_Q then $(a, b] \cap (a', b']$ either belongs to I_Q or is empty. Thus $\{I_Q, \varnothing\}$ is a π-system. Let $\mathcal{B}^*$ denote the σ-field generated by $\{I_Q, \varnothing\}$. I claim that $\mathcal{B}^* = \mathcal{B}$.

Let G be an arbitrary open subset of $\mathbb{R}^d$. Since Q is dense in $\mathbb{R}$, for each $x \in G$ there exist a_x and b_x in Q^d such that

$$x \in (a_x, b_x) \subset [a_x, b_x] \subset G.$$

Thus

$$G = \bigcup_{x \in G} (a_x, b_x].$$

Each $(a_x, b_x] \in I_Q$ and I_Q contains only countably many distinct intervals. Therefore every open subset $G \subset \mathbb{R}^d$ is a countable union of intervals from I_Q. Consequently $\mathcal{B}^*$ contains the open subset of $\mathbb{R}^d$, so that $\mathcal{B} \subset \mathcal{B}^*$. On the other hand, for each $(a, b] \in I_Q$ there exist decreasing sequences $\{b_{nj}\}$ in Q such that $b_{nj} \downarrow b_j$ as $n \to +\infty$ for each $j \in \{1, \cdots, d\}$. Set $b_n = (b_{n1}, \cdots, b_{nd})$. Since

$$(a, b] = \bigcap_{n=1}^{\infty} (a, b_n),$$

it follows that $\mathcal{B}^*$ is generated by a subclass of the open subsets of $\mathbb{R}^d$. Therefore $\mathcal{B}^* \subset \mathcal{B}$.

The measures ρ and μ are both σ-finite and defined on the σ-field $\mathcal{B}$ generated by the π-system $\{I_Q, \varnothing\}$. Moreover, according to (3.6), they agree on $\{I_Q, \varnothing\}$. They therefore agree on $\mathcal{B}$ [3, Thm. 10.3]. That is, if $u \in H^+$ has initial distribution μ, then $\mu = \rho$, where ρ is the unique element of M^+ such that $u = g * \rho$.

4. Generalizations. All of the results given above for nonnegative solutions of the heat conduction equation can be extended to nonnegative weak solutions of divergence structure linear parabolic equations. Consider the equation

$$\frac{\partial u}{\partial t} = \Lambda u, \tag{4.1}$$

where, with the convention that repeated indices are to be summed from 1 to d,

$$\Lambda \equiv \frac{\partial}{\partial x_j}\left\{A_{ij}(x,t)\frac{\partial}{\partial x_i} + A_j(x,t)\right\} + B_j(x,t)\frac{\partial}{\partial x_j} + C(x,t).$$

Assume that the coefficients of Λ are bounded measurable functions in $\mathbb{R}^d \times [0, T]$ for some $T \in \mathbb{R}^+$, and that there exists a constant $\lambda \in \mathbb{R}^+$ such that $A_{ij}(x,t)\xi_i\xi_j \geqq \lambda|\xi|^2$ for all $\xi \in \mathbb{R}^d$ and $(x,t) \in \mathbb{R}^d \times [0,T]$. The results which I will describe here actually hold under somewhat less stringent conditions on the coefficients of Λ. Further details can be found in [1].

A function $u = u(x,t)$ is said to be a *weak solution* of (4.1) in $S_T = \mathbb{R}^d \times (0,T)$ if

$$u \in L^\infty(\delta, T; L^2_{\mathrm{loc}}(\mathbb{R}^d)) \cap L^2(\delta, T; H^{1,2}_{\mathrm{loc}}(\mathbb{R}^d))$$

for all $\delta \in (0,T)$, and u satisfies the integral identity

$$\int_{S_T}\left(-u\frac{\partial\varphi}{\partial t} + A_{ij}\frac{\partial u}{\partial x_i}\frac{\partial\varphi}{\partial x_j} + A_j u\frac{\partial\varphi}{\partial x_j} - B_j\frac{\partial u}{\partial x_j}\varphi - Cu\varphi\right)dx\,dt = 0$$

for all test functions $\varphi \in C_0^1(S_T)$. Here $H^{1,2}_{\mathrm{loc}}(\mathbb{R}^d)$ denotes the space of functions on $\mathbb{R}^d$ whose gradient in the sense of distributions belongs to $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. Among the various representatives of any weak solution of (4.1) there is always one which is continuous in S_T. Thus it makes sense to speak of the value of a weak solution at a point.

In [1] it is shown that (4.1) possesses a weak fundamental solution $k_\Lambda(x,t;\xi,\tau)$. Thus, for example, the weak solution of (4.1) in S_T which satisfies $u(\cdot,0) = \psi$ (in the appropriate sense) is given by

$$u(x,t) = \int_{\mathbb{R}^d} k_\Lambda(x,t;\xi,0)\psi(\xi)\,d\xi.$$

One of the principal results in [1] is an estimate for k_Λ in terms of the fundamental solutions $g(x,\alpha t)$ of equations of the form

$$\frac{\partial u}{\partial t} = \alpha\Delta u.$$

Specifically, there exist constants $\alpha_1 \in \mathbb{R}^+$, $\alpha_2 \in \mathbb{R}^+$, and $C \geqq 1$, which depend only on T, λ and the bounds for the coefficients of Λ, such that

$$C^{-1}g(x-\xi, \alpha_1(t-\tau)) \leqq k_\Lambda(x,t;\xi,\tau) \leqq Cg(x-\xi, \alpha_2(t-\tau)). \tag{4.2}$$

Let $H^+_\Lambda \equiv H^+_\Lambda(S_T)$ denote the class of nonnegative weak solutions of (4.1) in S_T. The following generalization of the Widder representation theorem is proved in [1].

THEOREM A′. *If $u \in H_\Lambda^+$, there exists a unique $\rho \in M^+$ such that*

$$u(x, t) = \int_{\mathbb{R}^d} k_\Lambda(x, t; \xi, 0)\rho(d\xi).$$

If

$$u = \int_{\mathbb{R}^d} k_\Lambda \rho(d\xi) \in H_\Lambda^+(S_T)$$

then, in view of Theorem A′ and (4.2),

$$u(0, T) = \int_{\mathbb{R}^d} k_\Lambda(0, T; \xi, 0)\rho(d\xi) \geqq C^{-1} \int_{\mathbb{R}^d} g(-\xi, \alpha_1 T)\rho(d\xi).$$

Therefore

$$\int_{\mathbb{R}^d} e^{-|\xi|^2/4\alpha_1 T} \rho(d\xi) \leqq C(4\pi\alpha_1 T)^{d/2} u(0, T) < +\infty. \tag{4.3}$$

Note that this is the analogue of (1.2).

For $A \in \mathcal{B}$, define

$$\varkappa_{\Lambda,A}(\xi, t) \equiv \int_A k_\Lambda(x, t; \xi, 0)\, dx.$$

It follows from (4.2) that $\varkappa_{\Lambda,A}$ is bounded on S_T. Thus

$$\varkappa_{\Lambda,A}^+(\xi) = \limsup_{t \downarrow 0} \varkappa_{\Lambda,A}(\xi, t) \quad \text{and} \quad \varkappa_{\Lambda,A}^-(\xi) = \liminf_{t \downarrow 0} \varkappa_{\Lambda,A}(\xi, t)$$

are well defined on $\mathbb{R}^d$. Set

$$A_\Lambda^\# \equiv \{\xi \in \mathbb{R}^d : \varkappa_{\Lambda,A}^+(\xi) \neq \varkappa_{\Lambda,A}^-(\xi)\},$$

and define

$$\varkappa_{\Lambda,A}(\xi) = \varkappa_{\Lambda,A}^\pm(\xi) \quad \text{for } \xi \in \mathbb{R}^d \backslash A_\Lambda^\#.$$

[1, Lemma 8] and (4.2) imply that

$$\varkappa_{\Lambda,A}(\xi) = \begin{cases} 1 & \text{for } \xi \in \mathring{A}, \\ 0 & \text{for } \xi \in \mathbb{R}^d \backslash \bar{A}. \end{cases} \tag{4.4}$$

Therefore $A_\Lambda^\# \subset \partial A$.

THEOREM B′. *Let*

$$u(x, t) = \int_{\mathbb{R}^d} k_\Lambda(x, t; \xi, 0)\rho(d\xi) \in H_\Lambda^+.$$

If $A \in \mathcal{B}$ is bounded and $\rho(A_\Lambda^\#) = 0$ then

$$\lim_{t \downarrow 0} \int_A u(x, t)\, dx = \rho(\mathring{A}) + \int_{\partial A} \varkappa_{\Lambda,A}(\xi)\rho(d\xi).$$

The proof of Theorem B′ is essentially the same as the proof of Theorem B. By Tonelli's theorem

$$\int_A u(x, t)\, dx = \int_{\mathbb{R}^d} \varkappa_{\Lambda,A}(\xi, t)\, d\xi,$$

where, since $\rho(A_\Lambda^{\#})=0$, $\varkappa_{\Lambda,A}(\cdot,t)\to\varkappa_{\Lambda,A}$ ρ-almost everywhere. In view of (4.2),

$$\varkappa_{\Lambda,A}(\xi,t)\leqq C\int_A g(x-\xi,\alpha_2 t)\,dx.$$

By an argument similar to the one employed in § 2, one can show that there exist positive constants C' and T' depending only on C, T, λ and A such that

$$0\leqq\varkappa_{\Lambda,A}(\xi,t)\leqq C'\,e^{-|\xi|^2/4\alpha_1 T}\quad\text{for all }(\xi,t)\in\mathbb{R}^d\times(0,T'].$$

(For further details see [2].) Since, according to (4.3), the function $C'\,e^{-|\xi|^2/4\alpha_1 T}$ is ρ-integrable, the assertion follows from the dominated convergence theorem and (4.4).

The definition of initial distribution given in § 1 applies without change to $u\in H_\Lambda^+$. Apart from some measure theory, the proof of Theorem C given in § 3 uses only the representation and inversion theorems and the estimate (1.2). The analogues of all these results are available in the present case to prove the following result.

THEOREM C'. *$u\in H_\Lambda^+$ has initial distribution $\mu\in M^+$ if and only if*

$$u(x,t)=\int_{\mathbb{R}^d} k_\Lambda(x,t;\xi,0)\mu(d\xi).$$

Finally, there remains an interesting open question. Does the set $A_\Lambda^{\#}$ actually depend on the differential operator Λ or is it independent of Λ within the class of strictly elliptic divergence structure operators with bounded measurable coefficients? To answer this question, it is likely that one will need more refined estimates for the fundamental solution than those given by (4.2).

Appendix. For each $d\geqq 2$ there exist bounded open subsets $A\subset\mathbb{R}^d$ such that $A^{\#}\neq\varnothing$. To construct an example of such a set let

$$r_j=2^{-(j-1)^2},\quad l_j=2^{-(j-1/2)^2},\quad c_j=\tfrac{1}{2}(r_j+l_j),\quad d_j=\tfrac{1}{2}(r_{j+1}+l_j)$$

and

$$A_j=\{x\in\mathbb{R}^d\colon l_j<|x|<r_j\}$$

for $j=1,2,\cdots$. The set

$$A\equiv\bigcup_{j=1}^{\infty}A_j$$

is open and bounded. Moreover, $\{0\}\subset\partial A$. I claim that

$$\{0\}\subset A^{\#}. \tag{A.1}$$

Write

$$\gamma_A(0,t)=\sum_{j=1}^{\infty}\int_{A_j}g(x,t)\,dx,$$

and introduce d-dimensional spherical coordinates with $z=|x|/\sqrt{4t}$ to obtain

$$\gamma_A(0,t)=\frac{2}{\Gamma(d/2)}\sum_{j=1}^{\infty}\int_{l_j/\sqrt{4t}}^{r_j/\sqrt{4t}} z^{d-1}\,e^{-z^2}\,dz.$$

Note that the integrand $z^{d-1}\,e^{-z^2}$ achieves its maximum value for $z\in\mathbb{R}^+$ at $z=z^*\equiv\sqrt{(d-1)/2}$.

For each $k = 1, 2, \cdots$ choose t_k so that the interval

$$\left(\frac{l_k}{\sqrt{4t_k}}, \frac{r_k}{\sqrt{4t_k}}\right)$$

is centered at z^*, that is, choose

$$t_k = \left(\frac{c_k}{2z^*}\right)^2.$$

Then

$$\gamma_A(0, t_k) > \frac{2}{\Gamma(d/2)} \int_{l_k z^*/c_k}^{r_k z^*/c_k} z^{d-1} e^{-z^2} dz. \tag{A.2}$$

Since

$$\frac{r_k}{l_k} = 2^{k-3/4},$$

it follows that

$$\frac{r_k z^*}{c_k} = \frac{2z^*}{1 + l_k/r_k} \to 2z^*$$

and

$$\frac{l_k z^*}{c_k} = \frac{2z^*}{1 + r_k/l_k} \to 0$$

as $k \to +\infty$. Thus, in view of (A.2),

$$\gamma_A^+(0) \geqq \frac{2}{\Gamma(d/2)} \int_0^{2z^*} z^{d-1} e^{-z^2} dz \equiv U_d. \tag{A.3}$$

For each $k = 1, 2, \cdots$, choose s_k so that the interval

$$\left(\frac{r_{k+1}}{\sqrt{4s_k}}, \frac{l_k}{\sqrt{4s_k}}\right)$$

is centered at z^*; that is, choose

$$s_k = \left(\frac{d_k}{2z^*}\right)^2.$$

Observe that

$$\gamma_A(0, s_k) < \frac{2}{\Gamma(d/2)} \left\{ \int_0^{r_{k+1} z^*/d_k} z^{d-1} e^{-z^2} dz + \int_{l_k z^*/d_k}^{+\infty} z^{d-1} e^{-z^2} dz \right\}. \tag{A.4}$$

Now

$$\frac{l_k}{r_{k+1}} = 2^{k-1/4}.$$

Consequently,

$$\frac{r_{k+1} z^*}{d_k} = \frac{2z^*}{1 + l_k/r_{k+1}} \to 0$$

and

$$\frac{l_k z^*}{d_k} = \frac{2z^*}{1 + r_{k+1}/l_k} \to 2z^*$$

as $k \to +\infty$. It follows from (A.4) that

$$\gamma_{\bar{A}}(0) \leqq \frac{2}{\Gamma(d/2)} \int_{2z^*}^{+\infty} z^{d-1} e^{-z^2} dz \equiv L_d. \tag{A.5}$$

In view of (A.3) and (A.5), to prove (A.1) it suffices to prove that $L_d < U_d$ for all integers $d \geqq 2$. However, $L_d + U_d = 1$, so that it is enough to prove that

$$L_d < \tfrac{1}{2} \quad \text{for } d = 2, 3, \cdots. \tag{A.6}$$

It is easy to verify that

$$L_2 = 2 \int_{\sqrt{2}}^{+\infty} z e^{-z^2} dz = e^{-2} < \frac{1}{2}.$$

Moreover,

$$L_3 = \frac{2}{\sqrt{\pi}} \left(2e^{-4} + \int_2^{+\infty} e^{-z^2} dz \right).$$

Since $z^2 \geqq 2z$ for $z \geqq 2$ it follows that

$$L_3 \leqq \frac{2}{\sqrt{\pi}} \left(2e^{-4} + \int_2^{+\infty} e^{-2z} dz \right) = \frac{5}{\sqrt{\pi}} e^{-4} < \frac{1}{2}.$$

Thus it remains to be shown that (A.6) holds for $d \geqq 4$.

Write

$$z^{d-1} e^{-z^2} = z e^{-z^2/2} q(z),$$

where

$$q(z) \equiv z^{d-2} e^{-z^2/2}.$$

The function $q(z)$ achieves its maximum in $\mathbb{R}^+$ at $z = \sqrt{d-2}$ and is strictly decreasing for $z > \sqrt{d-2}$. Note that $\sqrt{d-2} < 2z^* = \sqrt{2(d-1)}$. Therefore $q(z) \leqq q(2z^*)$ for $z \geqq 2z^*$ and, in view of (A.5),

$$L_d \leqq \frac{2}{\Gamma(d/2)} q(2z^*) \int_{2z^*}^{+\infty} z e^{-z^2/2} dz = \frac{2}{\Gamma(d/2)} \{2(d-1)\}^{d/2-1} e^{-2(d-1)}. \tag{A.7}$$

Suppose that $d = 2m$ for an integer $m \geqq 2$. Then, according to (A.7),

$$L_d \leqq \frac{2^m}{\Gamma(m)} (2m-1)^{m-1} e^{-2(2m-1)}.$$

Robbins [4] has proved that for $n \geqq 1$

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n} e^{r_n}, \tag{A.8}$$

where

$$\frac{1}{12n+1} < r_n < \frac{1}{12n}.$$

Hence

$$\Gamma(m) = (m-1)! > \sqrt{2\pi} (m-1)^{m-1/2} e^{-(m-1)}$$

and

$$L_d < \frac{2^{m-1/2}}{\sqrt{\pi}} \frac{(2m-1)^{m-1}}{(m-1)^{m-1/2}} e^{-3m+1}.$$

It is easy to verify that for $m \geqq 2$

$$\frac{2m-1}{m-1} \leqq 3 \quad \text{and} \quad \frac{(m-1)^{1/2}}{2m-1} \leqq 2^{-3/2}.$$

Therefore

$$L_d \leqq \frac{1}{\sqrt{\pi}} 2^{m-2} 3^m e^{-3m+1} = \frac{1}{\sqrt{\pi}} \exp\left(m(\log 6 - 3) + 1 - 2\log 2\right) \leqq .0342\cdots < \frac{1}{2}$$

for $d = 2m$ and $m \geqq 2$.

For $d = 2m+1$,

$$L_d \leqq \frac{2^{2m}}{\Gamma(m+1/2)} m^{m-1/2} e^{-4m}.$$

Since

$$\Gamma\left(m+\frac{1}{2}\right) = \frac{(2m-1)!\Gamma(1/2)}{(m-1)!2^{2m-1}},$$

it follows from (A.8) that

$$L_d \leqq \frac{2^{4m-1}}{\sqrt{\pi}} e^{-3m} \frac{(m^2-m)^{m-1/2}}{(2m-1)^{2m-1/2}} e^{R(m)},$$

where

$$R(m) = \frac{12m+1}{12(m-1)(24m-11)}.$$

It is easily verified that for $m \geqq 2$,

$$\frac{m^2-m}{(2m-1)^2} \leqq \frac{1}{4}, \quad \left(\frac{m^2-m}{2m-1}\right)^{-1/2} \leqq \left(\frac{3}{2}\right)^{1/2}, \quad R(m) \leqq R(2) = .0563\cdots.$$

Therefore

$$L_d \leqq \frac{1}{\sqrt{\pi}} 2^{2m-3/2} 3^{1/2} e^{-3m+R(2)} = \frac{1}{\sqrt{\pi}} \exp\left(m(\log 4 - 3) + \frac{1}{2}\log 3 - \frac{3}{2}\log 2 + R(2)\right)$$

$$\leqq .01449\cdots < \frac{1}{2}$$

for $d = 2m+1$ and $m \geqq 2$. This completes the proof of (A.6).

Acknowledgment. I am indebted to Professor Calvin H. Wilcox for bringing these problems to my attention and for several helpful discussions.

REFERENCES

[1] D. G. ARONSON, *Non-negative solutions of linear parabolic equations*, Annali Scuola Normale Sup. Pisa, Classe di Sci., 22 (1968), pp. 607–694.

[2] ———, *Non-negative solutions of linear parabolic equations: an addendum*, Annali Scuola Normale Sup. Pisa, Classe di Sci., 25 (1971), pp. 222–228.

[3] P. BILLINGSLEY, *Probability and Measure*, John Wiley, New York, 1979.

[4] H. ROBBINS, *A remark on Stirling's Formula*, Amer. Math. Monthly, 62 (1955), pp. 26–29.

[5] D. V. WIDDER, *Positive temperature on an infinite rod*, Trans. Amer. Math. Soc., 55 (1944), pp. 85–95.

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

[7] C. H. WILCOX, *Positive temperatures with prescribed initial heat distributions*, Amer. Math. Monthly, 87 (1980), pp. 183–186.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0001 $01.00/0

A THEOREM CONCERNING UNIFORM SIMPLIFICATION AT A TRANSITION POINT AND THE PROBLEM OF RESONANCE*

YASUTAKA SIBUYA†

Abstract. Given sectors $\mathcal{S}_j = \{\varepsilon\,;\, a_j < \arg\varepsilon < b_j, 0 < |\varepsilon| < \rho\}$ $(1 \leqq j \leqq \nu)$ and functions $\delta_j (1 \leqq j \leqq \nu)$ such that (i) $\cup_j \mathcal{S}_j = \{\varepsilon\,;\, 0 < |\varepsilon| < \rho\}$, (ii) δ_j is holomorphic in $\mathcal{S}_j$, (iii) δ_j is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_j$, (iv) $|\delta_j(\varepsilon) - \delta_k(\varepsilon)| \leqq c_0 \exp\{-c_1/|\varepsilon|^\lambda\}$ in $\mathcal{S}_j \cap \mathcal{S}_k$ for some positive numbers c_0, c_1 and λ whenever $\mathcal{S}_j \cap \mathcal{S}_k \neq \emptyset$, we prove that $|\delta_j(\varepsilon)| \leqq c_2 \exp\{-c_1/|\varepsilon|^\lambda\}$ in $\mathcal{S}_j$ for some positive number c_2. Then, utilizing this result, we prove that the Matkowsky condition implies resonance in the sense of N. Kopell under a reasonable assumption. The sufficiency of the Matkowsky condition with regard to Ackerberg–O'Malley resonance has been an open question. This work gives an affirmative answer to this question in a reasonably general case.

1. Introduction.

The main result of this paper is the following theorem:

THEOREM 1.1. *Let*

$$\mathcal{S}_j = \{\varepsilon\,;\, a_j < \arg\varepsilon < b_j, 0 < |\varepsilon| < \rho\}, \qquad j = 1, \cdots, \nu \tag{1.1}$$

be sectors in the complex ε-plane, where ρ is a positive number and the a's and the b's are real numbers. Let $\delta_1(\varepsilon), \cdots, \delta_\nu(\varepsilon)$ be functions of ε. Assume that

(i) $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_\nu = \{\varepsilon\,;\, 0 < |\varepsilon| < \rho\}$;

(ii) *$\delta_j(\varepsilon)$ is holomorphic in $\mathcal{S}_j$;*

(iii) *$\delta_j(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_j$, i.e.,*

$$|\delta_j(\varepsilon)| \leqq K_N |\varepsilon|^N, \qquad N = 0, 1, \cdots \quad \text{in } \mathcal{S}_j$$

for some positive numbers K_N;

(iv) *if $\mathcal{S}_j \cap \mathcal{S}_k \neq \emptyset$, we have*

$$|\delta_j(\varepsilon) - \delta_k(\varepsilon)| \leqq c_0 \exp\left(\frac{-c_1}{|\varepsilon|^\lambda}\right) \quad \text{in } \mathcal{S}_j \cap \mathcal{S}_k, \tag{1.2}$$

for some positive numbers c_0, c_1 and λ.

Then there exists a positive number H such that

$$|\delta_j(\varepsilon)| \leqq H \exp\left(\frac{-c_1}{|\varepsilon|^\lambda}\right) \quad \text{in } \mathcal{S}_j, \qquad j = 1, 2, \cdots, \nu. \tag{1.3}$$

We shall prove this theorem in § 8. (For another proof, see J.-P. Ramis [5, Thm. 11(i), p. 189].) In other sections, utilizing Theorem 1.1, we shall treat the following problem.

We consider a differential equation

$$\varepsilon \frac{d^2 v}{dx^2} + F(x, \varepsilon)\frac{dv}{dx} + G(x, \varepsilon)v = 0, \tag{1.4}$$

where F and G are holomorphic in two complex variables x and ε in a domain

$$x \in \mathcal{D}_0, \qquad |\varepsilon| < \rho_0, \tag{1.5}$$

where $\mathcal{D}_0$ is a domain in the x-plane and ρ_0 is a positive number. We assume that $\mathcal{D}_0$

* Received by the editors August 7, 1980. This work was partially sponsored by the National Science Foundation under grant MCS79-01998 and by the U.S. Army under contract DAAG29-80-C-0041. This paper was prepared while the author was at the Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

contains a real interval

$$\mathcal{I}_0 = \{x\,;\, -a \leqq \mathrm{Re}\,(x) \leqq b,\ \mathrm{Im}\,(x) = 0\}, \tag{1.6}$$

where a and b are positive numbers. We also assume that

$$F(x, 0) = -2x. \tag{1.7}$$

We say that the differential equation (1.4) satisfies the *Matkowsky condition*, if there exists a nontrivial formal power series solution of (1.4),

$$v = \sum_{m=0}^{\infty} a_m(x)\varepsilon^m, \tag{1.8}$$

such that all the $a_m(x)$ are bounded on the real interval $\mathcal{I}_0$. We also say that the differential equation (1.4) exhibits *resonance in the sense of N. Kopell* on $\mathcal{I}_0$ if there exists a solution $v(x, \varepsilon)$ satisfying $v(b, \varepsilon) = 1$, such that $v(x, \varepsilon)$ converges uniformly on $\mathcal{I}_0$ as $\varepsilon \to +0$ to a nontrivial solution of

$$F(x, 0)\frac{dv}{dx} + G(x, 0)v = 0 \tag{1.9}$$

(cf. B. J. Matkowsky [4] and N. Kopell [2]).

We shall prove the following theorem:

THEOREM 1.2. *If $\mathcal{D}_0$ is a disk with the center at $x = 0$, i.e.,*

$$\mathcal{D}_0 = \{x\,;\, |x| < r_0\} \quad \textit{for some } r_0 > 0, \tag{1.10}$$

then the Matkowsky condition implies resonance in the sense of N. Kopell.

In our argument, the assumption that F and G are holomorphic in (x, ε) in a poly-disk (1.5) is indispensable. In our proof, we follow roughly the guideline given by R. McKelvey and R. Bohac [3]. It seems to us that our results yield a sharp estimate for eigenvalues studied by P. P. N. de Groen [1]. In § 2, we discuss a more general case.

Throughout this research, the author enjoyed lively discussions with N. Kopell, B. J. Matkowsky and P. P. N. de Groen.

2. A standard form. Let ρ_0 be a positive number and let $\mathcal{D}$ be a domain in the complex ξ-plane which contains a real interval

$$\mathcal{I} = \{\xi\,;\, -\alpha \leqq \mathrm{Re}\,(\xi) \leqq \beta,\ \mathrm{Im}\,(\xi) = 0\}, \tag{2.1}$$

where α and β are positive numbers.

We shall consider a linear differential equation:

$$\varepsilon\frac{d^2v}{d\xi^2} + f(\xi, \varepsilon)\frac{dv}{d\xi} + g(\xi, \varepsilon)v = 0, \tag{2.2}$$

where f and g are holomorphic in two variables ξ and ε in the domain

$$\xi \in \mathcal{D}, \qquad |\varepsilon| < \rho_0. \tag{2.3}$$

Set

$$f_0(\xi) = f(\xi, 0). \tag{2.4}$$

We assume that

$$f_0(0) = 0, \qquad f_0'(0) \neq 0, \tag{2.5}$$

$$\xi f_0(\xi) < 0 \quad \text{for } \xi \in \mathcal{I} \quad \text{if } \xi \neq 0. \tag{2.6}$$

Under this situation, we can write f_0 as

$$f_0(\xi) = \xi h(\xi), \tag{2.7}$$

where $h(\xi)$ is holomorphic in $\mathscr{D}$ and

$$h(\xi) < 0 \quad \text{for } \xi \in \mathscr{I}. \tag{2.8}$$

Let us change the independent variable by

$$x = \varphi(\xi) = \left\{ -\int_0^{\xi} f_0(t)\, dt \right\}^{1/2}. \tag{2.9}$$

Then, (2.2) becomes

$$\varepsilon \frac{d^2 v}{dx^2} + F(x, \varepsilon) \frac{dv}{dx} + G(x, \varepsilon) v = 0, \tag{2.10}$$

where

$$F(\varphi, \varepsilon) = (\varphi')^{-2}\{\varphi' f + \varepsilon \varphi''\}, \qquad G(\varphi, \varepsilon) = (\varphi')^{-2} g. \tag{2.11}$$

Since $f_0 = -2\varphi\varphi'$, we have

$$F(x, \varepsilon) = -2x + \varepsilon k(x, \varepsilon), \tag{2.12}$$

and $k(x, \varepsilon)$ and $G(x, \varepsilon)$ are holomorphic in a domain

$$x \in \mathscr{D}_0, \qquad |\varepsilon| < \rho_0, \tag{2.13}$$

where $\mathscr{D}_0$ is a domain in the x-plane which contains the real interval

$$\mathscr{I}_0 = \{x;\ -a \leqq \operatorname{Re}(x) \leqq b,\ \operatorname{Im}(x) = 0\}, \tag{2.14}$$

where

$$a = \sqrt{-\int_0^{-\alpha} f_0(t)\, dt}, \qquad b = \sqrt{-\int_0^{\beta} f_0(t)\, dt}. \tag{2.15}$$

Another transformation:

$$v = w \exp\left\{ -\frac{1}{2\varepsilon} \int_0^x F(t, \varepsilon)\, dt \right\}, \tag{2.16}$$

takes (2.10) to

$$\varepsilon^2 \frac{d^2 w}{dx^2} - \left\{ \frac{1}{4} F(x, \varepsilon)^2 + \varepsilon \left(\frac{1}{2} \frac{\partial F}{\partial x}(x, \varepsilon) - G(x, \varepsilon) \right) \right\} w = 0. \tag{2.17}$$

Note that

$$\frac{1}{4} F^2 + \varepsilon \left(\frac{1}{2} \frac{\partial F}{\partial x} - G \right) = x^2 + \varepsilon R(x, \varepsilon), \tag{2.18}$$

where R is holomorphic in (2.13).

Remark. To find the domain $\mathscr{D}_0$, we must take into account not only singularities of f and g, but also singularities of φ, i.e., the transformation (2.9). In particular, any zeros of f_0 would yield branch-points with respect to x.

3. Formal simplification. It is known that there exist three formal power series in ε:

$$A(x,\varepsilon)=\sum_{m=0}^{\infty}A_m(x)\varepsilon^m, \tag{3.1}$$

$$B(x,\varepsilon)=\sum_{m=0}^{\infty}B_m(x)\varepsilon^m, \tag{3.2}$$

$$C(\varepsilon)=\sum_{m=0}^{\infty}C_m\varepsilon^m, \tag{3.3}$$

such that

(i) $A_m(x)$ and $B_m(x)$ are holomorphic in the domain $\mathcal{D}_0$;
(ii) C_m are constants;
(iii) the formal transformation

$$w=A(x,\varepsilon)u+B(x,\varepsilon)\left(\varepsilon\frac{du}{dx}\right) \tag{3.4}$$

takes (2.17) to

$$\varepsilon^2\frac{d^2u}{dx^2}-\{x^2+\varepsilon C(\varepsilon)\}u=0; \tag{3.5}$$

(iv) we have

$$A_0(x)^2-(xB_0(x))^2=1 \quad \text{identically in } \mathcal{D}_0. \tag{3.6}$$

To effect the transformation (3.4), we differentiate both sides of (3.4) with respect to x. Then, we derive

$$\varepsilon\frac{dw}{dx}=(\varepsilon A'+(x^2+\varepsilon C)B)u+(A+\varepsilon B')\left(\varepsilon\frac{du}{dx}\right), \tag{3.7}$$

and

$$\begin{aligned}\varepsilon^2\frac{d^2w}{dx^2}&=(\varepsilon(\varepsilon A'+(x^2+\varepsilon C)B)'+(x^2+\varepsilon C)(A+\varepsilon B'))u\\&\quad+((\varepsilon A'+(x^2+\varepsilon C)B)+\varepsilon(A+\varepsilon B')')\left(\varepsilon\frac{du}{dx}\right),\end{aligned} \tag{3.8}$$

where $'$ denotes $\partial/\partial x$. Since $\varepsilon^2(d^2w/dx^2)=(x^2+\varepsilon R)w$, we derive the following equations on A, B and C:

$$\begin{aligned}(x^2+\varepsilon R)A&=\varepsilon(\varepsilon A'+(x^2+\varepsilon C)B)'+(x^2+\varepsilon C)(A+\varepsilon B'),\\(x^2+\varepsilon R)B&=(\varepsilon A'+(x^2+\varepsilon C)B)+\varepsilon(A+\varepsilon B')'.\end{aligned} \tag{3.9}$$

In particular, if we put

$$X=A_0, \qquad Y=xB_0,$$

we have

$$\frac{dX}{dx}=\frac{R_0(x)-C_0}{2x}Y, \qquad \frac{dY}{dx}=\frac{R_0(x)-C_0}{2x}X,$$

where $R_0(x) = R(x, 0)$. Hence

$$\frac{d(X^2 - Y^2)}{dx} = 0 \quad \text{identically.}$$

Choose $C_0 = R_0(0)$ and the initial condition: $X(0) = 1$, $Y(0) = 0$. Then, we can determine A_0, B_0 and C_0 so that (3.6) is satisfied. Other coefficients A_m, B_m and C_m can be determined in a similar way.

By virtue of (3.6), we can solve (3.4) and (3.7) with respect to u and $\varepsilon\, du/dx$:

$$\begin{aligned} u &= E_{11}(x, \varepsilon)w + E_{12}(x, \varepsilon)\left\{\varepsilon \frac{dw}{dx}\right\}, \\ \varepsilon \frac{du}{dx} &= E_{21}(x, \varepsilon)w + E_{22}(x, \varepsilon)\left(\varepsilon \frac{dw}{dx}\right), \end{aligned} \tag{3.10}$$

where E_{jk} are formal power series in ε whose coefficients are holomorphic in $\mathcal{D}_0$. In particular,

$$\begin{aligned} &E_{11}(x, 0) = E_{22}(x, 0) = A_0(x), \\ &E_{12}(x, 0) = -B_0(x), \qquad E_{21}(x, 0) = -x^2 B_0(x). \end{aligned} \tag{3.11}$$

Note that

$$C_0 = R_0(0) = -1 + 2\frac{g(0, 0)}{f_0'(0)}. \tag{3.12}$$

4. Outer expansions. A formal power series in ε:

$$v = \sum_{m=0}^{\infty} a_m(x)\varepsilon^m, \tag{4.1}$$

is called an outer expansion associated with the differential equation (2.10), if (4.1) formally satisfies (2.10). The power series (4.1) is an outer expansion if and only if

$$\begin{aligned} -2x\frac{da_0}{dx} + G_0(x)a_0 &= 0, \\ -2x\frac{da_m}{dx} + G_0(x)a_m &= L_m(x) - \frac{d^2 a_{m-1}(x)}{dx^2} \qquad m \geqq 1, \end{aligned} \tag{4.2}$$

where $G_0(x) = G(x, 0)$ and $L_m(x)$ is linear homogeneous in $a_0, \cdots, a_{m-1}$ and $da_0/dx, \cdots, da_{m-1}/dx$ with coefficients holomorphic in $\mathcal{D}_0$.

DEFINITION 4.1. The differential equation (2.10) is said to satisfy the Matkowsky condition, if there exists a nontrivial outer expansion (4.1) such that all the $a_m(x)$ are bounded on the real interval $\mathcal{I}_0$ (cf. (2.14)).

LEMMA 4.2. *The differential equation* (2.10) *satisfies the Matkowsky condition if and only if C_0 is a negative odd integer and*

$$C_m = 0, \qquad m \geqq 1. \tag{4.3}$$

Proof. The transformation

$$u = y \exp\left\{\frac{x^2}{2\varepsilon}\right\} \tag{4.4}$$

changes (3.5) to

$$\varepsilon\frac{d^2y}{dx^2}-2x\frac{dy}{dx}-(1+C)y=0. \tag{4.5}$$

By a straightforward computation, we can prove that the differential equation (4.5) satisfies the Matkowsky condition if and only if C_0 is a negative odd integer and $C_m=0$ for $m\geqq 1$.

Note also that, if all the a_m are bounded, then all the da_m/dx are bounded. Otherwise, d^2a_m/dx^2 would have much worse singularities at $x=0$, and hence a_{m+1} would be unbounded (cf. (4.2)).

Finally, by manipulating with the transformations (2.16), (3.4) and (3.7), and (3.10) together with (4.4), we can show that the differential equation (2.10) satisfies the Matkowsky condition if and only if the differential equation (4.5) satisfies the same condition. This completes the proof of Lemma 4.2.

5. Uniform simplification. Hereafter, we shall assume that

$$C_0=-p, \quad \text{where } p \text{ is a positive odd integer,} \tag{5.1}$$

$$C_m=0 \quad \text{for } m\geqq 1, \tag{5.2}$$

$$\mathscr{D}_0=\{x;\,|x|<r_0\} \quad \text{for some } r_0>0. \tag{5.3}$$

The assumption (5.3) means that $\mathscr{D}_0$ is a disk of radius r_0 with center at $x=0$.

Let us choose two positive numbers r_1 and r such that

$$0<r_1<r<r_0 \tag{5.4}$$

and that the disk

$$\mathscr{D}_1=\{x;\,|x|<r_1\} \tag{5.5}$$

contains the real interval $\mathscr{I}_0$ (cf. (2.14)).

Let us denote by $T(x,\varepsilon)$ the 2×2 matrix:

$$\begin{bmatrix} A(x,\varepsilon) & B(x,\varepsilon) \\ \varepsilon A'(x,\varepsilon)+(x^2-\varepsilon p)B(x,\varepsilon) & A(x,\varepsilon)+\varepsilon B'(x,\varepsilon)\end{bmatrix} \tag{5.6}$$

(cf. (3.4) and (3.7)). Set

$$U=\begin{bmatrix} u \\ \varepsilon\, du/dx\end{bmatrix}, \qquad W=\begin{bmatrix} w \\ \varepsilon\, dw/dx\end{bmatrix}. \tag{5.7}$$

Then the formal transformation

$$W=T(x,\varepsilon)U \tag{5.8}$$

takes the system

$$\varepsilon\frac{dW}{dx}=\begin{bmatrix} 0 & 1 \\ x^2+\varepsilon R(x,\varepsilon) & 0\end{bmatrix}W \tag{5.9}$$

to

$$\varepsilon\frac{dU}{dx}=\begin{bmatrix} 0 & 1 \\ x^2-\varepsilon p & 0\end{bmatrix}U. \tag{5.10}$$

The inverse of the matrix $T(x, \varepsilon)$ is given by

$$T(x, \varepsilon)^{-1} = \begin{bmatrix} E_{11}(x, \varepsilon) & E_{12}(x, \varepsilon) \\ E_{21}(x, \varepsilon) & E_{22}(x, \varepsilon) \end{bmatrix} \tag{5.11}$$

(cf. (3.10)).

Set

$$\mathscr{D}_2 = \{x\,;\, |x| < r\}. \tag{5.12}$$

It is known that there exist two positive numbers ρ_1 and ρ_2, a function $\delta(\varepsilon)$, and a 2×2 matrix $P(x, \varepsilon)$ such that

(i) $\delta(\varepsilon)$ is holomorphic in the sector

$$\mathscr{S} = \{\varepsilon\,;\, |\arg \varepsilon| < \rho_1,\ 0 < |\varepsilon| < \rho_2\}; \tag{5.13}$$

(ii) $\delta(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in $\mathscr{S}$, i.e.

$$|\delta(\varepsilon)| \leqq K_N |\varepsilon|^N, \qquad N = 0, 1, 2, \cdots \quad \text{in } \mathscr{S} \tag{5.14}$$

for some positive numbers K_N;

(iii) entries of P and P^{-1} are holomorphic in the domain

$$x \in \mathscr{D}_2, \qquad \varepsilon \in \mathscr{S}; \tag{5.15}$$

(iv) P (resp. P^{-1}) admits the matrix T (resp. T^{-1}) as an asymptotic expansion as $\varepsilon \to 0$ in $\mathscr{S}$ which is valid uniformly in $\mathscr{D}_2$;

(v) the transformation

$$W = P(x, \varepsilon) V \tag{5.16}$$

takes (5.9) to

$$\varepsilon \frac{dV}{dx} = \begin{bmatrix} 0 & 1 \\ x^2 - \varepsilon(p + \delta(\varepsilon)) & 0 \end{bmatrix} V \tag{5.17}$$

in the domain (5.15) (cf. Y. Sibuya [6]).

Utilizing this result and manipulating with rotations of the disk $\mathscr{D}_2$, we can prove the following lemma:

LEMMA 5.1. *There exist sectors*

$$\mathscr{S}_j = \{\varepsilon\,;\, a_j < \arg \varepsilon < b_j,\ 0 < |\varepsilon| < \rho_3\}, \qquad j = 1, 2, \cdots, k \tag{5.18-j}$$

(where ρ_3 is a positive number and the a's and the b's are real numbers), functions $\delta_1(\varepsilon), \cdots, \delta_k(\varepsilon)$, and 2×2 matrices $P_1(x, \varepsilon), \cdots, P_k(x, \varepsilon)$ such that $\mathscr{S}_1 \cup \cdots \cup \mathscr{S}_k = \{\varepsilon\,;\, 0 < |\varepsilon| < \rho_3\}$ and that

(i) *$\delta_j(\varepsilon)$ is holomorphic in $\mathscr{S}_j$;*

(ii) *$\delta_j(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in $\mathscr{S}_j$;*

(iii) *entries of P_j and P_j^{-1} are holomorphic in the domain*

$$x \in \mathscr{D}_2, \qquad \varepsilon \in \mathscr{S}_j; \tag{5.19-j}$$

(iv) *P_j (resp. P_j^{-1}) admits the matrix T (resp. T^{-1}) as an asymptotic expansion as $\varepsilon \to 0$ in $\mathscr{S}_j$ which is valid uniformly in $\mathscr{D}_2$;*

(v) *the transformation*

$$W = P_j(x, \varepsilon) V_j \tag{5.20}$$

takes (5.9) *to*

$$\varepsilon\frac{dV_j}{dx}=\begin{bmatrix}0 & 1\\ x^2-\varepsilon(p+\delta_j(\varepsilon)) & 0\end{bmatrix}V_j \tag{5.21-j}$$

in the domain (5.19-j).

6. An estimate for $\delta_j(\varepsilon)$. In this section, as an application of our main theorem (cf. Theorem 1.1), we shall derive an estimate

$$|\delta_j(\varepsilon)|\leqq H_j\exp\left(\frac{-r^2}{|\varepsilon|}\right)\quad\text{for } \varepsilon\in\mathcal{S}_j, \tag{6.1}$$

where H_j is a positive number. To do this, it is sufficient to prove that, if $\mathcal{S}_l\cap\mathcal{S}_j\neq\varnothing$, we have

$$|\delta_l(\varepsilon)-\delta_j(\varepsilon)|\leqq M_{lj}\exp\left(\frac{-r^2}{|\varepsilon|}\right)\quad\text{for } \varepsilon\in\mathcal{S}_l\cap\mathcal{S}_j, \tag{6.2}$$

where M_{lj} is a positive number. To derive an estimate (6.2), we need some preparation.

Let us consider the differential equation

$$\frac{d^2z}{dt^2}-(t^2-a)z=0,\quad\text{where } a \text{ is a parameter.} \tag{6.3}$$

This equation admits a solution

$$z=Z(t,a) \tag{6.4}$$

such that

(i) Z is an entire function in (t,a),

(ii) $\lim_{t\to+\infty}t^{(1-a)/2}e^{t^2/2}Z(t,a)=1$

uniformly in a if a is in a compact set in the a-plane. The solution $Z(t,a)$ is uniquely determined by (i) and (ii). The functions $Z((-i)t,-a)$, $Z(-t,a)$ and $Z(it,-a)$ are also solutions of (6.3). Set

$$\Psi_0(t,a)=\begin{bmatrix}Z(t,a) & Z((-i)t,-a)\\ Z'(t,a) & (-i)Z'((-i)t,-a)\end{bmatrix}, \tag{6.5-0}$$

$$\Psi_1(t,a)=\begin{bmatrix}Z((-i)t,-a) & Z(-t,a)\\ (-i)Z'((-i)t,-a) & -Z'(-t,a)\end{bmatrix}, \tag{6.5-1}$$

$$\Psi_2(t,a)=\begin{bmatrix}Z(-t,a) & Z(it,-a)\\ -Z'(-t,a) & iZ'(it,-a)\end{bmatrix}, \tag{6.5-2}$$

and

$$\Psi_{-1}(t,a)=\begin{bmatrix}Z(it,-a) & Z(t,a)\\ iZ'(it,-a) & Z'(t,a)\end{bmatrix}, \tag{6.5-(-1)}$$

where $'$ denotes $\partial/\partial t$. These four matrices are matrices of independent solutions of (6.3).

Set

$$\lambda_1(a)=2^{-a/2}e^{\pi i(a+1)/4}\frac{\sqrt{2\pi}}{\Gamma((1-a)/2)},\qquad \lambda_2(a)=(-i)e^{a\pi i/2}, \tag{6.6}$$

and

$$\mathscr{C}(a)=\begin{bmatrix}\lambda_1(a) & 1\\ \lambda_2(a) & 0\end{bmatrix}. \tag{6.7}$$

Then

$$\begin{aligned}&\Psi_0(t,a)=\Psi_1(t,a)\mathscr{C}(a), \qquad \Psi_1(t,a)=\Psi_2(t,a)\mathscr{C}(-a),\\ &\Psi_2(t,a)=\Psi_{-1}(t,a)\mathscr{C}(a), \qquad \Psi_{-1}(t,a)=\Psi_0(t,a)\mathscr{C}(-a).\end{aligned} \tag{6.8}$$

Fix l and j so that $\mathscr{S}_l\cap\mathscr{S}_j\neq\varnothing$. Choose a branch of $\varepsilon^{1/2}$ in the sector $\mathscr{S}_l\cap\mathscr{S}_j$. Set

$$\Lambda(\varepsilon)=\begin{bmatrix}1 & 0\\ 0 & \varepsilon^{1/2}\end{bmatrix}, \tag{6.9}$$

and

$$\begin{aligned}&\Phi_{l,h}(x,\varepsilon)=\Lambda(\varepsilon)\Psi_h\left(\frac{x}{\varepsilon^{1/2}},p+\delta_l(\varepsilon)\right),\\ &\Phi_{j,h}(x,\varepsilon)=\Lambda(\varepsilon)\Psi_h\left(\frac{x}{\varepsilon^{1/2}},p+\delta_j(\varepsilon)\right),\end{aligned} \qquad h=-1,0,1,2. \tag{6.10-h}$$

Then, $\Phi_{l,0}(x,\varepsilon)$, $\Phi_{l,1}(x,\varepsilon)$, $\Phi_{l,2}(x,\varepsilon)$ and $\Phi_{l,-1}(x,\varepsilon)$ (resp. $\Phi_{j,0}(x,\varepsilon)$, $\Phi_{j,1}(x,\varepsilon)$, $\Phi_{j,2}(x,\varepsilon)$ and $\Phi_{j,-1}(x,\varepsilon)$) are fundamental matrix solutions of (5.21-l) (resp. (5.21-j)) such that

$$\begin{aligned}&\Phi_{l,0}(x,\varepsilon)=\Phi_{l,1}(x,\varepsilon)\mathscr{C}(p+\delta_l(\varepsilon)),\\ &\Phi_{l,1}(x,\varepsilon)=\Phi_{l,2}(x,\varepsilon)\mathscr{C}(-p-\delta_l(\varepsilon)),\\ &\Phi_{l,2}(x,\varepsilon)=\Phi_{l,-1}(x,\varepsilon)\mathscr{C}(p+\delta_l(\varepsilon)),\\ &\Phi_{l,-1}(x,\varepsilon)=\Phi_{l,0}(x,\varepsilon)\mathscr{C}(-p-\delta_l(\varepsilon)),\end{aligned} \tag{6.11-l}$$

and

$$\begin{aligned}&\Phi_{j,0}(x,\varepsilon)=\Phi_{j,1}(x,\varepsilon)\mathscr{C}(p+\delta_j(\varepsilon)),\\ &\Phi_{j,1}(x,\varepsilon)=\Phi_{j,2}(x,\varepsilon)\mathscr{C}(-p-\delta_j(\varepsilon)),\\ &\Phi_{j,2}(x,\varepsilon)=\Phi_{j,-1}(x,\varepsilon)\mathscr{C}(p+\delta_j(\varepsilon)),\\ &\Phi_{j,-1}(x,\varepsilon)=\Phi_{j,0}(x,\varepsilon)\mathscr{C}(-p-\delta_j(\varepsilon)).\end{aligned} \tag{6.11-j}$$

Set

$$J=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}, \tag{6.12}$$

and

$$\begin{aligned}&Q_{l,h}(x,\varepsilon)=\Phi_{l,h}(x,\varepsilon)\exp\left\{(-1)^h\frac{x^2}{2\varepsilon}J\right\},\\ &Q_{j,h}(x,\varepsilon)=\Phi_{j,h}(x,\varepsilon)\exp\left\{(-1)^h\frac{x^2}{2\varepsilon}J\right\},\end{aligned} \qquad h=-1,0,1,2. \tag{6.13}$$

It is known that, if (x,ε) is in a domain

$$x\in\mathscr{D}_2, \quad \varepsilon\in\mathscr{S}_l\cap\mathscr{S}_j, \quad \left|\arg\left(\frac{x}{\varepsilon^{1/2}}\right)-\frac{1}{4}\pi-\frac{1}{2}h\pi\right|\leqq\frac{1}{2}\pi-\nu, \tag{6.14-h}$$

where ν is a small positive number, we have

$$\|Q_{j,h}(x,\varepsilon)\| \leqq H|\varepsilon|^q, \qquad \|Q_{j,h}(x,\varepsilon)^{-1}\| \leqq H|\varepsilon|^q, \tag{6.15}$$

where H is a positive number depending on ν, q is a real number and $\|\cdot\|$ denotes a usual norm of matrices. Furthermore, the matrix

$$Q_{j,h}(x,\varepsilon) - Q_{l,h}(x,\varepsilon) \tag{6.16}$$

is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_l \cap \mathcal{S}_j$ uniformly in the domain (6.14-h). (For these results see, for example, Y. Sibuya [6], [7].)

Let $P_l(x,\varepsilon)$ and $P_j(x,\varepsilon)$ be the matrices given in Lemma 5.1. Then, $P_l(x,\varepsilon)\Phi_{l,0}(x,\varepsilon)$ and $P_j(x,\varepsilon)\Phi_{j,0}(x,\varepsilon)$ are two fundamental matrix solutions of (5.9) in the domain

$$x \in \mathcal{D}_2, \qquad \varepsilon \in \mathcal{S}_l \cap \mathcal{S}_j. \tag{6.17}$$

Therefore, there exists a 2×2 matrix $L(\varepsilon)$ such that

$$P_l(x,\varepsilon)\Phi_{l,0}(x,\varepsilon) = P_j(x,\varepsilon)\Phi_{j,0}(x,\varepsilon)L(\varepsilon). \tag{6.18}$$

Note that $L(\varepsilon)$ does not depend on x. It follows from (6.18) that

$$\exp\left\{-\frac{x^2}{2\varepsilon}J\right\}L(\varepsilon)\exp\left\{\frac{x^2}{2\varepsilon}J\right\} = Q_{j,0}(x,\varepsilon)^{-1}P_j(x,\varepsilon)^{-1}P_l(x,\varepsilon)Q_{l,0}(x,\varepsilon). \tag{6.19}$$

Hence, the matrix

$$\exp\left\{-\frac{x^2}{2\varepsilon}J\right\}L(\varepsilon)\exp\left\{\frac{x^2}{2\varepsilon}J\right\} - 1_2 \tag{6.20}$$

is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_l \cap \mathcal{S}_j$ uniformly in the domain (6.14-0), where 1_2 is the 2×2 identity matrix.

In the same way (manipulating with the connection formulas (6.11-l) and (6.11-j)), we can prove that the matrix

$$\exp\left\{\frac{x^2}{2\varepsilon}J\right\}L_1(\varepsilon)\exp\left\{-\frac{x^2}{2\varepsilon}J\right\} - 1_2 \tag{6.21}$$

is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_l \cap \mathcal{S}_j$ uniformly in the domain (6.14-1), where

$$L_1(\varepsilon) = \mathcal{C}(p+\delta_j(\varepsilon))L(\varepsilon)\mathcal{C}(p+\delta_l(\varepsilon))^{-1}. \tag{6.22}$$

Also, the matrix

$$\exp\left\{\frac{x^2}{2\varepsilon}J\right\}L_2(\varepsilon)\exp\left\{-\frac{x^2}{2\varepsilon}J\right\} - 1_2 \tag{6.23}$$

is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_l \cap \mathcal{S}_j$ uniformly in the domain (6.14-(-1)), where

$$L_2(\varepsilon) = \mathcal{C}(-p-\delta_j(\varepsilon))^{-1}L(\varepsilon)\mathcal{C}(-p-\delta_l(\varepsilon)). \tag{6.24}$$

Set

$$\begin{aligned} L(\varepsilon) &= \begin{bmatrix} c_{11}(\varepsilon) & c_{12}(\varepsilon) \\ c_{21}(\varepsilon) & c_{22}(\varepsilon) \end{bmatrix}, \\ L_1(\varepsilon) &= \begin{bmatrix} \hat{c}_{11}(\varepsilon) & \hat{c}_{12}(\varepsilon) \\ \hat{c}_{21}(\varepsilon) & \hat{c}_{22}(\varepsilon) \end{bmatrix}, \\ L_2(\varepsilon) &= \begin{bmatrix} \tilde{c}_{11}(\varepsilon) & \tilde{c}_{12}(\varepsilon) \\ \tilde{c}_{21}(\varepsilon) & \tilde{c}_{22}(\varepsilon) \end{bmatrix}. \end{aligned} \tag{6.25}$$

Then

$$\hat{c}_{12}(\varepsilon)=\frac{\{\lambda_1(p+\delta_j(\varepsilon))c_{11}(\varepsilon)+c_{21}(\varepsilon)\}}{\lambda_2(p+\delta_l(\varepsilon))} - \frac{\lambda_1(p+\delta_l(\varepsilon))\{\lambda_1(p+\delta_j(\varepsilon))c_{12}(\varepsilon)+c_{22}(\varepsilon)\}}{\lambda_2(p+\delta_l(\varepsilon))} \tag{6.26}$$

and

$$\tilde{c}_{21}(\varepsilon)=\lambda_1(-p-\delta_l(\varepsilon))c_{11}(\varepsilon)+\lambda_2(-p-\delta_l(\varepsilon))c_{12}(\varepsilon) - \frac{\lambda_1(-p-\delta_j(\varepsilon))}{\lambda_2(-p-\delta_j(\varepsilon))}\{\lambda_1(-p-\delta_l(\varepsilon))c_{21}(\varepsilon)+\lambda_2(-p-\delta_l(\varepsilon))c_{22}(\varepsilon)\}. \tag{6.27}$$

Utilizing the fact that, for any $\varepsilon \in \mathcal{S}_l \cap \mathcal{S}_j$, there exists an $x \in \mathcal{D}_2$ such that

(i) (x, ε) is in the domain (6.14-h),

(ii) $x/\varepsilon^{1/2}$ takes either a real value or a purely imaginary value,

we derive from (6.20), (6.21) and (6.23) that

1) $c_{11}(\varepsilon)-1$ and $c_{22}(\varepsilon)-1$ are asymptotically zero

as $\varepsilon \to 0$ in $\mathcal{S}_l \cap \mathcal{S}_j$;

2) $|c_{12}(\varepsilon)| \leqq c \exp\left(\frac{-r^2}{|\varepsilon|}\right)$, $|c_{21}(\varepsilon)| \leqq c \exp\left(\frac{-r^2}{|\varepsilon|}\right)$

for $\varepsilon \in \mathcal{S}_l \cap \mathcal{S}_j$, where c is a positive constant;

3) $|\hat{c}_{12}(\varepsilon)| \leqq c \exp\left(\frac{-r^2}{|\varepsilon|}\right)$ for $\varepsilon \in \mathcal{S}_l \cap \mathcal{S}_j$;

4) $|\tilde{c}_{21}(\varepsilon)| \leqq c \exp\left(\frac{-r^2}{|\varepsilon|}\right)$ for $\varepsilon \in \mathcal{S}_l \cap \mathcal{S}_j$.

Set $\mu(a)=\lambda_1(a)/\lambda_2(a)$. Then

$$\begin{aligned}&|\mu(-p-\delta_l(\varepsilon))c_{11}(\varepsilon)-\mu(-p-\delta_j(\varepsilon))c_{22}(\varepsilon)| \\ &\quad=|\mu(-p-\delta_l(\varepsilon))\{c_{11}(\varepsilon)-c_{22}(\varepsilon)\}+\{\mu(-p-\delta_l(\varepsilon))-\mu(-p-\delta_j(\varepsilon))\}c_{22}(\varepsilon)| \\ &\quad\leqq \tilde{c} \exp\left(\frac{-r^2}{|\varepsilon|}\right) \quad \text{in } \mathcal{S}_l \cap \mathcal{S}_j\end{aligned}$$

for some $\tilde{c}>0$. Since $\mu(-p) \neq 0$, we have

$$|c_{11}(\varepsilon)-c_{22}(\varepsilon)| \leqq c_1|\delta_l(\varepsilon)-\delta_j(\varepsilon)|+c_2 \exp\left(\frac{-r^2}{|\varepsilon|}\right) \tag{6.28}$$

in $\mathcal{S}_l \cap \mathcal{S}_j$ for some $c_1>0$ and $c_2>0$. On the other hand,

$$\begin{aligned}&|\lambda_1(p+\delta_j(\varepsilon))c_{11}(\varepsilon)-\lambda_1(p+\delta_l(\varepsilon))c_{22}(\varepsilon)| \\ &\quad=|\{\lambda_1(p+\delta_j(\varepsilon))-\lambda_1(p+\delta_l(\varepsilon))\}c_{11}(\varepsilon)+\lambda_1(p+\delta_l(\varepsilon))\{c_{11}(\varepsilon)-c_{22}(\varepsilon)\}| \\ &\quad\leqq \hat{c} \exp\left(\frac{-r^2}{|\varepsilon|}\right) \quad \text{in } \mathcal{S}_l \cap \mathcal{S}_j\end{aligned}$$

for some $\hat{c}>0$. Since $\lambda_1(p)=0$ and $(d\lambda_1/da)(p)\neq 0$, we have

$$|\delta_l(\varepsilon)-\delta_j(\varepsilon)|\leqq c_3|\lambda_1(p+\delta_l(\varepsilon))|\,|c_{11}(\varepsilon)-c_{22}(\varepsilon)|+c_4\exp\left(\frac{-r^2}{|\varepsilon|}\right)\quad\text{in } \mathcal{S}_l\cap\mathcal{S}_j \tag{6.29}$$

for some $c_3>0$ and $c_4>0$. An estimate (6.2) follows from (6.28) and (6.29).

7. Resonance. In this section we shall prove Theorem 1.2. To do this, we return to § 5. We proved there that the transformation (5.16) takes the system (5.9) to (5.17) in the domain (5.15). The function $\delta(\varepsilon)$ satisfies the condition (5.14). We replace (5.14) by

$$|\delta(\varepsilon)|\leqq H\exp\left(\frac{-r^2}{|\varepsilon|}\right)\quad\text{in } \mathcal{S} \tag{7.1}$$

for some positive number H.

Set

$$\begin{aligned}\Phi_h(x,\varepsilon)&=\Lambda(\varepsilon)\Psi_h(x/\varepsilon^{1/2},p+\delta(\varepsilon)),\\ \tilde{\Phi}_h(x,\varepsilon)&=\Lambda(\varepsilon)\Psi_h(x/\varepsilon^{1/2},p),\end{aligned}\qquad h=-1,0,1,2. \tag{7.2}$$

Then, $\Phi_h(x,\varepsilon)$ (resp. $\tilde{\Phi}_h(x,\varepsilon)$) are fundamental matrix solutions of (5.17) (resp. (5.10)) such that

$$\begin{aligned}\Phi_0(x,\varepsilon)&=\Phi_1(x,\varepsilon)\mathscr{C}(p+\delta(\varepsilon)),\\ \Phi_1(x,\varepsilon)&=\Phi_2(x,\varepsilon)\mathscr{C}(-p-\delta(\varepsilon)),\\ \Phi_2(x,\varepsilon)&=\Phi_{-1}(x,\varepsilon)\mathscr{C}(p+\delta(\varepsilon)),\\ \Phi_{-1}(x,\varepsilon)&=\Phi_0(x,\varepsilon)\mathscr{C}(-p-\delta(\varepsilon)),\end{aligned} \tag{7.3}$$

and

$$\begin{aligned}\tilde{\Phi}_0(x,\varepsilon)&=\tilde{\Phi}_1(x,\varepsilon)\mathscr{C}(p),\\ \tilde{\Phi}_1(x,\varepsilon)&=\tilde{\Phi}_2(x,\varepsilon)\mathscr{C}(-p),\\ \tilde{\Phi}_2(x,\varepsilon)&=\tilde{\Phi}_{-1}(x,\varepsilon)\mathscr{C}(p),\\ \tilde{\Phi}_{-1}(x,\varepsilon)&=\tilde{\Phi}_0(x,\varepsilon)\mathscr{C}(-p).\end{aligned} \tag{7.4}$$

Set

$$S(x,\varepsilon)=\Phi_0(x,\varepsilon)\tilde{\Phi}_0(x,\varepsilon)^{-1}. \tag{7.5}$$

Then the transformation

$$V=S(x,\varepsilon)U \tag{7.6}$$

takes (5.17) to (5.10). Hence, the main part of the proof is to show that $S(x,\varepsilon)-1_2$ is asymptotically zero as $\varepsilon\to 0$ in $\mathcal{S}$ uniformly in $\mathcal{D}_1$. Note that $r_1<r$. To do this we manipulate in a way similar to the arguent in § 6, utilizing the fact that

(i) $\delta(\varepsilon)\exp\{x^2/\varepsilon\}$ and $\delta(\varepsilon)\exp\{-x^2/\varepsilon\}$ are asymptotically zero as $\varepsilon\to 0$ in $\mathcal{S}$ uniformly in $\mathcal{D}_1$;

(ii) $\mathscr{C}(p+\delta(\varepsilon))\mathscr{C}(p)^{-1}-1_2=O(\delta(\varepsilon))$;

(iii) $\mathscr{C}(-p-\delta(\varepsilon))\mathscr{C}(-p)^{-1}-1_2=O(\delta(\varepsilon))$.

The details are left to the reader.

8. Proof of Theorem 1.1. We shall prove Theorem 1.1 in the case when $\nu=3$. The general case can be treated in the same manner. We shall consider three sectors $\mathcal{S}_1,\mathcal{S}_2$,

$\mathcal{S}_3$ as shown in Fig. 1. We denote by $\mathcal{S}_{1,2}, \mathcal{S}_{2,3}, \mathcal{S}_{3,1}$ the intersections $\mathcal{S}_1 \cap \mathcal{S}_2, \mathcal{S}_2 \cap \mathcal{S}_3$, $\mathcal{S}_3 \cap \mathcal{S}_1$, respectively.

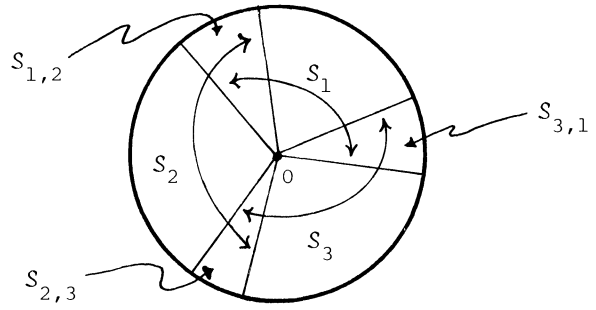


FIG. 1

The three functions $\delta_1(\varepsilon), \delta_2(\varepsilon), \delta_3(\varepsilon)$ are holomorphic in $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, respectively. Furthermore,

(8.1) $\delta_j(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in $\mathcal{S}_j$,

and

$$|\delta_{j+1}(\varepsilon) - \delta_j(\varepsilon)| \leqq c_0 \exp\left(\frac{-c_1}{|\varepsilon|^\lambda}\right) \quad \text{in } \mathcal{S}_{j,j+1}, \tag{8.2}$$

where c_0, c_1, λ are positive numbers and $\mathcal{S}_{3,4} = \mathcal{S}_{3,1}$, $\delta_4 = \delta_1$. We shall denote $\delta_{j+1}(\varepsilon) - \delta_j(\varepsilon)$ by $\sigma_j(\varepsilon)$.

We consider a sufficiently small disk:

$$\mathcal{D} = \{\varepsilon\,;\, |\varepsilon| \leqq \rho_0\}. \tag{8.3}$$

We choose three line-segments l_1, l_2, l_3 starting from $\varepsilon = 0$ in such a way that

(8.4) $l_j \subset \mathcal{S}_{j,j+1}$ (cf. Fig. 2).

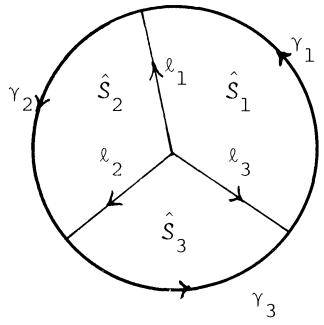


FIG. 2

Three line-segments l_1, l_2, l_3 divide the disk $\mathcal{D}$ (cf. (8.3)) into three open sectors $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$, $\hat{\mathcal{S}}_3$ (cf. Fig. 2). The boundaries of $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$, $\hat{\mathcal{S}}_3$ are respectively

(8.5-j) $l_{j-1} + \gamma_j - l_j, \qquad j = 1, 2, 3,$

where $l_0 = l_3$, and the γ's are circular arcs such that

$$\gamma_1 + \gamma_2 + \gamma_3 = \mathscr{C} = \{\varepsilon\,;\, |\varepsilon| = \rho_0\}. \tag{8.6}$$

The line-segments l_j and the circular arcs γ_j are oriented as indicated in Fig. 2. We assume that ρ_0 is so small that

$$\bar{\hat{\mathscr{S}}}_j \subset \mathscr{S}_j, \tag{8.7}$$

where $\bar{\hat{\mathscr{S}}}_j$ denotes the closure of $\hat{\mathscr{S}}_j$.

Set, for $\varepsilon \in \hat{\mathscr{S}}_1 \cup \hat{\mathscr{S}}_2 \cup \hat{\mathscr{S}}_3$,

$$\delta(\varepsilon) = \delta_j(\varepsilon) \quad \text{if } \varepsilon \in \hat{\mathscr{S}}_j. \tag{8.8}$$

Since

$$\frac{1}{2\pi i}\int_{l_{j-1}+\gamma_j-l_j} \frac{\delta_j(\xi)}{\xi-\varepsilon}\,d\xi = \begin{cases} \delta_j(\varepsilon), & \varepsilon \in \hat{\mathscr{S}}_j, \\ 0 & \varepsilon \notin \hat{\mathscr{S}}_j, \end{cases}$$

we have

$$\delta(\varepsilon) = \frac{1}{2\pi i}\sum_{j=1}^{3}\int_{l_{j-1}+\gamma_j-l_j} \frac{\delta_j(\varepsilon)}{\xi-\varepsilon}\,d\xi \quad \text{in } \hat{\mathscr{S}}_1 \cup \hat{\mathscr{S}}_2 \cup \hat{\mathscr{S}}_3.$$

Utilizing

$$\frac{1}{\xi-\varepsilon} = \sum_{m=0}^{N} \xi^{-(m+1)}\varepsilon^m + \frac{\varepsilon^{N+1}}{\xi^{N+1}(\xi-\varepsilon)},$$

we derive

$$\delta(\varepsilon) = \frac{1}{2\pi i}\sum_{m=0}^{N}\left\{\sum_{j=1}^{3}\int_{l_{j-1}+\gamma_j-l_j} \xi^{-(m+1)}\,\delta_j(\xi)\,d\xi\right\}\varepsilon^m + \left\{\frac{1}{2\pi i}\sum_{j=1}^{3}\int_{l_{j-1}+\gamma_j-l_j} \frac{\delta_j(\xi)}{\xi^{N+1}(\xi-\varepsilon)}\,d\xi\right\}\varepsilon^{N+1}.$$

Since $\delta(\varepsilon)$ is asymptotically zero as $\varepsilon \to 0$ in $\hat{\mathscr{S}}_1 \cup \hat{\mathscr{S}}_2 \cup \hat{\mathscr{S}}_3$, the first term must be zero, and hence

$$\delta(\varepsilon) = \left\{\frac{1}{2\pi i}\sum_{j=1}^{3}\int_{l_{j-1}+\gamma_j-l_j} \frac{\delta_j(\varepsilon)}{\xi^{N+1}(\xi-\varepsilon)}\,d\xi\right\}\varepsilon^{N+1}.$$

Thus we arrive at the following formula:

$$\delta(\varepsilon) = \frac{1}{2\pi i}\left\{\sum_{j=1}^{3}\int_{l_j} \frac{\sigma_j(\xi)}{\xi^N(\xi-\varepsilon)}\,d\xi + \int_{\mathscr{C}} \frac{\delta(\xi)}{\xi^N(\xi-\varepsilon)}\,d\xi\right\}\varepsilon^N \tag{8.9}$$

for $\varepsilon \in \hat{\mathscr{S}}_1 \cup \hat{\mathscr{S}}_2 \cup \hat{\mathscr{S}}_3$ and $N = 1, 2, 3, \cdots$, where $\sigma_j = \delta_{j+1} - \delta_j$.

Construct three open sectors $\tilde{\mathscr{S}}_1, \tilde{\mathscr{S}}_2, \tilde{\mathscr{S}}_3$ as shown in Fig. 3, where $0 < \rho_1 < \rho_0$ and θ is a small positive number. Then

$$\left|\int_{\mathscr{C}} \frac{\delta(\xi)}{\xi^N(\xi-\varepsilon)}\,d\xi\right| \leqq \frac{C_0}{\rho_0^{N-1}}\,\frac{1}{\rho_0-\rho_1}$$

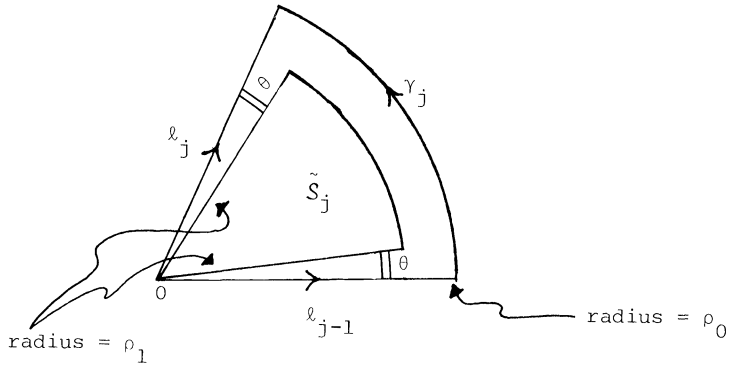


FIG. 3

and

$$\left|\int_{l_j} \frac{\sigma_j(\xi)}{\xi^N(\xi-\varepsilon)}\,d\varepsilon\right| \leqq \frac{c_0}{\sin\theta}\int_0^{\rho_0} t^{-N-1}\exp(-c_1 t^{-\lambda})\,dt$$

$$< \frac{c_0}{\lambda\sin\theta}\int_0^{+\infty} \tau^{(N\lambda^{-1}-1)}\exp(-c_1\tau)\,d\tau$$

$$= \frac{c_0}{\lambda\sin\theta} c_1^{-(N/\lambda)}\Gamma\left(\frac{N}{\lambda}\right)$$

for $\varepsilon \in \tilde{\mathscr{S}}_1 \cup \tilde{\mathscr{S}}_2 \cup \tilde{\mathscr{S}}_3$, where C_0 is a positive number. Since

$$\Gamma\left(\frac{N}{\lambda}\right) \leqq C_1\left(\frac{N}{\lambda}\right)^{(N/\lambda)} e^{-(N/\lambda)}$$

for some $C_1 > 0$, we have

$$|\delta(\varepsilon)| \leqq C_2\left(\frac{|\varepsilon|^\lambda N}{c_1\lambda}\right)^{(N/\lambda)} e^{-(N/\lambda)} \tag{8.10}$$

for $\varepsilon \in \tilde{\mathscr{S}}_1 \cup \tilde{\mathscr{S}}_2 \cup \tilde{\mathscr{S}}_3$; C_1 is a positive number. For a given ε, choose N so that

$$\frac{N}{\lambda} < \frac{c_1}{|\varepsilon|^\lambda} \leqq \frac{N+1}{\lambda}.$$

Then, it follows from (8.10) that

$$|\delta(\varepsilon)| \leqq C_2\, e^{1/\lambda}\exp\left(\frac{-c_1}{|\varepsilon|^\lambda}\right). \tag{8.11}$$

Choosing l_1, l_2, l_3 in various ways, we can complete the proof of Theorem 1.1.

REFERENCES

[1] P. P. N. DE GROEN, *The nature of resonance in a singular perturbation problem of turning point type*, this Journal, 11 (1980), pp. 1–22.

[2] N. KOPELL, *A geometric approach to boundary layer problems exhibiting resonance*, SIAM J. Appl. Math., 37 (1979), pp. 436–458.

[3] R. McKELVEY AND R. BOHAC, *Ackerberg–O'Malley resonance revisited*, Rocky Mountain J. Math., 6 (1976), pp. 637–650.
[4] B. J. MATKOWSKY, *On boundary layer problems exhibiting resonance*, SIAM Rev., 17 (1975), pp. 82–100.
[5] J. P. RAMIS, *Dévissage Gevrey*, Astérisque, 59–60 (1978), pp. 173–204.
[6] Y. SIBUYA, *Uniform simplification in a full neighborhood of a transition point*, Memoirs AMS, 149, American Mathematical Society, Providence, RI, 1974.
[7] ———, *Global Theory of a Second Order Linear Ordinary Differential Equation with a Polynomial Coefficient*, North–Holland, Amsterdam, 1975.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0002 $01.00/0

ASYMPTOTIC BEHAVIOR OF SOLUTIONS TO MULTIPLE LOOP POSITIVE FEEDBACK SYSTEMS*

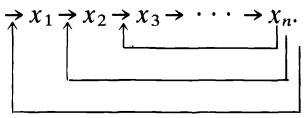
JAMES F. SELGRADE†

Abstract. This paper shows that the asymptotic behavior of solutions to a system of ordinary differential equations which models multiple loop positive feedback in biochemical control circuits is similar to the asymptotic behavior for single loop positive feedback systems. Specifically, the positive orthant is positively invariant and positive-time solutions are bounded. The critical points in the positive orthant are ordered by strict inequality. Each critical point is either asymptotically stable or unstable. Each nondegenerate unstable critical point has two orbits leaving it in opposite directions and each of these orbits is asymptotic to the adjacent critical point. The regions of attraction of the critical points are studied. In particular, for dimensions two and three, the stable manifolds of the unstable critical points separate the positive orthant into regions of attraction. Thus the orbit of each point is asymptotic to some critical point.

1. Introduction. Product feedback occurs in many biological and biochemical processes [1], [4], [5], [6], [8], [9], [13], [14], [16], [17], [18]. Here we study a nonlinear system of ordinary differential equations for multiple loop end-product positive feedback (see [1] for negative feedback). The n-dimensional system has the form

$$\text{(1)} \qquad \begin{cases} \dot{x}_1 = f_1(x_1, x_n) - h_1(x_1), \\ \dot{x}_i = f_i(x_{i-1}, x_i, x_n) - h_i(x_i), \qquad 2 \leqq i \leqq n-1, \\ \dot{x}_n = f_n(x_{n-1}, x_n) - h_n(x_n), \end{cases}$$

where, for $1 \leqq i \leqq n$, x_i is a real function of time t and f_i, h_i are C^1 functions with properties to be discussed shortly. This system has been used to model a cellular process for control of gene expression in enzyme synthesis [6], [13], [18] and also to model a system of enzyme-catalyzed reactions where enzyme activity, rather than synthesis, is induced [18]. In these cases, the variables in (1) represent chemical concentrations and the system of reactions has the structure



Let the positive orthant in $\mathbb{R}^n$ be denoted by $\mathcal{H}$:

$$\mathcal{H} \equiv \{x = (x_1, x_2, \cdots, x_n) : x_i \geqq 0, 1 \leqq i \leqq n\}.$$

For notational convenience, we consider the subscripts on the variables x_i modulo n, i.e., let $x_0 \equiv x_n$. For $1 \leqq i \leqq n$, we assume f_i and h_i are C^1 functions in a neighborhood of $\mathcal{H}$ satisfying the following conditions on $\mathcal{H}$:

(A1) $f_i > 0$ if $x_{i-1} > 0$ and $h_i > 0$ if $x_i > 0$ with $h_i(0) = 0$;

(A2) f_i is bounded and $h_i(x_i) \to \infty$ as $x_i \to \infty$;

(A3) $h_i' > 0$ if $x_i \geqq 0$; $\partial f_i / \partial x_{i-1} > 0$ if $x_{i-1} > 0$;
$\partial f_i / \partial x_i \leqq 0$; $\partial f_i / \partial x_n \geqq 0$ for $i > 1$.

Assumption (A1) means that the presence of x_{i-1} causes the production of x_i and that x_i inhibits its own production. (A2) prevents the solutions to (1) from becoming arbitrarily large. (A3) makes the feedback positive and allows the possibility of feedback by x_n to all the other variables. We add a condition on the critical points of (1):

* *Received by the editors May 6, 1980, and in revised form November 19, 1980.*

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

(A4) Equation (1) has exactly k critical points in $\mathcal{H}$ and $\partial f_i/\partial x_i = 0$ at each critical point.

According to (A4), at a critical point the linear variation of the ith component of the vector field in the x_i variable is contained in h_i'. These assumptions are analogous to those of Mees and Rapp [13] for negative feedback.

A critical point is *degenerate* if the Jacobian matrix, at the critical point, of the right-hand side of (1) has a zero eigenvalue. If at least one of the critical points is degenerate, we refer to this situation as the *degenerate case*. Then the *nondegenerate case* will be the case where no critical point is degenerate. Let $x, y \in \mathbb{R}^n$. There is a partial order on $\mathbb{R}^n$ defined by $x \leqq y$ ($x < y$) if and only if $x_i \leqq y_i$ ($x_i < y_i$) for all i, $1 \leqq i \leqq n$.

For (1) with assumptions (A1) through (A4), we show that $\mathcal{H}$ is invariant for the positive-time solution flow, and positive-time orbits in $\mathcal{H}$ are bounded. Also the critical points are ordered $\mathcal{C}_1 < \mathcal{C}_2 < \cdots < \mathcal{C}_k$ and each $\mathcal{C}_i$ is either asymptotically stable or unstable. In the nondegenerate case, the critical points alternate between stable and unstable. Each unstable point generates an orbit asymptotic (in positive time) to the adjacent stable critical point. In the degenerate case, there may be several adjacent unstable critical points. In either case, information about the domains of attraction of the critical points can be obtained. If $u, v \in \mathcal{H}$ and $u \leqq v$ then define the *rectangular box*

$$B(u, v) \equiv \{x \in \mathcal{H}: u \leqq x \leqq v\}.$$

It follows that all orbits in the interior of $B(\mathcal{C}_i, \mathcal{C}_{i+1})$ are asymptotic to the same critical point, either $\mathcal{C}_i$ or $\mathcal{C}_{i+1}$. Also, if $n = 2$ or $n = 3$, the stable manifolds of the unstable critical points separate $\mathcal{H}$ into regions of attraction. Thus the orbit of each point in $\mathcal{H}$ is asymptotic to some critical point.

These results are the same as those for the single loop feedback system [17] where $h_i(x_i) \equiv x_i$ and $f_i(x_{i-1}, x_i, x_n) \equiv x_{i-1}$, $i > 1$. However the critical points of (1) do not lie on a line as in [17]. Also, for $n = 3$, the effect of x_1 and x_3 on $\dot{x}_2$ complicates the spiralling of orbits in $B(\mathcal{C}_1, \mathcal{C}_k)$. Below we concentrate on the new arguments needed to prove these results for (1).

2. Background. If $\mathcal{S}$ is a subset of $\mathbb{R}^n$, let Int $\mathcal{S}$, $\partial\mathcal{S}$, and Cl $\mathcal{S}$ denote its topological interior, boundary, and closure, respectively. If $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}^n$, define $\mathcal{S}_1 \backslash \mathcal{S}_2$ to be the set of points in $\mathcal{S}_1$ that are not in $\mathcal{S}_2$. Let θ denote the origin in $\mathbb{R}^n$. In vector form, an autonomous system of ordinary differential equations in $\mathbb{R}^n$ is denoted

$$\dot{x} = G(x). \tag{2}$$

The C^1 function G is called a *vector field*. The unique solution to (2) at time t with initial condition $x \in \mathbb{R}^n$ is written $x \cdot t$. The solution curve is referred to as the *orbit* of x. When we are discussing the components of a solution curve, the functional notation $x(t)$ is often more convenient than $x \cdot t$. We assume solutions exist for all $t \geqq 0$.

If $T \subset \mathbb{R}$ then define $x \cdot T \equiv \bigcup_{t \in T} x \cdot t$. Likewise, if $\mathcal{S} \subset \mathbb{R}^n$ and $T \subset \mathbb{R}$, $\mathcal{S} \cdot T \equiv \bigcup_{x \in \mathcal{S}} x \cdot T$. $\mathcal{S}$ is *invariant* if $\mathcal{S} \cdot t \subset \mathcal{S}$ for all $t \in \mathbb{R}$ and $\mathcal{S}$ is *positively invariant* if $\mathcal{S} \cdot t \subset \mathcal{S}$ for all $t \geqq 0$. For $\Lambda \subset \mathbb{R}^n$, define the *$\omega$-limit set* of Λ by

$$\omega(\Lambda) \equiv \bigcap_{t > 0} \text{Cl}\,(\Lambda \cdot [t, \infty));$$

$\omega(\Lambda)$ is an invariant set. A compact set $A \subset \mathbb{R}^n$ is called an *attractor* if A has a closed neighborhood $N \subset \mathbb{R}^n$ such that $\omega(N) = A$. A bounded set $\mathcal{S} \subset \mathbb{R}^n$ is an *attracting region* if $\mathcal{S}$ is positively invariant and has a closed neighborhood N such that $\omega(N) \subset \mathcal{S}$. Thus an attracting region contains an attractor. An attracting region which is rectangular is

called an *attracting box*. The *domain of attraction* of an attracting region $\mathscr{S}$, dom $\mathscr{S}$, is defined by

$$\text{dom } \mathscr{S} \equiv \{x \in \mathbb{R}^n : \omega(x) \subset \mathscr{S}\}.$$

If a critical point $\mathscr{C}$ of (2) is an attracting region then $\mathscr{C}$ is an attractor and $\mathscr{C}$ is called an *asymptotically stable* critical point.

From (2) we get a system of equations on $\mathbb{R}^n \times \mathbb{R}^n$ given by

$$\dot{x} = G(x), \qquad \dot{v} = DG(x)v \tag{3}$$

where $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ and $DG(x)$ is the derivative matrix of G at x. The solution flow to (3) is called the *tangent flow*. The second equation of (3) is referred to as the *linearized equations* of (2) and a solution is written $v \cdot t$ or $v(t)$.

In [17] we show:

LEMMA 2.1. *Let $\mathscr{D} \subset \mathbb{R}^n$ be a domain which is positively invariant under the solution flow of* (2). *Suppose that $\partial G_i/\partial x_j \geqq 0$ on $\mathscr{D}$ for all i and j, $i \neq j$. Let $x \cdot t$ denote the solution to* (2) *where $x \in \mathscr{D}$. If $\theta \leqq G(x)$ then $x \cdot t \leqq x \cdot s$ for all $0 \leqq t \leqq s$. If $G(x) \leqq \theta$ then $x \cdot s \leqq x \cdot t$ for all $0 \leqq t \leqq s$. In either case, if the positive orbit of x is bounded then $\omega(x)$ is one critical point.*

If either property in Lemma 2.1 holds then the orbit of x is said to be *monotone*. This monotonicity is useful for finding attracting regions.

An $n \times n$ matrix A is *irreducible* if it leaves invariant no nontrivial coordinate subspaces of $\mathbb{R}^n$. A is *positive* if all its entries are positive.

The next result is due to Perron and can be found in Gantmacher [3, p. 53].

THEOREM 2.2. *A positive matrix A always has a positive eigenvalue μ (called the principal eigenvalue of A) which is a simple root of the characteristic equation and exceeds the moduli of all other eigenvalues of A. To μ there corresponds an eigenvector (called the principal eigenvector of A) with positive components.*

3. Some general results for (1). Let F denote the vector field of (1). The next two results follow directly from our assumptions but the proofs are somewhat different from those in [17].

PROPOSITION 3.1. *If* (1) *satisfies* (A1) *then* Int $\mathscr{H}$ *is positively invariant.*

Proof. Take $x \in \text{Int } \mathscr{H}$ and suppose the positive orbit of x leaves Int $\mathscr{H}$. Then there is a first time s so that the orbit of x meets $\partial\mathscr{H}$. Thus at least one component of $x(s)$, say $x_k(s)$, is zero; and $x_j(t) > 0$ for all j and t, $0 \leqq t < s$. Since $f_k(x_{k-1}(t), x_k(t), x_n(t)) \geqq 0$ for all $0 \leqq t \leqq s$, we have that $x_k(t)$ satisfies the differential inequality $\dot{y} \geqq -h_k(y)$ for all $0 \leqq t < s$. Thus $x_k(t) \geqq y(t)$ for all $0 \leqq t < s$ where $y(t)$ is the solution to $\dot{y} = -h_k(y)$ with $y(0) = x_k(0)$. But $y(t) > 0$ for all $t \geqq 0$ since $h_k(0) = 0$. Hence $x_k(s) \geqq y(s) > 0$, which is a contradiction. □

PROPOSITION 3.2. *If* (1) *satisfies* (A1) *then $\mathscr{H}$ is positively invariant. In fact, $(\mathscr{H} \backslash \theta) \cdot t \subset$ Int $\mathscr{H}$ for all $t > 0$. If $F(\theta) \neq \theta$ then $\mathscr{H} \cdot t \subset$ Int $\mathscr{H}$ for all $t > 0$.*

Proof. The first assertion follows from Proposition 3.1 because of the continuity of the solution flow. Thus the orbit of each point in $\partial\mathscr{H} \backslash \theta$ must enter Int $\mathscr{H}$ immediately or remain in $\partial\mathscr{H}$ temporarily. However, examining F on $\partial\mathscr{H}$ shows the latter is impossible. The reason is that for each face or edge of $\partial\mathscr{H} \backslash \theta$ we have two coordinates x_{i-1} and x_i such that $x_i = 0$ for this edge but $x_{i-1} \neq 0$. Hence $F_i(x) = f_i(x_{i-1}, x_i, x_n) > 0$ on this edge and so the vector field is not tangent to this edge. This completes the proof. □

Since (1) satisfies (A3), the positive-time solution flow preserves the partial order on Int $\mathscr{H}$, [2], [10], [12], [16], [17], i.e., if $x, y \in \mathscr{H}$ and $x \leqq y$ then $x \cdot t \leqq y \cdot t$ for all $t \geqq 0$. Also Lemma 2.1 applies to (1).

PROPOSITION 3.3. *If* (1) *satisfies* (A2) *and* (A3) *then positive orbits of points in* $\mathcal{H}$ *are bounded.*

Proof. From (A2) each point y of $\mathcal{H}$, all of whose components are large enough, has $F(y) \leqq \theta$. Lemma 2.1 gives that the positive orbit of y is monotone nonincreasing. Thus the positive orbit of each point in $\mathcal{H}$ is contained within the box $B(\theta, y)$ for some such y. □

Let $f_{i,j} \equiv \partial f_i / \partial x_j$. Using (A4), the linearized equations of (1) at a critical point $\mathcal{C} = (c_1, \cdots, c_n)$ are:

$$(4) \qquad \begin{cases} \dot{v}_1 = v_n f_{1,n}(c_1, c_n) - v_1 h_1'(c_1), \\ \dot{v}_i = v_{i-1} f_{i,i-1}(c_{i-1}, c_i, c_n) + v_n f_{i,n}(c_{i-1}, c_i, c_n) - v_i h_i'(c_i), \qquad 1 < i < n. \\ \dot{v}_n = v_{n-1} f_{n,n-1}(c_{n-1}, c_n) - v_n h_n'(c_n), \end{cases}$$

If the critical point $\mathcal{C}$ is not the origin θ, we have that $f_{i,i-1} > 0$ and $h_i' > 0$ for $i \geqq 1$ and $f_{i,n} \geqq 0$ for $i > 1$ because of (A3). The right-hand side of (4), $DF(\mathcal{C})v$, is a linear vector field which satisfies (A1). Proposition 3.2 implies that for $v \neq \theta$ and for all $t > 0$

$$(5) \qquad \theta < \exp(tDF(\mathcal{C}))v.$$

Hence the fundamental matrix $\exp(tDF(\mathcal{C}))$ is positive for all $t > 0$. If $\mathcal{C} = \theta$, it may happen that each $f_{i,i-1}$ is zero. If so, $DF(\theta)$ is upper triangular and thus θ is asymptotically stable. If $\mathcal{C} = \theta$ but some $f_{i,i-1}$ is nonzero, then we need a condition to guarantee that $\exp(tDF(\theta))$ is positive. This condition is irreducibility. If $\mathcal{C} \neq \theta$, $DF(\mathcal{C})$ is irreducible by virtue of (A3) but $DF(\theta)$ may not be. So if θ is a critical point we must assume $DF(\theta)$ is irreducible to get that $\exp(tDF(\theta))$ is positive—see [12, Lemma 4]. To avoid this special case in subsequent discussion we now assume that $F(\theta) \neq \theta$. However, our results remain valid if $F(\theta) = \theta$ with the additional assumption that $DF(\theta)$ is irreducible, needed if θ is not asymptotically stable.

Since the trace of $DF(x)$ is negative for all $x \in \mathcal{H}$, for each critical point $\mathcal{C}$ the matrix $DF(\mathcal{C})$ has at least one eigenvalue with negative real part. So $\mathcal{C}$ has at least a 1-dimensional stable manifold.

Henceforth, we assume $F(\theta) \neq \theta$ and (A1) through (A4) and we restrict our attention to the flow of (1) in Int $\mathcal{H}$.

4. Critical points. The k critical points of (1) correspond to the zeros of a function of one variable. From (A1), (A2), and (A3) we have that $h_i(0) = 0$ and h_i is strictly increasing without bound. Also $f_i > 0$ if $x_{i-1} > 0$; and f_i is nonincreasing as a function of x_i and strictly increasing as a function of x_{i-1}. For $c \geqq 0$ we proceed recursively to define the functions $p_i(c)$, $1 \leqq i < n$. Let $p_1(c)$ be the unique, positive value for x_1 solving the equation

$$h_1(x_1) = f_1(x_1, c).$$

For $1 < i < n$, let $p_i(c)$ be the unique, positive solution to

$$h_i(x_i) = f_i(p_{i-1}(c), x_i, c).$$

Each p_i is a strictly increasing function for $c > 0$ because of (A3). Let $p(c) \equiv (p_1(c), p_2(c), \cdots, p_{n-1}(c), c)$ represent a curve in $\mathcal{H}$. The critical points of (1) are situated along p precisely where $h_n(c) = f_n(p_{n-1}(c), c)$. Define $g(c)$, $c \geqq 0$, by

$$(6) \qquad g(c) \equiv -h_n(c) + f_n(p_{n-1}(c), c).$$

The vector field F along the curve p is given by

$$F(p(c)) = (-h_1(p_1(c)) + f_1(p_1(c), c), \cdots, -h_n(c) + f_n(p_{n-1}(c), c))$$
$$= (0, 0, \cdots, 0, g(c)).$$

Thus $F(p(c)) \leqq \theta$ if $g(c) < 0$, $F(p(c)) = \theta$ if $g(c) = 0$, and $F(p(c)) \geqq \theta$ if $g(c) > 0$. So the sign of $g(c)$ determines the direction of F along p, and the critical points of F are in "1–1" correspondence with the zeros of g. From our basic assumptions, it follows that $g(0) > 0$ and $g(c) \to -\infty$ as $c \to \infty$. Let $0 < c_1 < c_2 < \cdots < c_k$ denote the zeros of g. Then the critical points of (1) are $\mathscr{C}_j = p(c_j)$, $1 \leqq j \leqq k$. Since p is a strictly increasing function of c, we have $\theta < \mathscr{C}_1 < \mathscr{C}_2 < \cdots < \mathscr{C}_k$. Also, a tedious computation involving the chain rules gives that

$$\det DF(\mathscr{C}_j) = (-1)^{n+1} g'(c_j) \prod_{i<n} h_i'(p_i(c_j)). \tag{7}$$

So the sign of $g'(c_j)$ provides some information about the eigenvalues of $DF(\mathscr{C}_j)$.

Using Lemma 2.1 and the fact that the sign of g determines the direction of F along the curve p, we find attracting boxes in $\mathscr{H}$. Detailed proofs of these results can be found in [17].

LEMMA 4.1. *Let $c > 0$. If $g(c) < 0$ then the positive orbit of the point $p(c)$ is monotone and asymptotic to the largest critical point less than $p(c)$. If $g(c) > 0$ then the positive orbit of $p(c)$ is monotone and asymptotic to the smallest critical point greater than $p(c)$.*

LEMMA 4.2. *Fix j, $1 \leqq j \leqq k$, and let $c_{j-1} = 0$ and $\mathscr{C}_{j-1} = \theta$ if $j = 1$. If $g(c) > 0$ for $c_{j-1} < c < c_j$ then $B(\mathscr{C}_j, \mathscr{C}_k)$ is an attracting box with* $\operatorname{Int} B(\mathscr{C}_{j-1}, \infty) \subset \operatorname{dom} B(\mathscr{C}_j, \mathscr{C}_k)$.

Proof. Choose c' and c'' where $c_{j-1} < c' < c_j$ and $c_k < c''$. Since $g(c') > 0$ and $g(c'') < 0$, $\theta \leqq F(p(c'))$ and $F(p(c'')) \leqq \theta$. Lemma 4.1 gives that $p(c') \cdot t \nearrow \mathscr{C}_j$ and $p(c'') \cdot t \searrow \mathscr{C}_k$ as $t \to \infty$. If $x \in B(p(c'), p(c''))$ then $p(c') \cdot t \leqq x \cdot t \leqq p(c'') \cdot t$ for all $t \geqq 0$. Thus $\omega(B(p(c'), p(c''))) \subset B(\mathscr{C}_j, \mathscr{C}_k)$, which implies the result. □

LEMMA 4.3. *Fix j, and let $c_{j+1} = \infty$ and $\mathscr{C}_{j+1} = \infty$ if $j = k$. If $g(c) < 0$ for $c_j < c < c_{j+1}$ then $B(\theta, \mathscr{C}_j)$ is an attracting box with* $\operatorname{Int} B(\theta, \mathscr{C}_{j+1}) \subset \operatorname{dom} B(\theta, \mathscr{C}_j)$.

THEOREM 4.4. *Consider $\mathscr{C}_i \leqq \mathscr{C}_j$. If $g(c) > 0$ for $c_{i-1} < c < c_i$ and $g(c) < 0$ for $c_j < c < c_{j+1}$ then $B(\mathscr{C}_i, \mathscr{C}_j)$ is an attracting box with* $\operatorname{Int} B(\mathscr{C}_{i-1}, \mathscr{C}_{j+1}) \subset \operatorname{dom} B(\mathscr{C}_i, \mathscr{C}_j)$. *In particular, if $\mathscr{C}_i = \mathscr{C}_j$ then $\mathscr{C}_j$ is an asymptotically stable critical point.*

Proof. Lemma 4.2 and Lemma 4.3 imply that $B(\mathscr{C}_i, \mathscr{C}_k)$ and $B(\theta, \mathscr{C}_j)$ are attracting boxes. Since the intersection of attracting boxes is an attracting region and $B(\mathscr{C}_i, \mathscr{C}_j) = B(\mathscr{C}_i, \mathscr{C}_k) \cap B(\theta, \mathscr{C}_j)$, our result follows easily. □

COROLLARY 4.5. *Suppose $\mathscr{C}_j$ is a nondegenerate critical point. If $g'(c_j) > 0$ then $DF(\mathscr{C}_j)$ has a positive eigenvalue and so $\mathscr{C}_j$ is unstable. Also, $\mathscr{C}_j$ is asymptotically stable if and only if $g'(c_j) < 0$.*

Proof. The first assertion follows from (7). The second assertion follows from the first and from Theorem 4.4. □

THEOREM 4.6. *Let $\mathscr{C}_j$ be any critical point of* (1). *Then $\mathscr{C}_j$ is asymptotically stable or unstable. $\mathscr{C}_j$ is unstable if and only if either $g(c) < 0$ for $c_{j-1} < c < c_j$ or $g(c) > 0$ for $c_j < c < c_{j+1}$.*

Proof. From (5) we have that $\exp(tDF(\mathscr{C}_j))$ is a positive matrix for all $t > 0$. The Perron theorem implies that the principal eigenvector lies in $U \equiv \operatorname{Int} B(\mathscr{C}_{j-1}, \mathscr{C}_j) \cup \operatorname{Int}(\mathscr{C}_j, \mathscr{C}_{j+1})$. Let μ denote the principal eigenvalue of $\exp(DF(\mathscr{C}_j))$. We argue three cases determined by the position of μ relative to the unit circle.

If $|\mu| > 1$ then $\mathscr{C}_j$ is unstable with a 1-dimensional strong unstable manifold contained in U. Thus $g(c) < 0$ for $c_{j-1} < c < c_j$ and $g(c) > 0$ for $c_j < c < c_{j+1}$. If $|\mu| < 1$

then $\mathscr{C}_j$ is asymptotically stable, and so $g(c)>0$ for $c_{j-1}<c<c_j$ and $g(c)<0$ for $c_j<c<c_{j+1}$ by Lemma 4.1.

If $|\mu|=1$ then the Perron theorem gives that $\mu=1$ and that all other eigenvalues of $\exp(DF(\mathscr{C}_j))$ have norms less than 1. Hence $\mathscr{C}_j$ has a 1-dimensional center manifold and a $(n-1)$-dimensional stable manifold. By the equivalency extension theorem of J. Palis and F. Takens [15], the flow in a neighborhood of $\mathscr{C}_j$ is equivalent (i.e., there is a homeomorphism taking orbits onto orbits) to the product flow on the Cartesian product of this center manifold and the stable manifold. Since this center manifold is tangent at $\mathscr{C}_j$ to the principal eigenvector of $\exp(DF(\mathscr{C}_j))$, this center manifold is contained in U and so the flow on it is determined by g near c_j. Thus, if $g(c)<0$ for $c_{j-1}<c<c_j$ or $g(c)>0$ for $c_j<c<c_{j+1}$, the flow on at least one side of this center manifold is leaving $\mathscr{C}_j$ as t increases and hence $\mathscr{C}_j$ is unstable. Otherwise, the flow on this center manifold is positively asymptotic to $\mathscr{C}_j$ and, by the Palis–Takens result, so is the flow in a neighborhood of $\mathscr{C}_j$, i.e., $\mathscr{C}_j$ is asymptotically stable. Thus the theorem is proved. □

If $\mathscr{C}_j$ is an unstable critical point, let $E_j^s\subset\mathbb{R}^n$ denote the invariant vector space determined by all eigenvalues of $\exp(DF(\mathscr{C}_j))$ having norm less than 1. Let M_j denote the stable manifold of $\mathscr{C}_j$ tangent to E_j^s at $\mathscr{C}_j$. Two distinct points in $\mathbb{R}^n$, $x\neq y$, are *related* if $x\leqq y$ or $y\leqq x$. Being related is invariant under the positive-time flow of (1). The next two results are proved in [17].

LEMMA 4.7. *If $\mathscr{C}_j$ is an unstable critical point then the vector space E_j^s contains no related points. In particular, if $v\neq\theta\in\mathbb{R}^n$ and either $\theta\leqq v$ or $v\leqq\theta$, then $v\notin E_j^s$.*

THEOREM 4.8. *Let $\mathscr{C}_j$ be an unstable critical point with an $(n-1)$-dimensional stable manifold M_j. Then M_j contain no related points.*

5. Two-dimensional case. Here we assume $n=2$ in (1). From the Poincaré–Bendixson theorem [7, p. 151], we know an orbit in $\mathscr{H}$ is positively asymptotic to a critical point or a periodic orbit. As in [17], it follows that:

LEMMA 5.1. *$\mathscr{H}$ contains no nonconstant periodic solutions to* (1).

THEOREM 5.2. *The orbit of each point in $\mathscr{H}$ is positively asymptotic to some critical point. If $\mathscr{C}_j$ is an unstable critical point, its stable manifold M_j is 1-dimensional. Each M_j separates $\mathscr{H}$ into regions of orbits positively asymptotic to $\mathscr{C}_{j-1}$, $\mathscr{C}_j$, or $\mathscr{C}_{j+1}$ depending on the sign of g near c_j.*

6. Three-dimensional case. The results for $n=3$ are the same as $n=2$ except the stable manifolds of unstable critical points are 2-dimensional. The argument proceeds by induction on the number of critical points and this induction is similar to that in [17]. The difference occurs in the analysis of a special case with three critical points which is crucial to the induction argument. Here we study this three critical point case. $\mathscr{C}_1$ is assumed to be asymptotically stable. Since $g(c)<0$ for $c_1<c<c_2$, $\mathscr{C}_2$ is unstable. Thus either $g(c)<0$ for $c_2<c<c_3$ or $g(c)>0$ for $c_2<c<c_3$. In the nondegenerate case, the latter occurs and so $\mathscr{C}_3$ is asymptotically stable.

First we show that the stable manifold of an unstable critical point is 2-dimensional.

LEMMA 6.1. *Let $n=3$ in* (1). *The principal eigenvalue of* $\exp(DF(\mathscr{C}))$ *for any degenerate critical point $\mathscr{C}$ is one. Also, if $\mathscr{C}_j$ is an unstable critical point, then its stable manifold M_j is 2-dimensional.*

Proof. Let $\mathscr{C}_j$ be a critical point of (1). Recall that $f_{i,k}\equiv\partial f_i/\partial x_k\geqq 0$. For $1\leqq i\leqq 3$, $h_i>0$ is evaluated at the appropriate component of $\mathscr{C}_j$. An eigenvalue λ of $DF(\mathscr{C}_j)$ is a root of the polynomial

$$\begin{aligned}(8)\qquad &\lambda^3+(h_1+h_2+h_3)\lambda^2+(h_1h_2+h_1h_3+h_2h_3-f_{3,2}f_{2,3})\lambda\\ &+h_1h_2h_3-f_{1,3}f_{2,1}f_{3,2}-h_1f_{3,2}f_{2,3}.\end{aligned}$$

Define the following list of real numbers from the coefficients in (8):

$$a_3 \equiv h_1h_2h_3 - f_{1,3}f_{2,1}f_{3,2} - h_1f_{3,2}f_{2,3},$$

$$T_0 \equiv 1,$$

$$T_1 \equiv h_1 + h_2 + h_3,$$

$$T_2 \equiv (h_1h_2 + h_1h_3 + h_2h_3 - f_{3,2}f_{2,3})T_1 - a_3.$$

According to the Routh–Hurwitz criterion [11, p. 15], the number of roots of (8) with positive real parts is equal to the number of sign changes in the sequence $\{T_0, T_1, T_1T_2, a_3\}$. Since T_0 and T_1 are positive, to have two roots with positive real parts T_1T_2 must be negative and a_3 must be positive. The sign of T_1T_2 is the same as the sign of T_2. But $T_2 < 0$ implies, after a computation, that $h_2h_3 - f_{3,2}f_{2,3} < 0$; and $h_2h_3 - f_{3,2}f_{2,3} < 0$ gives $a_3 < 0$. Hence (8) has at most one root with positive real part. If $\mathscr{C}_j$ is degenerate then $a_3 = 0$. Thus $h_2h_3 - f_{3,2}f_{2,3} > 0$ and so $T_2 > 0$. Hence the principal eigenvalue of a degenerate critical point is one. This completes the proof. □

We return to the special three critical point case. Theorem 4.4 asserts that $B(\mathscr{C}_1, \mathscr{C}_3)$ is an attracting box with the positive orbits of all points in $\mathscr{H}$ entering $B(\mathscr{C}_1, \mathscr{C}_3)$. We divide $B(\mathscr{C}_1, \mathscr{C}_3)$ into eight subboxes by planes through $\mathscr{C}_2$ and parallel to the coordinate planes. $B(\mathscr{C}_1, \mathscr{C}_2)$ and $B(\mathscr{C}_2, \mathscr{C}_3)$ are two such boxes and the other six are defined as follows (see Fig. 1):

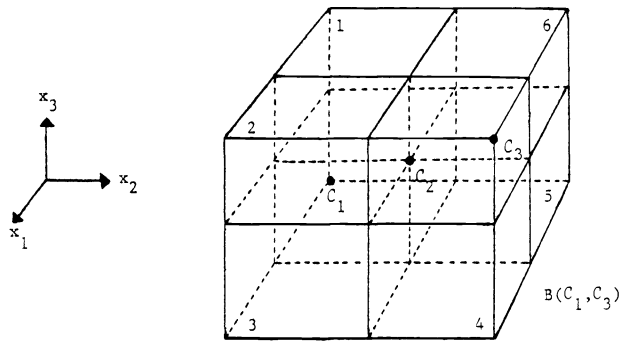


FIG. 1

$$\text{Box } (1) \equiv \{x : p_1(c_1) \leqq x_1 \leqq p_1(c_2),\ p_2(c_1) \leqq x_2 \leqq p_2(c_2),\ c_2 \leqq x_3 \leqq c_3\}$$

$$\text{Box } (2) \equiv \{x : p_1(c_2) \leqq x_1 \leqq p_1(c_3),\ p_2(c_1) \leqq x_2 \leqq p_2(c_2),\ c_2 \leqq x_3 \leqq c_3\}$$

$$\text{Box } (3) \equiv \{x : p_1(c_2) \leqq x_1 \leqq p_1(c_3),\ p_2(c_1) \leqq x_2 \leqq p_2(c_2),\ c_1 \leqq x_3 \leqq c_2\}$$

$$\text{Box } (4) \equiv \{x : p_1(c_2) \leqq x_1 \leqq p_1(c_3),\ p_2(c_2) \leqq x_2 \leqq p_2(c_3),\ c_1 \leqq x_3 \leqq c_2\}$$

$$\text{Box } (5) \equiv \{x : p_1(c_1) \leqq x_1 \leqq p_1(c_2),\ p_2(c_2) \leqq x_2 \leqq p_2(c_3),\ c_1 \leqq x_3 \leqq c_2\}$$

$$\text{Box } (6) \equiv \{x : p_1(c_1) \leqq x_1 \leqq p_1(c_2),\ p_2(c_2) \leqq x_2 \leqq p_2(c_3),\ c_2 \leqq x_3 \leqq c_3\}$$

All orbits in Int $B(\mathscr{C}_1, \mathscr{C}_2)$ are asymptotic to $\mathscr{C}_1$ and all orbits in Int $(\mathscr{C}_2, \mathscr{C}_3)$ are asymptotic to either $\mathscr{C}_2$ or $\mathscr{C}_3$ depending on whether $g(c) < 0$ for $c_2 < c < c_3$ or $g(c) > 0$ for $c_2 < c < c_3$. To study the behavior of orbits in the remaining subboxes we must obtain information about the direction of the vector field F on the faces and interiors of these boxes. If $x_1 = p_1(c_2)$ then $F_1(x) = -h_1(p_1(c_2)) + f_1(p_1(c_2), x_3) =$

$-f_1(p_1(c_2), c_2)+f_1(p_1(c_2), x_3)$, so we get

$$(9) \qquad \begin{cases} F_1(x)>0 & \text{if } x_3>c_2 \quad \text{and} \quad x_1 \leqq p_1(c_2), \\ F_1(x)<0 & \text{if } x_3<c_2 \quad \text{and} \quad x_1 \geqq p_1(c_2). \end{cases}$$

If $x_3=c_2$ then $F_3(x)=-h_3(c_2)+f_3(x_2, c_2)$, so

$$(10) \qquad \begin{cases} F_3(x)>0 & \text{if } x_2>p_2(c_2) \quad \text{and} \quad x_3 \leqq c_2, \\ F_3(x)<0 & \text{if } x_2<p_2(c_2) \quad \text{and} \quad x_3 \geqq c_2. \end{cases}$$

F_2 is slightly more complicated. Divide the portion of the plane $x_2=p_2(c_2)$ in $B(\mathscr{C}_1, \mathscr{C}_3)$ into four open faces (see Fig. 2):

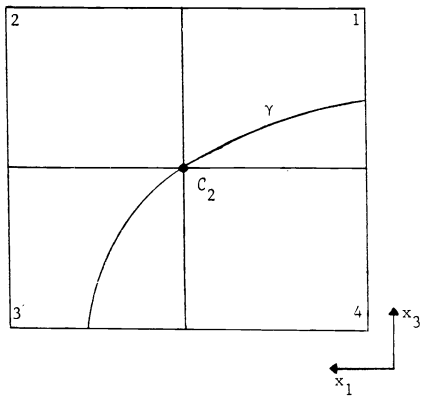


FIG. 2

$$\text{Face (1)} \equiv \{x: p_1(c_1)<x_1<p_1(c_2), x_2=p_2(c_2), c_2<x_3<c_3\}$$

$$\text{Face (2)} \equiv \{x: p_1(c_2)<x_1<p_1(c_3), x_2=p_2(c_2), c_2<x_3<c_3\}$$

$$\text{Face (3)} \equiv \{x: p_1(c_2)<x_1<p_1(c_3), x_2=p_2(c_2), c_1<x_3<c_2\}$$

$$\text{Face (4)} \equiv \{x: p_1(c_1)<x_1<p_1(c_2), x_2=p_2(c_2), c_1<x_3<c_2\}$$

If $x_2=p_2(c_2)$ then $F_2(x)=-h_2(p_2(c_2))+f_2(x_1, p_2(c_2), x_3)$. Since f_2 is increasing in x_1 and nondecreasing in x_3, the set of points where $F_2=0$ is a set γ containing $\mathscr{C}_2$ and is contained in Face (1) $\cup$ Face (3) and in the line $\{x_1=p_1(c_2), x_2=p_2(c_2)\}$. γ separates the plane $x_2=p_2(c_2)$; $F_2(x)>0$ if x is greater than γ and $F_2(x)<0$ if x is less than γ.

Thus, on the faces of $B(\mathscr{C}_1, \mathscr{C}_2)$ or $B(\mathscr{C}_2, \mathscr{C}_3)$, F points into $B(\mathscr{C}_1, \mathscr{C}_2)$ or $B(\mathscr{C}_2, \mathscr{C}_3)$. Hence $B(\mathscr{C}_1, \mathscr{C}_2)\backslash\mathscr{C}_2 \subset \operatorname{dom} \mathscr{C}_1$, and the positive orbit of each point in $B(\mathscr{C}_2, \mathscr{C}_3)$ is asymptotic to $\mathscr{C}_2$ or $\mathscr{C}_3$. As in [17], we prove:

LEMMA 6.2. $A \equiv B(\mathscr{C}_1, \mathscr{C}_2) \cup B(\mathscr{C}_2, \mathscr{C}_3)$ *is an attracting region.*

In the other six subboxes, certain component functions are Lyapunov functions. From (9), the x_1-component function is increasing along orbits in Box (1) $\cup$ Box (6) and decreasing in Box (3) $\cup$ Box (4). From (10), the x_3-component function is increasing in Box (4) $\cup$ Box (5) and decreasing in Box (1) $\cup$ Box (2). This does not prevent an orbit from passing through Face (3) from Box (3) into Box (4) and then returning to Box (3) through Face (3). However, we do have:

LEMMA 6.3. *Neither* Int (Box (3) $\cup$ Box (4)) *nor* Int (Box (1) $\cup$ Box (6)) *contains any compact invariant sets.*

Proof. Suppose Int (Box (3) ∪ Box (4)) contains a compact invariant set I. The set of points $\{x: F_1(x) = 0\}$ intersects Box (3) ∪ Box (4) only in the line $\{x: x_1 = p_1(c_2), x_3 = c_2\}$. Hence I is bounded away from $\{x: F_1(x) = 0\}$, and there is some $\beta > 0$ so that $F_1(x) \leqq -\beta$ for all $x \in I$. If $x(0) \in I$ then $\dot{x}_1(t) = F_1(x(t)) \leqq -\beta$ for all $t \geqq 0$. Thus, for $s > 0$, we have

$$x_1(s) - x_1(0) = \int_0^s \dot{x}_1(t)\, dt \leqq -\beta s.$$

So $x_1(s) \to -\infty$ as $s \to \infty$. This is a contradiction since the orbit of $x(0)$ is bounded. A similar argument holds for Int (Box (1) ∪ Box (6)). □

Lemma 6.3 and the preceding discussion imply that the positive orbit of a point in Box (3) ∪ Box (4) is in dom A or leaves Box (3) ∪ Box (4) in finite time. Thus an orbit starting in Box (2) either enters dom A or enters Box (3) ∪ Box (4) by (10). Then this orbit enters either dom A or Box (5) by (9). This orbit then enters dom A or Box (1) ∪ Box (6) by (10). Finally, this orbit enters either dom A or Box (2) by (9) and Lemma 6.3. Therefore, any orbit not in dom A must spiral around A through the six numbered boxes. We study this spiralling by taking a section for the flow and analyzing the return map. The remaining argument is similar to that in [17] and we just sketch it here.

Choose the face $\mathcal{S}$, shared by Box (1) and Box (2) (see Fig. 3),

$$\mathcal{S} \equiv \{x: x_1 = p_1(c_2),\ p_2(c_1) \leqq x_2 \leqq p_2(c_2),\ c_2 \leqq x_3 \leqq c_3\}.$$

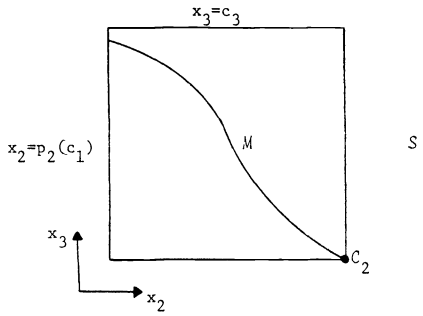


FIG. 3

Let $\mathcal{P}$ denote the plane $x_1 = p_1(c_2)$. $\text{Int}_{\mathcal{P}} \mathcal{S}$ denotes the interior of $\mathcal{S}$ with respect to $\mathcal{P}$. Recall that M_2 is the 2-dimensional stable manifold of $\mathcal{C}_2$. M_2 intersects $\mathcal{P}$ transversely by Lemma 4.7 and Theorem 4.8. Hence $M_2 \cap \mathcal{P}$ is a 1-dimensional C^1 manifold. Let $\mathcal{M}$ denote the connected component of $M_2 \cap \text{Int}_{\mathcal{P}} \mathcal{S}$ such that $\mathcal{C}_2 \subset \text{Cl}\, \mathcal{M}$.

LEMMA 6.4. *Each point of $\mathcal{S} \setminus \mathcal{M}$ related to some point of $\mathcal{M}$ belongs to* dom A. *If the point is above $\mathcal{M}$ then its positive orbit is asymptotic to $\mathcal{C}_2$ or $\mathcal{C}_3$. If the point is below $\mathcal{M}$ then it belongs to* dom $\mathcal{C}_1$.

LEMMA 6.5. *$\mathcal{M}$ is a closed subset of $\text{Int}_{\mathcal{P}} \mathcal{S}$. In fact, M must have a limit point on the edge $x_2 = p_2(c_1)$ or the edge $x_3 = c_3$.*

THEOREM 6.6. *Let $n = 3$ and $k = 3$ in (1). Suppose $\mathcal{C}_1$ is asymptotically stable and M_2 is the stable manifold of $\mathcal{C}_2$. If a point is below M_2 then its positive orbit is asymptotic to $\mathcal{C}_1$. If $\mathcal{C}_3$ is asymptotically stable (i.e., if and only if $g(c) > 0$ for $c_2 < c < c_3$), then the positive orbit of a point above M_2 is asymptotic to $\mathcal{C}_3$. If $\mathcal{C}_3$ is unstable (i.e., if and only if $g(c) < 0$ for $c_2 < c < c_3$), then the positive orbit of a point between M_2 and M_3 is asymptotic to $\mathcal{C}_2$ and the positive orbit of a point above M_3 is asymptotic to $\mathcal{C}_3$.*

7. Conclusion. For $n = 2$ and $n = 3$ we have shown that the stable manifolds of unstable critical points separate the regions of attraction of the critical points. In higher dimensions, unstable critical points may not have codimension one stable manifolds [18] and so cannot separate the space. Even for $n = 4$ there exists the possibility of a degenerate critical point with a 1-dimensional unstable manifold. In general, the dynamics of (1) with $n \geqq 4$ appears complicated.

Acknowledgment. The author would like to thank Robert Martin for a helpful conversation concerning Proposition 3.1.

REFERENCES

[1] D. J. ALLWRIGHT, *A global stability criterion for simple control loops*, J. Math. Biol., 4 (1977), pp. 363–373.
[2] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1965.
[3] J. R. GANTMACHER, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1964.
[4] L. GLASS, *Classification of biological networks by their qualitative dynamics*, J. Theoret. Biol., 54 (1975), pp. 85–107.
[5] L. GLASS AND J. S. PASTERNACK, *Stable oscillations in mathematical models of biological control systems*, J. Math. Biol., 6 (1978), pp. 207–223.
[6] J. S. GRIFFITH, *Mathematics of cellular control processes*, II. *Positive feedback to one gene*, J. Theoret. Biol., 20 (1968), pp. 209–216.
[7] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
[8] S. P. HASTINGS, *On the uniqueness and global asymptotic stability of periodic solutions for a third order system*, Rocky Mountain J. Math., 7 (1977), pp. 513–538.
[9] S. P. HASTINGS, J. J. TYSON AND D. WEBSTER, *Existence of periodic solutions for negative feedback cellular control systems*, J. Differential Equations, 25 (1977), pp. 39–64.
[10] M. W. HIRSCH, *On limit sets in competitive or cooperative dynamical systems*, to appear.
[11] G. A. KORN AND T. M. KORN, *Manual of Mathematics*, McGraw-Hill, New York, 1967.
[12] R. H. MARTIN, *Asymptotic stability and critical points for nonlinear quasimonotone parabolic systems*, J. Differential Equations, 30 (1978), pp. 391–423.
[13] A. J. MEES AND P. E. RAPP, *Periodic metabolic systems: Oscillations in multiple-loop negative feedback biochemical control networks*, J. Math. Biol., 5 (1978), pp. 99–114.
[14] H. G. OTHMER, *The qualitative dynamics of a class of biochemical control circuits*, J. Math. Biol., 3 (1976), pp. 53–78.
[15] J. PALIS AND J. TAKENS, *Topological equivalences of normally hyperbolic dynamical systems*, Topology, 16 (1977), pp. 335–345.
[16] J. F. SELGRADE, *Mathematical analysis of a cellular control process with positive feedback*, SIAM J. Appl. Math., 36 (1979), pp. 219–229.
[17] ———, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38 (1980), pp. 80–103.
[18] J. J. TYSON AND H. G. OTHMER, *The dynamics of feedback control circuits in biochemical pathways*, Progr. Theoret. Biol., 5 (1978), pp. 1–62.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0003 $01.00/0

A FREE BOUNDARY PROBLEM DESCRIBING TRANSITION IN A SUPERCONDUCTOR*

LUIS A. CAFFARELLI†, AVNER FRIEDMAN‡ AND AUGUSTO VISINTIN§

Abstract. Isothermal transition of superconducting material in the presence of an external supercritical magnetic field is considered. In the one-dimensional case a weak formulation is given and then the existence and uniqueness of a solution are derived. Regularity results are obtained for both the solution u (which represents the intensity of the magnetic field) and for the free boundary, which separates the superconductor from the normal conductor.

1. The physical problem. Consider a half space $\{(x, y, z), x > 0\}$ of initially superconductor material in the presence of an external uniform magnetic field of intensity $l \leqq u_c$ (u_c = the critical value, $u_c > 0$). Suppose an external magnetic field of intensity $u_e(t) > u_c$ is applied for $t > 0$ at the boundary $x = 0$ of the superconductor. Assume that both external fields are parallel to the x-axis. Then the intensity of the magnetic field will depend only on x and t; we denote it by $u(x, t)$. Thus

$$u(x, t) \geqq 0 \quad \text{for } x \geqq 0, \quad t \geqq 0 \tag{1.1}$$

and

$$u(x, 0) = u_0(x) \qquad \text{for } x \geqq 0, \tag{1.2}$$

where $u_0(x)$ is the initial intensity. We assume that the initial intensity is stationary, that is (cf. (1.8))

$$u_{0xx} + \alpha u_0 = 0 \qquad \text{for } x > 0, \quad \alpha > 0.$$

It follows that

$$u_0(x) = l e^{-\alpha x} \qquad \text{for } x > 0 \tag{1.3}$$

and $0 < l < u_c$. Also

$$u(0, t) = u_e(t) \qquad \text{for } t > 0, \tag{1.4}$$

where $u_e(t) > u_c$.

Under the external field with intensity u_e, the superconductor is switched gradually into a normal conductor; we assume the transition to be isothermal. The two phases are separated by an unknown curve $x = s(t)$ on which u attains the critical value,

$$u(x(t), t) = u_c \qquad \text{for } t > 0. \tag{1.5}$$

Furthermore, by Maxwell's equations and London's equations [4], on this interface there holds a discontinuity relation for the magnetic flux:

$$u_x(s(t) + 0, t) - \beta u_x(s(t) - 0, t) = -\alpha \int_{s(t)}^{\infty} u(\xi, t)\, d\xi \quad \text{for } t > 0, \tag{1.6}$$

where β is a physical constant, $0 \leqq \beta \leqq 1$; we refer to [1] for the derivation of (1.6) and for further details on the physical phenomenon.

* Received by the editors November 4, 1980. This work was partially supported by the National Science Foundation under grants 7406375 A01 and MCS 791 5171 and by L.A.N. in Pavia.

† Courant Institute of Mathematical Science, New York University, New York, New York 10012.

‡ Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

§ University of Pavia, Corso Strada Nuova 65, 27100 Pavia, Italy.

In the normal conductor Maxwell's equations yield

$$u_t - u_{xx} = 0 \qquad \text{for } 0 < x < s(t), \quad t > 0 \tag{1.7}$$

whereas in the superconductor, by Maxwell's equations and London's equations,

$$\beta u_t - u_{xx} + \alpha u = 0 \qquad \text{for } x > s(t), \quad t > 0. \tag{1.8}$$

The problem (1.1)–(1.8) has been formulated by Cohen and Miranker [1], who have obtained results on the behavior of the free boundary using the asymptotic techniques of singular perturbation theory.

In this paper we begin with a weak formulation of the problem. We show that this problem has a solution u and that we can correspond to it a free boundary $x = s(t)$ such that (1.1)–(1.8) hold ((1.6) is established for almost all t). Further, the solution is unique and the free boundary is monotone increasing and Hölder continuous (exponent $\frac{1}{2}$).

2. Weak formulation. We assume that $\beta \neq 0$. Set

$$\beta(\xi) = \begin{cases} 1 & \text{if } \xi < u_c, \\ \beta & \text{if } \xi > u_c, \end{cases}$$

$$a(\xi) = \int_0^{\xi} \beta(\eta)\, d\eta,$$

$$\chi(x, t) = \begin{cases} 0 & \text{if } x < s(t), \\ 1 & \text{if } x > s(t). \end{cases}$$

Let $T > 0$ and set $Q = \mathbb{R}^+ \times (0, T)$,

$$\Omega = Q \cap \{(x, t);\, x > s(t)\}.$$

LEMMA 2.1. *If (u, s) is a solution of the physical problem* (1.1)–(1.8) *then*

$$\beta u_t - (a(u))_{xx} - \alpha\left(\left(\int_x^{\infty} u(\xi, t)\, d\xi\right)\chi\right)_x = 0 \quad \text{in } \mathscr{D}'(Q). \tag{2.1}$$

Proof. Denote by $\langle \cdot, \cdot \rangle$ the duality between $\mathscr{D}(Q)$ and $\mathscr{D}'(Q)$. Then for any $v \in \mathscr{D}(Q)$,

$$\begin{aligned}
&\left\langle \beta u_t - a(u)_{xx} - \alpha\left(\left(\int_x^{\infty} u\right)\chi\right)_x, v\right\rangle \\
&\quad = \iint_Q \left[-\beta u v_t - a(u) v_{xx} + \alpha\left(\int_x^{\infty} u\right)\chi v_x\right] dx\, dt \\
&\quad = \iint_Q \left[\beta u_t v + a(u)_x v_x\right] dx\, dt + \alpha \iint_{\Omega} \left(\int_x^{\infty} u\right) v_x\, dx\, dt \\
&\quad \qquad \text{(since } u \text{ is continuous)} \qquad (2.2) \\
&\quad = \iint_Q [\beta u_t - a(u)_{xx}] v\, dx\, dt \\
&\qquad - \int_0^T \left[\lim_{x \to s(t)+0} a(u)_x - \lim_{x \to s(t)-0} a(u)_x v\, dt\right] \\
&\qquad - \alpha \int_0^T \left(\int_{s(t)}^{\infty} u\right) v\, dt + \iint_Q uv\, dx\, dt \\
&\quad = \iint_{Q/\Omega} \beta(u_t - u_{xx}) v\, dx\, dt + \iint_{\Omega} (\beta u_t - u_{xx} + \alpha u) v\, dx\, dt
\end{aligned}$$

by (1.6); finally, the last two integrals vanish by (1.7), (1.8).

We dénote by $H(\xi)$ the Heaviside graph.

We shall assume that

$$u_e \in L^2(0, T) \tag{2.3}$$

and introduce a weak formulation of the physical problem.

Problem (P1). Find functions u, χ such that

$$u \in L^2(0, T; H^1(\mathbb{R}^+)), \qquad u \geqq 0 \text{ a.e. in } Q, \tag{2.4}$$

$$\int_0^\infty u(s, t)\, dx < \infty \qquad \text{for a.e. } t \in (0, T), \tag{2.5}$$

$$\chi \text{ is measurable in } Q; \qquad \chi \in H(u_c - u) \text{ a.e. in } Q, \tag{2.6}$$

$$u(0, t) = u_e(t) \qquad \text{a.e. in } t \in (0, T), \tag{2.7}$$

$$\beta \frac{d}{dt} \int_0^\infty uv\, dx + \int_0^\infty \left[a(u)_x v_x + \alpha \left(\int_x^\infty u(\xi, t)\, d\xi \right) \chi v_x \right] dx = 0 \tag{2.8}$$

in $\mathscr{D}'(0, T)$ $\forall v \in H_0^1(\mathbb{R}^+)$,

$$u(0) = u_0. \tag{2.9}$$

Notice that (2.8) implies (2.1). It follows that

$$u_t \in L^2(0, T; H^{-1}(\mathbb{R}^+)); \tag{2.10}$$

hence [3] $u \in C([0, T); H^{-1}(\mathbb{R}^+))$, so that (2.9) has a meaning in $H^{-1}(\mathbb{R}^+)$.

Setting

$$U_0(x) = \int_x^\infty u_0(\xi)\, d\xi, \qquad x > 0,$$

we now introduce another weak formulation of the physical problem.

Problem (P2). Find functions U, χ such that

$$U \in L^2(0, T; H^1(\mathbb{R}^+)), \qquad U_x \leqq 0 \text{ a.e. in } Q, \tag{2.11}$$

$$\chi \text{ is measurable in } Q; \qquad \chi \in H(u_c + U_x) \text{ a.e. in } Q, \tag{2.12}$$

$$\beta \frac{d}{dt} \int_0^\infty Uv\, dx + \int_0^\infty [-a(-U_x)v_x + \alpha U \chi v]\, dx = \int_0^T a(u_e(t))v(0, t)\, dt \tag{2.13}$$

in $\mathscr{D}'(0, T)$ $\forall v \in H^1(\mathbb{R}^+)$,

$$U(0) = U_0. \tag{2.14}$$

Notice that (2.13) yields

$$\beta U_t - (-a(-U_x))_x + \alpha U \chi = 0 \qquad \text{in } \mathscr{D}'(Q) \tag{2.15}$$

from which we get [3]

$$U_t \in L^2(0, T; (H^1(\mathbb{R}^+))'). \tag{2.16}$$

Hence $U \in C([0, T); (H^1(\mathbb{R}^+))')$ so that (2.14) has meaning in $(H^1(\mathbb{R}^+))'$.

Notice also that (2.13) includes the boundary condition

$$U_x(0, t) = -u_e(t) \quad \text{in } (0, T).$$

LEMMA 2.2. *(u, χ) is a solution of* (P1) *if and only if (U, χ) is a solution of* (P2) *where u, U are related by*

$$U(x,t)=\int_x^\infty u(\xi,t)\,d\xi, \tag{2.17}$$

$$u=-U_x. \tag{2.18}$$

The proof is rather immediate.

3. Existence.

THEOREM 3.1. *If*

$$u_e\in W^{1,1}(0,T) \tag{3.1}$$

then there exists at least one solution of (P2) (*and hence, by Lemma* 2.2, *also of* (P1)).

Proof. The proof consists of several steps.

Step 1. *Approximation of* (P2). For any small $\varepsilon>0$, let $H_\varepsilon\in C^\infty(\mathbb{R})$, $a_\varepsilon\in C^\infty(\mathbb{R})$, $u_{e\varepsilon}\in C^1[0,T]$, $u_{0\varepsilon}\in C^2(\mathbb{R}^+)\cap L^1(\mathbb{R}^+)$ be such that

$$H_\varepsilon(\xi)=\begin{cases}0 & \text{if } \xi<0,\\ 1 & \text{if } \xi>\varepsilon,\end{cases}\qquad H_\varepsilon'\geqq 0,$$

$$a_\varepsilon(\xi)=a(\xi)\qquad \text{if } \xi<u_c \text{ or if } \xi>u_c+\varepsilon,$$

$$\tfrac{1}{2}\beta<a'(\xi)<2,$$

$$u_{e\varepsilon}(t)>u_c \text{ if } t>0,\quad u_{0\varepsilon}(x)>0 \text{ if } x>0,\quad u_{e\varepsilon}(0)=u_{0\varepsilon}(0)=u_c,$$

$$a_\varepsilon\to a \quad \text{in } C^\infty(\mathbb{R}\backslash\{u_c\}),$$

$$u_{e\varepsilon}\to u_e \quad \text{in } W^{1,1}(0,T),$$

$$u_{0\varepsilon}\to u_0 \quad \text{in } C^2(\mathbb{R}^+)\cap L^1(\mathbb{R}^+).$$

Set

$$U_{0\varepsilon}(x)=\int_x^\infty u_{0\varepsilon}(\xi)\,d\xi$$

and consider the following problem:

Problem $(\mathrm{P2})_\varepsilon$. Find $U_\varepsilon\in C^{2,1}_{x,t}(Q)$ such that

$$\beta U_{\varepsilon t}-(-a_\varepsilon(-U_{\varepsilon x}))_x+\alpha U_\varepsilon H_\varepsilon(u_C+U_{\varepsilon x})=0 \quad \text{in } Q, \tag{3.2}$$

$$U_{\varepsilon x}(0,t)=-u_{e\varepsilon}(t) \quad \text{for } 0<t<T, \tag{3.3}$$

$$U_\varepsilon(x,0)=U_{0\varepsilon}(x) \quad \text{for } x>0. \tag{3.4}$$

This problem has at least one solution, which can be obtained by standard techniques [2], by approximating Q by bounded domains

$$Q_m=Q\cap\{(x,t);\,|x|<m\}$$

and solving (3.2)–(3.4) in Q_m with, say, $U_\varepsilon=U_{0\varepsilon}$ on $|x|=m$, $0<t<T$.

By the maximum principle, $U_\varepsilon\geqq 0$.

Step 2. *A priori estimates.* We shall denote various positive constants independent of ε by C.

Multiply (3.2) by U_ε and integrate with respect to (x,t). Setting

$$\chi_\varepsilon=H_\varepsilon(u_c+U_{\varepsilon x})$$

and noting that $\alpha U^2 \chi_\varepsilon \geqq 0$, we get

$$\frac{\beta}{2}\int_0^\infty ((U_\varepsilon(t))^2-(U_{0\varepsilon})^2)\,dx+\int_0^t d\tau\int_0^\infty a_\varepsilon(-U_{\varepsilon x})(-U_{\varepsilon x})\,dx \tag{3.5}$$

$$\leqq \int_0^t u_{e\varepsilon}(\tau)U_\varepsilon(0,\tau)\,d\tau$$

$$\leqq C\left\{\int_0^t \|U_\varepsilon\|_{H^1(Q)}\,d\tau\right\}^{1/2}.$$

Hence

$$\|U_\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^+))\cap L^2(0,T;H^1(\mathbb{R}^+))} \leqq C. \tag{3.6}$$

Next we differentiate (3.2) with respect to x. Setting $u_\varepsilon = -U_{\varepsilon x}$ we have

$$\beta u_{\varepsilon t}-(a_\varepsilon(u_\varepsilon))_{xx}-\alpha(U_\varepsilon\chi_\varepsilon)_x=0 \quad \text{in } Q. \tag{3.7}$$

The function u_ε also satisfies the boundary conditions

$$u_\varepsilon(0,t)=u_{e\varepsilon}(t) \qquad \text{for } 0<t<T, \tag{3.8}$$

$$u_\varepsilon(x,0)=u_{o\varepsilon}(x) \qquad \text{for } x>0. \tag{3.9}$$

We now multiply (3.7) by $u_\varepsilon - u_{e\varepsilon}(t)e^{-x}$ and integrate with respect to x, t. We get

$$\beta\int_0^\infty (u_\varepsilon^2(t)-u_{0\varepsilon}^2)\,dx+\int_0^t d\tau\int_0^\infty (a_\varepsilon(u_\varepsilon))_x(u_{\varepsilon x}+u_{e\varepsilon}(t)e^{-x})\,dx$$

$$=\beta\int_0^\infty (u_\varepsilon u_{e\varepsilon})(x,t)\,e^{-x}\,dx-\beta\int_0^t\int_0^\infty u_\varepsilon u_{e\varepsilon}\,e^{-x}\,dx\,dt-\alpha\int_0^t\int_0^\infty U_\varepsilon\chi_\varepsilon u_{\varepsilon x}\,dx.$$

The right-hand side is bounded by

$$C\|u_\varepsilon\|_{L^2(Q)}+C\left(\int_0^\infty u_\varepsilon^2(t)\,dx\right)^{1/2}+C\left\{\int_0^t \|u_{\varepsilon x}\|^2_{L^2(\mathbb{R}^+)}\,d\tau\right\}^{1/2}$$

where (3.6) was used. It now easily follows that

$$\|u_\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^+))\cap L^2(0,T;H^1(\mathbb{R}^+))} \leqq C. \tag{3.10}$$

Recalling (3.2), (3.6) we thus altogether get the estimate

$$\|U_\varepsilon\|_{H^1(0,T;L^2(\mathbb{R}^+))\cap L^2(0,T;H^2(\mathbb{R}^+))} \leqq C. \tag{3.11}$$

Next we want to show that

$$-U_{\varepsilon x}=u_\varepsilon \geqq 0 \quad \text{in } Q. \tag{3.12}$$

For this purpose we rewrite (3.7) in the form

$$\beta u_{\varepsilon t}-(a_\varepsilon(u_\varepsilon))_{xx}+\alpha\chi_\varepsilon u_\varepsilon+\alpha U_\varepsilon H_\varepsilon'(u_c-u_\varepsilon)u_{\varepsilon x}=0 \qquad \text{in } Q. \tag{3.13}$$

This is a parabolic equation for u_ε, and the initial-boundary conditions are given by (3.8), (3.9); notice that these conditions are compatible at (0, 0). Applying the maximum principle we get $u_\varepsilon \geqq 0$, so that (3.12) is valid.

Step 3: *Taking* $\varepsilon \to 0$. From (3.11) and the definition of H_ε it follows that there exists a sequence $\varepsilon_m \downarrow 0$ and functions U, χ such that if $\varepsilon = \varepsilon_m \downarrow 0$,

$$U_\varepsilon \to U \quad \text{in } H^1(0,T;L^2(\mathbb{R}^+))\cap L^2(0,T;H^2(\mathbb{R})) \text{ weakly},$$

$$\chi_\varepsilon \to \chi \quad \text{in } L^\infty(Q) \text{ in the weak star topology},$$

and by compact imbedding and interpolation [3]

$$U_{\varepsilon x} \to U_x \quad \text{in } L^2(Q) \text{ strongly and a.e. in } Q.$$

Since $a(\xi)$ is Lipschitz continuous, we then have

$$a_\varepsilon(U_{\varepsilon x}) \to a(\dot{U}_x) \quad \text{in } L^2(Q) \text{ strongly and a.e. in } Q.$$

Taking $\varepsilon = \varepsilon_m \to 0$ in (3.2), we get (2.13) in the sense of $\mathscr{D}'(Q)$. Next, (3.12) gives $U_x \leqq 0$. Finally,

$$H_\varepsilon(u_c + U_{\varepsilon x}) \to H(u_c + U_x) \quad \text{a.e. in } Q \setminus \{(x, t);\ U_x(x, t) \neq u_c\},$$

and (2.12) follows. We have thus proved that U is a solution of (P2). □

Remark 1. From (3.11) it follows that the solution U of (P2), which we have established, satisfies

$$U \in H^1(0, T: L^2(\mathbb{R}^+)) \cap L^2(0, T; H^2(\mathbb{R}^+)). \tag{3.14}$$

Remark 2. If in Theorem 3.1 we assume that $U_0 \in L^P(0, \infty)$, then $U \in L^\infty(0, T; L^P(0, \infty))$. Indeed, we may take in $(\mathrm{P2})_\varepsilon$ $U_{0\varepsilon} \to U_0$ in $L^P(0, \infty)$. Now multiply (3.2) by $|U_\varepsilon|^{P-2}U_\varepsilon$ and integrate with respect to (x, t). Performing integration by parts, as in the case, $P = 2$, the assertion easily follows.

4. Regularity of the solution. In this section we assume, in addition to (3.1), that

$$u_e'(t) \geqq 0 \qquad \text{if } 0 < t < T. \tag{4.1}$$

Denote by u the solution of (P1) constructed in Theorem 3.1 and denote by Γ the set

$$\Gamma = \{(x, t) \in Q;\ u(x, t) = u_c\}.$$

In this section we prove:

THEOREM 4.1. *Assume that* (3.1), (4.1) *hold. Then* Γ *is given by* $x = s(t)$ *and*

$s(t)$ *is monotone increasing and continuous,*

$u \geqq 0, \quad u_x \leqq 0, \quad (a(u))_{xx} \geqq 0, \quad u_t \geqq 0,$

$u \in C_{x,t}^{\sigma,\sigma/2}$ *for any* $0 < \sigma < 1$, *locally in* Q.

$u_x \in L^\infty_{\mathrm{loc}}(Q), \qquad u_t \in L^2_{\mathrm{loc}}(Q).$

The functions u, s *satisfy* (1.1)–(1.5), (1.7), (1.8); $u_x(s(t) \pm 0, t)$ *exist for all* $t \in (0, T)$ *and* (1.6) *holds for a.e.* t.

Thus the solution of the weak formulation (P2) satisfies essentially all the equations of the physical problem. In the following section we shall establish uniqueness for u.

In order to prove Theorem 4.1 we begin with the solutions u_ε of $(\mathrm{P2})_\varepsilon$. We have already proved that

$$u_\varepsilon \geqq 0 \qquad \text{in } Q.$$

LEMMA 4.2. *Assume that*

$$u_{e\varepsilon}'(t) \geqq 0 \qquad \text{for } 0 < t < T, \tag{4.2}$$

$$u_{0\varepsilon}''(0) = u_{e\varepsilon}'(0) \tag{4.3}$$

$$U_{0\varepsilon}''(x) \geqq \alpha U_{0\varepsilon}(x) \qquad \text{for } x > 0. \tag{4.4}$$

Then

$$U_{\varepsilon t} \geqq 0 \qquad \text{in } Q, \tag{4.5}$$

$$U_{\varepsilon xx} \geqq 0 \qquad \text{in } Q. \tag{4.6}$$

Proof. Differentiating (3.2) with respect to t, we get for $U_{\varepsilon t}$ the equation

$$\beta(U_{\varepsilon t})_t - (a'(-U_{\varepsilon x})U_{\varepsilon tx})_x + \alpha\chi_\varepsilon U_{\varepsilon t} + \alpha U_\varepsilon H'_\varepsilon(u_c + U_{\varepsilon x})U_{\varepsilon tx} = 0 \quad \text{in } Q, \tag{4.7}$$

with boundary conditions

$$\begin{aligned} U_{\varepsilon tx}(0, t) &= -u'_{e\varepsilon}(t) \leqq 0 \qquad \text{for } 0 < t < T, \\ \beta U_{\varepsilon t}(x, 0) &= (-a_\varepsilon(-U'_{0\varepsilon}(x)))' - \alpha U_{0\varepsilon}(x)\chi_\varepsilon(x, 0) \\ &\geqq a'_\varepsilon(-U'_{0\varepsilon}(x))(-U'_{0\varepsilon x}(x)) - \alpha U_{0\varepsilon}(x) \\ &\geqq 0 \qquad \text{for } x > 0, \end{aligned}$$

where $\chi_\varepsilon \leqq 1$ and (4.4) were used in the last two inequalities. The condition (4.3) assures the compatibility of the boundary conditions at (0, 0). We can now apply the maximum principle to deduce (4.5). The inequality (4.6) follows from (3.2) and (4.5).

LEMMA 4.3. *Assume* (4.2)–(4.4) *and*

$$u''_{0\varepsilon}(x) \geqq \alpha u_{0\varepsilon}(x) \qquad \text{for } x > 0. \tag{4.8}$$

Then

$$U_{\varepsilon xt} \leqq 0 \qquad \text{in } Q. \tag{4.9}$$

Proof. Differentiating (4.7) with respect to x we obtain for $U_{\varepsilon xt}$ the equation

$$\begin{aligned} &\beta(U_{\varepsilon xt})_t - (a'_\varepsilon(U_{\varepsilon x})U_{\varepsilon tx})_{xx} + \alpha H'_\varepsilon(u_c + U_{\varepsilon x})U_{\varepsilon x}U_{\varepsilon xt} + \alpha\chi_\varepsilon U_{\varepsilon xt} \\ &\quad + \alpha H''_\varepsilon(u_c + U_{\varepsilon x})U_\varepsilon U_{\varepsilon xt} + \alpha H'_\varepsilon(u_c + U_{\varepsilon x})U_\varepsilon(U_{\varepsilon xt})_x \\ &\quad = -\alpha H'_\varepsilon(u_c + U_{\varepsilon x})U_{\varepsilon t}U_{\varepsilon xx} \leqq 0 \qquad \text{in } Q \end{aligned}$$

where (4.5), (4.6) were used in the last inequality. As for the boundary conditions,

$$U_{\varepsilon xt}(0, t) = -u'_{e\varepsilon}(t) \leqq 0 \qquad \text{for } 0 < t < T$$

and, by differentiating (3.2) with respect to x,

$$\begin{aligned} \beta U_{\varepsilon xt}(x, 0) &= (-a_\varepsilon(-U'_{0\varepsilon}(x)))'' - \alpha U'_{0\varepsilon}(x)\chi_\varepsilon(x, 0) - \alpha U_{0\varepsilon}(x)H'_\varepsilon(u_c + U'_{0\varepsilon}(x))U''_{0\varepsilon}(x) \\ &\leqq (-a_\varepsilon(-U'_{0\varepsilon}(x)))'' - \alpha U'_{0\varepsilon}(x) \leqq 0 \qquad \text{for } x > 0 \end{aligned}$$

where (4.8) was used in the last inequality and $\chi_\varepsilon \leqq 1$, $-U'_{o\varepsilon} \geqq 0$, $U_{0\varepsilon} \geqq 0$, $U''_{0\varepsilon} \geqq 0$ were used in the preceding inequality. The assumption (4.3) gives the compatibility of the boundary conditions at (0, 0). We can now apply the maximum principle and obtain (4.9).

LEMMA 4.4. *There holds*

$$U_t \geqq 0, \quad u = -U_x \geqq 0, \quad u_x = -U_{xx} \leqq 0, \quad u_t = -U_{xt} \geqq 0 \quad \text{a.e. in } Q. \tag{4.10}$$

Proof. It suffices to prove that $u_{0\varepsilon}$, $u_{e\varepsilon}$ can be chosen so that (4.2)–(4.4) and (4.8) hold; for then we take $\varepsilon \to 0$ in (4.5), (4.6) and (4.9). Notice that (4.4) is obtained from (4.8) by integration.

Define $\bar{u}_{0\varepsilon}$ by

$$\bar{u}''_{0\varepsilon} - \alpha \bar{u}_{0\varepsilon} = 0 \qquad \text{if } 0 < x < \delta,$$

$$\bar{u}_{0\varepsilon}(0) = u_c,$$

$$\bar{u}_{0\varepsilon}(x) = u_0(x) \qquad \text{if } x > \delta, \ \delta \text{ small.}$$

Then, since $u''_0 + \alpha u_0 = 0$ for $x > 0$ and $u_0(0) = l < u_c$,

$$\bar{u}''_{0\varepsilon} - \alpha \bar{u}_{0\varepsilon} \geqq 0 \qquad \text{in } \mathscr{D}'(0, \infty)$$

and we can "smooth" $\bar{u}_{0\varepsilon}$ in a neighborhood of $x = \delta$ so that the new function, $u_{0\varepsilon}$, satisfies (4.8); notice that $\bar{u}''_0$ near $x = \delta$ is proportional to a Dirac measure at $x = \delta$.

Since also $u_e(0) > u_c$, we can easily construct approximating functions $u_{e\varepsilon}(t)$ such that

$$u_{e\varepsilon}(t) = u_c + u''_{0\varepsilon}(0)t \qquad \text{if } 0 < t < \bar{\delta}$$

and $u'_{e\varepsilon}(t) \geqq 0$ if $t > \bar{\delta}$; notice that $u'_{e\varepsilon}(t) = u''_{0\varepsilon}(0) > 0$ if $0 < t < \bar{\delta}$. The parameters δ, $\bar{\delta}$ tend to zero as $\varepsilon \to 0$. We have now constructed $u_{0\varepsilon}$, $u_{e\varepsilon}$ satisfying (4.2), (4.3) and (4.8); this completes the proof of the lemma.

LEMMA 4.5. *The set* Γ *(the free boundary) is a curve* $x = s(t)$, *and* $s(t)$ *is monotone increasing.*

Proof. Recall that

$$\Gamma = \{(x, t) \in Q; \ u \equiv -U_x = u_c\}.$$

Since $U_{xx} \leqq 0$, $U_{xt} \geqq 0$, it follows that Γ is a graph with respect to both axes. To prove that Γ is a curve $x = s(t)$ it remains to show that it is impossible to have

$$u(x_1, t_0) = u(x_2, t_0) \quad \text{for } x_1 < x_2, \quad (x_i, t_0) \in Q. \tag{4.11}$$

Suppose (4.11) holds. Then in the rectangle

$$R = (x_1, x_2) \times (0, t_0)$$

we have $\chi = 1$ and therefore

$$\beta u_t - u_{xx} + \alpha u = 0.$$

Moreover u takes a minimum in $\bar{R}$ at (x_1, t_0) and $u_x(x_1, t_0) = 0$; this contradicts the maximum principle.

Since $u(x, t) \geqq u_c$ $(\leqq u_c)$ if $x < s(t)$ $(x > s(t))$, it follows (recalling that $u_e(t) > u_c$, $u_0(x) < u_c$) that

$$s(t) > 0 \qquad \text{if } t > 0. \tag{4.12}$$

Set

$$\Omega = \{(x, t) \in Q; \ x > s(t)\}.$$

In view of Lemma 4.5,

$$\beta u_t - u_{xx} + \alpha u = 0 \qquad \text{in } \Omega, \tag{4.13}$$

$$u_t - u_{xx} = 0 \qquad \text{in } Q \backslash \bar{\Omega}. \tag{4.14}$$

By standard potential theory estimates [2] we deduce that locally in Q

$$\begin{aligned} &u(x, t) \text{ is Hölder continuous in } x \text{ (exponent } \sigma) \\ &\quad \text{and in } t \text{ (exponent } \sigma/2) \text{ for any } 0 < \sigma < 1. \end{aligned} \tag{4.15}$$

From Lemma 4.5 we also obtain $\chi_x \leqq 0$. Since also $u \geqq 0$, $u_t \geqq 0$, the equation (2.1) for u yields

$$(a(u))_{xx} \geqq 0 \qquad \text{in } \mathscr{D}'(Q). \tag{4.16}$$

Thus, in particular,

$$u_x \text{ is increasing in } x, \qquad \text{in } Q\backslash\Gamma. \tag{4.17}$$

Recalling that $u_x \leqq 0$ in Q we have that

$$u_x(s(t)-0, t) \equiv \lim_{x \uparrow s(t)} u_x(x, t) \tag{4.18}$$

exists and is finite for any $0 < t \leqq T$. Similarly

$$u_x(s(t)+0, t) = \lim_{x \uparrow s(t)} u_x(x, t) \tag{4.19}$$

exists for any $0 < t \leqq T$, but we do not know, as yet, that it is finite.

Multiplying (2.1) by $v \in \mathscr{D}(Q)$, $v \geqq 0$ and reversing the calculations in (2.2) (in taking the limit of u_x as $x \to s(t)+0$ we use the monotone convergence theorem) we get

$$\int_0^T \left[u_x(s(t)+0, t) - \beta u_x(s(t)-0, t) + \alpha \int_{s(t)}^{\infty} u(\xi, t)\, d\xi \right] v(s(t), t)\, dt = 0.$$

It follows that

$$u_x(s(t)+0, t) - \beta u_x(s(t)-0, t) = -\alpha \int_{s(t)}^{\infty} u(\xi, t)\, d\xi \qquad \text{a.e. in } t. \tag{4.20}$$

Now, for any $\eta > 0$, if $\eta < t < T$ then

$$|u_x(x, t)| < C \quad \text{if } x = 0 \text{ or } x \geqq x_0, \quad C = C(\eta) \ (x_0 \text{ large enough}).$$

From (4.18), (4.20) and (4.17) we then also have

$$|u_x(x, t)| \leqq C \qquad \text{if } x \in \mathbb{R}^+,$$

for a.e. $t \in (\eta, T)$, with another constant C. Since u_x is continuous in $Q\backslash\Gamma$, it follows that

$$|u_x(x, t)| \leqq C \qquad \text{if } (x, t) \in Q\backslash\Gamma,\ t > \eta; \tag{4.21}$$

thus $u_x \in L^{\infty}_{\text{loc}}(Q)$.

To prove that $u_t \in L^2_{\text{loc}}(Q)$, approximate the free boundary $x = s(t)$ by a curve Γ_δ: $x = s_\delta(t)$ given by $u = u_c + \delta (\delta > 0)$. Denote by R the domain bounded by Γ_δ, $x = x_0$ and $t = t_0$; it lies to the left of Γ.

Multiplying $u_t - u_{xx} = 0$ by u_t and integrating over R, we get

$$\iint_R (u_t)^2 + \frac{1}{2} \iint_R (u_x^2)_t - \int_{\partial R \cap \Gamma_\delta} u_x u_t\, dt + \int_{\partial R \cap \{x = x_0\}} u_x u_t\, dt = 0. \tag{4.22}$$

Notice that

$$\frac{1}{2} \iint_R (u_x^2)_t = \frac{1}{2} \int_{\partial R \cap \{t = t_0\}} (u_x)^2\, dx - \frac{1}{2} \int_{\partial R \cap \Gamma_\delta} (u_x)^2\, dx,$$

$$-\int_{\partial R \cap \Gamma_\delta} u_x u_t\, dt = \int_{\partial R \cap \Gamma_\delta} (u_x)^2\, dx \qquad (\text{since } \dot{s}_\delta = -u_t/u_x)$$

and finally, by (4.21),

$$\int_{\partial R\cap\{x=x_0\}} u_x u_t\,dt = -\int_{\partial R\cap\{x=x_0\}} |u_x| u_t\,dt \geqq -C\int_{\partial R\cap\{x=x_0\}} u_t \geqq -Cu_c.$$

Combining these facts we obtain from (4.22) the inequality

$$\iint_R (u_t)^2 \leqq Cu_c.$$

This completes the proof of Theorem 4.1. □

5. Uniqueness.

THEOREM 5.1. *Let U, V be two continuous solutions of* (P2) *with continuous x-derivative and assume that the sets*

$$\Gamma_U = \{(x,t)\in Q;\ -U_x(x,t) = u_c\},$$
$$\Gamma_V = \{(x,t)\in Q;\ -V_x(x,t) = u_c\}$$

are given by continuous curves

$$x = s_U(t) \text{ and } x = s_V(t) \text{ respectively.}$$

Then $U \equiv V$ in Q.

Proof. Suppose the assertion is not true and assume, for definiteness, that $U - V > 0$ at some points of Q. Then $U - V$ must take its positive maximum in $\bar{Q}$ at some point (x_0, t_0) in Q. The point (x_0, t_0) cannot lie to the left of both Γ_U, Γ_V since in that region U and V satisfy the same parabolic equation. Similarly (x_0, t_0) cannot lie to the right of both free boundaries.

Suppose next that (x_0, t_0) lies between Γ_U and Γ_V or on only one of these curves. Then

$$-U_x \geqq u_c \text{ and } -V_x \leqq u_c \qquad \text{(or conversely)},$$

with at least one inequality being strict. Hence

$$(U-V)_{(x_0,t_0)} \neq 0, \tag{5.1}$$

which is impossible.

It remains to consider the case that (x_0, t_0) belongs to $\Gamma_U \cap \Gamma_V$. Then U and V satisfy the same parabolic equation in the rectangle

$$R_0 : (x_0, x_0+1)\times\left(\frac{t_0}{2}, t_0\right)$$

and the maximum of $U - V$ in $\bar{R}_0$ is attained at (x_0, t_0). But then, the maximum principle gives (5.1), which is impossible. □

6. Further smoothness of the free boundary

LEMMA 6.1. *For any $\eta > 0$ there exists a positive constant c such that*

$$u_x < -c \quad \text{in } (Q\backslash\Gamma)\cap\{t>\eta\}\cap\{x<1/\eta\}. \tag{6.1}$$

Thus u_x is locally strictly negative.

Proof. Set

$$u_x^+(t) = u_x(s(t)+0, t),$$
$$u_x^-(t) = u_x(s(t)-0, t).$$

Thus

$$u_x^+(t) - u_x^- = \alpha \int_{s(t)}^{\infty} u(\xi, t)\, d\xi.$$

In view of $u_x \leqq 0$, $u_{xx} \geqq 0$, the assertion of the lemma would follow from an inequality of the form:

$$u_x^-(t) < -c, \qquad c > 0. \tag{6.2}$$

Denote by u^0 the solution of

$$u_{xx}^0 = \alpha u^0 \qquad \text{if } x > s(t)$$

$$u^0 \to u_c, \qquad u_x^0 \to u_x^+ \quad \text{as } x \downarrow s(t).$$

Since $u_t > 0$ away from Γ,

$$u_{xx} = \alpha u + u_t > \alpha u \qquad \text{if } x > s(t). \tag{6.3}$$

Hence, by comparison, $u < u^0$ if $x > s(t)$. Therefore

$$u_x^+ = u_x^0 = -\int_{s(t)}^{\infty} u_{xx}^0 = -\alpha \int_{s(t)}^{\infty} u^0 < -\int_{s(t)}^{\infty} u,$$

so that

$$u_x^- = u_x^+ + \alpha \int_{s(t)}^{\infty} u\, d\xi < 0,$$

and (6.2) readily follows.

THEOREM 6.2. *The function $s(t)$ is Hölder continuous (exponent $\frac{1}{2}$) in $\eta < t < T$, for any $\eta > 0$.*

Proof. Denote by R the region bounded by

$$x = s(t_0), \quad t = t_0 + h \quad \text{and} \quad \Gamma \qquad (h > 0),$$

and let k be defined by

$$s(t_0 + h) = s(t_0) + k;$$

notice that $k > 0$. The assertion of the theorem is equivalent to the inequality

$$\frac{k^2}{h} \leqq C, \tag{6.4}$$

where C is a constant independent of h, provided

$$\eta < t_0 < t_0 + h \leqq T, \qquad 0 < h < 1.$$

Setting $\tilde{u} = u - u_c$ we can write

$$\iint_R \tilde{u}_t = \left| \int_{\partial R \cap \{t = t_0 + h\}} \tilde{u} \right| = \left| \int_{s(t_0)}^{s(t_0)+k} dx \int_x^{s(t)+k} u_x(\xi, t_0 + h)\, d\xi \right| \tag{6.5}$$

$$\geqq ck^2, \quad \text{by lemma 6.1} \qquad (\text{since } \tilde{u}_x = u_x).$$

The left-hand side of (6.5) is equal to

$$\left| \iint_R \tilde{u}_{xx} \right| = \left| \int_{t_0}^{t_0+h} u_x(s(t) - 0, t)\, dt - \int_{t_0}^{t_0+h} u_x(s(t_0), t)\, dt \right|$$

$$\leqq Ch,$$

since $u_x \in L^\infty_{\text{loc}}(Q)$. It follows that $Ch > ck^2$, and (6.4) is proved. □

REFERENCES

[1] H. COHEN AND W. L. MIRANKER, *Boundary-layer behavior in the superconductor transition problem*, J. Math. Phys., 2 (1961), pp. 575–585.

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

[3] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Problems and Applications*, vol. I, Grundlagen Math. Wiss., 181, Springer, Berlin, 1972.

[4] F. LONDÒN, *Superfluids*, vol. 1, John Wiley, New York, 1950.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0004 $01.00/0

ON THE REDUCTION OF CONNECTION PROBLEMS FOR DIFFERENTIAL EQUATIONS WITH AN IRREGULAR SINGULAR POINT TO ONES WITH ONLY REGULAR SINGULARITIES, I.*

W. BALSER,† W. B. JURKAT‡ AND D. A. LUTZ§

Abstract. It is shown that a complete system of Stokes' multipliers for a system of linear differential equations having an irregular singularity of Poincaré rank one can be calculated in terms of connection relations of certain associated functions near their regular singularities. Solutions of the differential equation are expressed as Laplace integrals of the associated functions, and the formal solutions can be summed as generalized factorial series in explicitly given half-planes of convergence.

Introduction. It is well known (see [5], [11]) that solutions of the *standard differential equation of Poincaré rank one*, i.e.,

$$\frac{dx}{dz} = (\Lambda + A_1 z^{-1})x, \tag{0.1}$$

(where x is n-dimensional ($n \geqq 2$) and $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$ has all distinct entries) can be expressed in terms of convergent Laplace integrals and the solutions are asymptotically equal to formal solutions in sectors which are slightly larger than half-planes (we call these enlarged half-planes). Some integrals which have been used have the form

$$\int Y(t)\, e^{zt}\, dt,$$

where the matrix $Y(t)$ consists of certain solutions of the associated differential equation

$$\frac{dy}{dt} = (\Lambda - tI)^{-1}(I + A_1)y, \tag{0.2}$$

which has singularities only at the regular singular points $\lambda_1, \cdots, \lambda_n, \infty$, and the contours of integration are loops from ∞ along certain rays which encircle the finite singularities.

In the context of a theory of invariants for meromorphic differential equations which has recently been developed by the authors (see [1], [2], [12]), the integrals represent the *normal solutions* (see § 5) of (0.1) (which are uniquely determined by their asymptotic in enlarged half-planes) and the corresponding normalized connection system (i.e., the Stokes' multipliers which relate the normal solutions in consecutive sectors) are invariants of the differential equation. The problem of calculating Stokes' multipliers is usually referred to as a *lateral connection problem* for the differential equation.

In this paper we will show how such lateral connection problems can be reduced to solving connection problems for certain solutions of (0.2) at their regular singularities in the finite complex plane. These solutions, which we call *associated functions*, are expressed locally by convergent expansions which can be calculated explicitly from the

* Received by the editors December 3, 1979, and in revised form November, 7, 1980. Supported in part by grants from the National Science Foundation.

† Universität Ulm, 7900Ulm, West Germany.

‡ Department of Mathematics, Syracuse University, Syracuse, New York 13210.

§ Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

formal solutions of (0.1). The complete analytic structure of these solutions can be determined on their full Riemann surfaces through the differential equation (0.2).

For a "general" differential equation of the form

$$\frac{dx}{dz} = \left(\sum_0^\infty A_\nu z^{-\nu}\right)x, \tag{0.3}$$

where A_0 has all distinct eigenvalues and the power series converges for $|z|$ sufficiently large, associated functions can also be constructed locally by convergent expansions which are formed in an analogous way using the columns of a formal fundamental solution matrix for (0.3). Although there does not appear to be a simple differential equation for these associated functions in the general case, their complete analytic structure on their full Riemann surfaces (including their behavior at ∞) can be determined via a transfer which expresses these associated functions as convolution transforms of associated functions corresponding to standard differential equations. Here, we use a result of Birkhoff–Turrittin which asserts the existence of a meromorphic transformation which takes (0.3) to some standard differential equation (0.1). The normal solutions of (0.3) can be expressed as certain convergent Laplace integrals of these associated functions, the lateral connection problem for (0.3) can be solved in terms of connection problems for the associated functions at their (finite) regular singularities, and the normal solutions may even be summed as convergent generalized factorial series expansions in explicitly given half-planes.

Our methods extend some results of Birkhoff [7] (for a scalar second-order differential equation) and Jurkat, Lutz, Peyerimhoff [13] (for two-dimensional systems) to the n-dimensional case. In these cases the associated differential equation has only two finite singularities, the standard differential equation can be solved explicitly in terms of Kummer functions (see [4]), and the asymptotic of the coefficients in the formal solutions (both for standard and general differential equations) alone is sufficient to produce the Stokes' multipliers. When $n > 2$ one requires more information from the formal solution than just the asymptotic of the coefficients and our procedure involving the associated functions and their connection formulae may be considered as a natural extension of the procedure which was used in the two-dimensional case. In particular, the transfer of behavior of associated functions (from a standard differential equation to a general one) may be thought of as a refinement and extension of a perturbation theorem of Jurkat, Lutz, Peyerimhoff [13, p. 447] of formal series whose coefficients have a given asymptotic behavior.

The differential equation (0.2) has also been considered by K. Okubo [15], who called it the "hypergeometric system." He recognized that the constants which relate the associated functions at their singularities are involved in computing the monodromy group of (0.2) at ∞, but did not consider the relation between (0.2) and (0.1). Okubo [14] has also treated the central connection problem for standard differential equations satisfying certain additional hypotheses. He has reduced the calculation of the central connection coefficients (which then can be used to calculate the Stokes' multipliers in the case of a standard differential equation) to that of calculating the solution of a central connection problem for a related system of linear difference equations, for which convergent representations of solutions are obtained.

Our development proceeds along the following lines. We first investigate the structure of solution matrices $Y(t)$ of the associated differential equation and construct a fundamental solution matrix $Y^*(t)$ which has the property that its kth column is regular everywhere in the finite complex plane except at $t = \lambda_k$. Next we show how to determine the analytic continuation of solutions onto their full Riemann surfaces (§ 2)

and then obtain connection formulas between various solutions (§ 3). The matrix $Y^*(t)$ plays an especially important role in these calculations. In § 4 the transfer of behavior of the associated functions from a standard differential equation to a general one is performed, and in § 5 we use the associated functions to express normal solutions of (0.3) as certain convergent Laplace integrals and factorial series. Here, we also show the equivalence between the lateral connection problem for normal solutions of (0.3) and one for the associated functions. This treatment is based on some notes of the second author.

Three types of contours arise naturally in the Laplace integrals of normal solutions. Loop contours may be used with a different path for each column of $Y(t)$, the loops may be deformed to obtain standard Laplace integrals along various rays, and using the matrix $Y^*(t)$ it is possible to select a common path for all the columns. This last type is especially important in showing the equivalence of the connection problems mentioned above.

We shall use the notation $[A(z)]$ to abbreviate the system of linear differential equations $dx/dz = A(z)x$ and we note here that we always assume that the system is at least two-dimensional, since in the trivial case $n = 1$ the formal solutions converge and yield all the actual solutions explicitly.

1. The associated differential equation and its solutions. Consider the standard differential equation (0.1) where $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$ and the λ_j are all distinct. We assume that

(i) $\operatorname{diag} A_1 \equiv \Lambda' = \operatorname{diag}\{\lambda_1', \cdots, \lambda_n'\}$ *where none of the* λ_j' *is an integer.* Note that condition (i) can always be brought about by making a scalar transformation $x = z^{\gamma} I\tilde{x}$, which transforms (0.1) into

$$\frac{d\tilde{x}}{dz} = [\Lambda + (A_1 - \gamma I)z^{-1}]\tilde{x}$$

and selecting γ to be (mod 1) incongruent to the λ_j'.

With (0.1) we associate the differential equation

$$\frac{dy}{dt} = (\Lambda - tI)^{-1}(I + A_1)y \equiv B(t)y, \tag{1.1}$$

which may be obtained from (0.1) by formally expressing a solution x as a Laplace integral of the form $\int y(t)\, e^{zt}\, dt$ (see also § 5).

The coefficient matrix $B(t)$ has singularities in the finite complex plane at $\lambda_1, \cdots, \lambda_n$ which are first order poles, and has a first order zero at ∞. Hence $\lambda_1, \cdots, \lambda_n$, ∞ are singularities of the first kind for the *associated differential equation* $[B]$; therefore they are regular singular points of its solutions. Expanding $B(t)$ at $t = \lambda_k$ we have

$$B(t) = (t - \lambda_k)^{-1} \sum_{\nu=0}^{\infty} B_{\nu,k}(t - \lambda_k)^{\nu},$$

where $B_{0,k}$ has all entries equal to zero except for the kth row which is equal to the negative of the kth row of $(I + A_1)$. Since $B_{0,k}$ has $n-1$ linearly independent eigenvectors corresponding to the eigenvalue zero, and $-(\lambda_k' + 1)$ (which is not an integer) as the only nonzero eigenvalue, then at $t = \lambda_k$ there exist $n-1$ linearly independent regular solution vectors of $[B]$ and a unique singular solution of the form

$$y_k(t) = (t - \lambda_k)^{-(\lambda_k' + 1)} \sum_{\nu=0}^{\infty} h_k(\nu)(t - \lambda_k)^{\nu}, \tag{1.2}$$

where $h_k(0)/\Gamma(\lambda_k'+1)=e_k$, the kth unit vector (see, e.g., [19, Chapt. II]). By calculating the recursion formulas for the coefficients $h_k(\nu)$ and comparing them with the recursion formulas for the coefficients $f_k(\nu)$ in the formal solution vector

$$\sum_0^\infty f_k(\nu)z^{-\nu}z^{\lambda_k'}e^{\lambda_k z}, \qquad f_k(0)=e_k,$$

of the differential equation $[\Lambda+A_1z^{-1}]$ it can be verified that $h_k(\nu)=f_k(\nu)\Gamma(\lambda_k'+1-\nu)$. This can also be seen from the asymptotic expansion of the Lapace integral $\int e^{zt}y_k(t)\,dt$ (see § 5).

Each of the singular solutions $y_k(t)$ has a branch at $t=\lambda_k$. In order to be able to determine their behavior at the other singularities $\lambda_j (j\neq k)$, we consider a t-plane together with parallel cuts from each λ_k to ∞ along the ray $\arg(t-\lambda_k)=\eta$ for some fixed real number η which may be arbitrarily chosen so that none of these cuts passes through another one of the singularities, i.e.,

$$\arg(\lambda_j-\lambda_k)\not\equiv\eta \bmod 2\pi, \qquad 1\leqq j,k\leqq n, \quad j\neq k.$$

Every such η will be called *admissible*. If we now specify, for $k=1,\cdots,n$,

$$\log(t-\lambda_k)=\log|t-\lambda_k|+i\eta-0$$

for all t such that $\arg(t-\lambda_k)=\eta-0$, then we denote the t-plane with these cuts and choices of the logarithms by $\mathcal{P}_\eta$. The solution $y_k(t)$ (for each k, $1\leqq k\leqq n$) may now be defined near λ_k according to our selection of $\log(t-\lambda_k)$ and then (by means of analytic continuation) defines a single valued function in $\mathcal{P}_\eta$. We will write $y_k(t;\eta)$ if we wish to emphasize the dependence of $y_k(t)$ on the location of the cuts, and we define

$$Y(t)=[y_1(t),\cdots,y_n(t)]=Y(t;\eta).$$

The functions $y_k(t;\eta)$, $1\leqq k\leqq n$, will be called the *associated functions* corresponding to $[A(z)]$ and an admissible direction η.

Considering $y_k(t)$ near $t=\lambda_j$ (for $1\leqq j,k\leqq n$), we see that there exist *unique constants* $c_{jk}=c_{jk}(\eta)$ such that

$$y_k(t)=c_{jk}y_j(t)+\operatorname{reg}(t-\lambda_j) \tag{1.3}$$

(where by reg (t) we generically denote a matrix function of appropriate size which is regular at $t=0$). We form the matrix

$$C=C(\eta)=(c_{jk}), \qquad 1\leqq j,k\leqq n, \tag{1.4}$$

and we note from (1.3) that $c_{jj}=1$, $1\leqq j\leqq n$.

In order to relate the quantities in C to the Stokes' multipliers of the differential equation (0.1) (§ 5), we will as an aid construct another solution matrix $Y^*(t)$ which is intimately related to C. To prove the existence of $Y^*(t)$, we require the following additional assumption:

(ii) *None of the eigenvalues of A_1 is a negative integer.* Note that by selecting γ appropriately in the scalar transformation $x=z^\gamma I\tilde{x}$, we can always arrange that (i) and (ii) hold simultaneously. Moreover, we could even arrange that A_1 satisfied the stronger assumption that

(ii′) *None of the eigenvalues of A_1 are integers.* We shall see in § 2 that assumption (ii′) is equivalent to the invertibility of C.

Remark 1.1. Assumption (ii) holds if and only if $[B]$ has no nontrivial polynomial solution. To see this, realize that $p(t)=\sum_{\nu=0}^d c_\nu t^\nu$ for some integer $d\geqq 0$ and constant

vectors $c_\nu (0 \leqq \nu \leqq d)$, $c_d \neq 0$, is a solution of $[B]$ if and only if

$$
\begin{aligned}
&-(tI-\Lambda)p'(t) = (I+A_1)p(t), \quad \text{i.e.,} \\
&-dc_d = (A_1+I)c_d, \\
&-\nu c_\nu + \Lambda c_{\nu+1}(\nu+1) = (A_1+I)c_\nu, \qquad 0 \leqq \nu \leqq d-1.
\end{aligned} \tag{1.5}
$$

Hence, for a polynomial solution to exist it is necessary that A_1+I have $-d$ as an eigenvalue. Conversely, if $-d$ is taken to be the largest nonpositive integer eigenvalue of A_1+I and c_d is any corresponding eigenvector, then using the invertibility of $(A_1+(\nu+1)I)$, $0 \leqq \nu \leqq d-1$, we recursively calculate $c_{d-1}, \cdots, c_0$ satisfying (1.5) and hence have constructed a polynomial solution.

The nonexistence of a polynomial solution of $[B]$ will be used in the following proposition to construct the solution matrix $Y^*(t)$.

PROPOSITION 1. *Let $[B(t)]$ satisfy assumption* (i) *and let η be admissible. Then there exists a solution matrix $Y^*(t) = Y^*(t;\eta) = [y_1^*(t;\eta), \cdots, y_n^*(t;\eta)]$ for $[B(t)]$ satisfying*

$$y_k^*(t) = \operatorname{reg}(t-\lambda_j) \quad \text{for } t \in \mathscr{P}_\eta, \quad j \neq k, \quad 1 \leqq j,k \leqq n, \tag{1.6}$$

$$y_k^*(t) = y_k(t) + \operatorname{reg}(t-\lambda_k) \quad \text{for } t \in \mathscr{P}_\eta, \quad 1 \leqq k \leqq n, \tag{1.7}$$

if and only if (ii) *holds.*

Moreover, $Y^*(t)$ is uniquely defined by (1.6), (1.7), *is a fundamental solution of $[B]$, and satisfies*

$$Y(t) = Y^*(t)C. \tag{1.8}$$

Proof. Since all vector solutions of $[B]$ form an n-dimensional vector space whereas those staying regular at λ_j form an $(n-1)$-dimensional subspace, we see that for each fixed k, $1 \leqq k \leqq n$, the subspace of solutions staying regular at all λ_j, $j \neq k$, is obtained as an intersection of $n-1$ such subspaces, and by a well-known dimension formula from linear algebra we conclude that this intersection is at least one-dimensional. On the other hand, since every solution of $[B]$ is of the form $cy_k(t) + \operatorname{reg}(t-\lambda_k)$ for a suitable constant c, we conclude that the space of solutions which are regular at all λ_j, $j \neq k$, has dimension greater than one only if $[B]$ has a solution which is regular everywhere and therefore must be a polynomial, due to the fact that ∞ is a regular singularity.

Hence the foregoing discussion shows: If (ii) holds, i.e., if $[B]$ has no polynomial solutions, then for every k, $1 \leqq k \leqq n$, there exists a nontrivial solution $y_k^*(t)$ which is regular at λ_j for all $j \neq k$, $1 \leqq j \leqq n$, and therefore must be singular at λ_k. Such a solution is defined up to a multiplicative constant, which may be selected such that (1.7) holds. This shows that a solution $Y^*(t)$ exists and is unique whenever (ii) holds. Conversely, assume that $Y^*(t)$ satisfying (1.6), (1.7) exists. Let c be a constant vector such that $Y^*(t)c \equiv 0$ for $t \in \mathscr{P}_\eta$. Then for every k, $1 \leqq k \leqq n$, if we let t tend to λ_k, all but one column of $Y^*(t)$ remain regular, hence $Y^*(t)c \equiv 0$ implies $c_k = 0$, $1 \leqq k \leqq n$, which shows that $Y^*(t)$ is fundamental. In the same manner, if c is such that $Y^*(t)c$ stays regular at λ_j for all $j \neq k$ (k fixed), then $c_j = 0$ for all $j \neq k$, hence $Y^*(t)c = y_k^*(t)c_k$, hence the space of solutions being regular at λ_j ($j \neq k$) is one-dimensional, which implies (ii) by means of the foregoing discussion.

Finally, since $Y^*(t)$ (in case of existence) is fundamental, there exists a unique constant matrix, say $\tilde{C}$, such that

$$Y(t) = Y^*(t)\tilde{C},$$

and using (1.7) and (1.3) we see that $\tilde{C} = C$ by discussing the behavior of the kth column near $t = \lambda_j$.

Remark 1.2. Note that C may be constructed for any differential equation (1.1) satisfying assumption (i), and from the proof of Proposition 1 it is seen that under this assumption alone there exist nontrivial solution vectors which are regular at all of the singularities $\lambda_1, \cdots, \lambda_n$ except possibly one of them. So (1.6) can always be satisfied, whereas (1.7) requires assumption (ii).

2. The analytic continuation of solutions of the associated differential equation. Our goal in the next two sections is to determine how the matrices $Y^*(t; \eta)$ vary with respect to t (i.e., determining the analytic continuation when crossing one of the cuts, for example) and η (i.e., comparing $Y^*(t; \eta)$ for different choices of admissible η). The analytic continuation of $Y^*(t)$ will be treated first and it will be applied in the next section to study how the matrices $Y^*(t; \eta)$ are related for various η. As a consequence of our techniques we will characterize the invertibility of the matrix $C(\eta)$ in this section and in the next one we will see how the matrices $C(\eta)$ are related for various choices of η. One could, in principle, make these calculations for $Y(t; \eta)$ directly, but the utilization of $Y^*(t; \eta)$ is a technical convenience which simplifies the ensuing algebraic computations. Whenever we speak of $Y^*(t; \eta)$ we will always implicitly be making assumptions (ii), or (ii'), which guarantees its existence.

LEMMA 1. *Let* (i), (ii) *be satisfied and let* η *be admissible. Then for every fixed* k, $1 \leqq k \leqq n$, *the analytic continuation of* $Y^*(t; \eta)$ *across the cut* $\arg(t - \lambda_k) = \eta$ *in the positive sense is given by*

$$Y^*(t; \eta)(I + C_k^*), \tag{2.1}$$

where C_k^* *has all zero columns except for the* kth *column which is equal to the* kth *column of* $C(\eta)$ *multiplied by* $(e^{-2\pi i \lambda_k'} - 1)$.

Proof. From (1.8) we see (note that $c_{kk} = 1$)

$$y_k(t) = \sum_{j=1}^{n} y_j^*(t) c_{jk}, \quad \text{hence } y_k^*(t) = y_k(t) - \sum_{j \neq k} y_j^*(t) c_{jk},$$

for every fixed k, $1 \leqq k \leqq n$. Since $y_j^*(t)$ stays regular at $t = \lambda_k$ for $j \neq k$, it is sufficient to discuss how $y_k(t)$ changes when crossing the kth cut. From (1.2) we see that $y_k(t)$ goes into $y_k(t)\, e^{-2\pi i \lambda_k'}$ when crossing the cut in the positive direction, hence $y_k^*(t)$ goes into

$$e^{-2\pi i \lambda_k'} y_k^*(t) + (e^{-2\pi i \lambda_k'} - 1) \sum_{j \neq k} y_j^*(t) c_{jk}$$

$$= y_k^*(t) + (e^{-2\pi i \lambda_k'} - 1) \sum_{j=1}^{n} y_j^*(t) c_{jk},$$

and since all the other columns of $Y^*(t)$ remain unchanged, this completes the proof.

Remark 2.1. By exactly the same arguments as in the foregoing proof (or by inverting the matrix $(I + C_k^*)$), one finds that the analytic continuation of $Y^*(t; \eta)$ across the cut $\arg(t - \lambda_k) = \eta$ *in the negative sense* is given by $Y^*(t; \eta)(I + \tilde{C}_k^*)$, where $\tilde{C}_k^*$ has all zero columns except for the kth column which is equal to the kth column of $C(\eta)$ multiplied by $(e^{2\pi i \lambda_k'} - 1)$. This will be used in the next section.

The matrix $Y(t)$ may be thought of as the naturally constructed solution of (1.1) while the matrix $Y^*(t)$ is the uniquely constructed fundamental solution which is particularly convenient with respect to its analytic structure. The proof of Proposition 1 provides a means of constructing $Y^*(t)$, but if C is invertible it may be more natural to

construct $Y^*(t)$ by means of solving (1.8) to obtain

$$Y^*(t) = Y(t)C^{-1}. \tag{2.2}$$

The following proposition shows that C is invertible if and only if (ii′) holds.

PROPOSITION 2. *Let A_1 satisfy* (i) *and let C be defined according to* (1.3). *Then if C is invertible, the matrix*

$$Y^*(t) = Y(t)C^{-1}$$

must satisfy (1.6), (1.7). *Furthermore, C is invertible if and only if* (ii′) *holds.*

Proof. Assume C is invertible, and define $Y^*(t) = Y(t)C^{-1}$. Then it follows from (1.3) that for every j, $1 \leqq j \leqq n$,

$$Y(t) = [y_1(t), \cdots, y_n(t)] = [y_j(t)c_{j1}, \cdots, y_j(t)c_{jn}] + \operatorname{reg}(t-\lambda_j);$$

hence

$$Y^*(t) = Y(t)C^{-1} = [0, \cdots, 0, y_j(t), 0, \cdots, 0] + \operatorname{reg}(t-\lambda_j),$$

which is equivalent to (1.6), (1.7). Therefore we conclude from Proposition 1 that A_1 must satisfy (ii) whenever C is invertible, and since (ii′) always implies (ii) we may now assume that (ii) holds (since otherwise (ii′) fails and C is not invertible).

It follows by solving $[B(t)]$ at $t = \infty$ that (ii′) fails if and only if $[B]$ has a vector solution which is single-valued in a deleted neighborhood of ∞. Since every vector solution of $[B]$ is of the form $Y^*(t)c$ for a constant column vector c, then the analytic continuation of $Y^*(t)c$ around ∞ (in the counter-clockwise sense) may be expresssed as $Y^*(t)C^*c$, where $Y^*(t)C^*$ denotes the analytic continuation of Y^* around ∞. Hence Y^*c is single valued if and only if $C^*c = c$. Hence there exists a single-valued vector solution of $[B]$ if and only if C^* has one as an eigenvalue. If we assume, without loss in generality, that the numeration of the λ_k is such that the ray $\arg(t-\lambda_{k+1}) = \eta$ lies to the right (when going towards ∞) of the ray $\arg(t-\lambda_k) = \eta$ for $1 \leqq k \leqq n-1$, then according to Lemma 1 we find, beginning on the right-hand side of all the cuts,

$$C^* = (I + C_1^*) \cdots (I + C_n^*).$$

Now let $x = (x_1, \cdots, x_n)$ be any constant row vector. Then by induction one proves that, for $k = 1, \cdots, n$,

$$x(I + C_1^*) \cdots (I + C_k^*) = x + \sum_{\nu=1}^{k} \alpha_\nu \delta_\nu,$$

where δ_ν is the νth unit row vector and the constants α_ν are independent of k and are recursively given by

$$\begin{aligned} \alpha_1 &= (e^{-2\lambda_1'\pi i} - 1) \sum_{r=1}^{n} x_r c_{r1}, \\ \alpha_{\nu+1} &= (e^{-2\lambda_{\nu+1}'\pi i} - 1)\left\{ \sum_{r=1}^{n} x_r c_{r,\nu+1} + \sum_{\lambda=1}^{\nu} \alpha_\lambda c_{\lambda,\nu+1} \right\} \qquad \nu = 1, \cdots, n-1. \end{aligned} \tag{2.3}$$

For $k = n$ we see that C^* has one as an eigenvalue if and only if there is a nontrivial row vector x such that

$$x + \sum_{\nu=1}^{n} \alpha_\nu \delta_\nu = x,$$

which holds if and only if $\alpha_1 = \cdots = \alpha_n = 0$. Since none of the λ_k' is an integer, this occurs (in view of (2.3)) if and only if $xC' = 0$; that is, C^* has one as an eigenvalue if and only if C is not invertible.

Remark 2.2. By a similar argument one can show that

$$\det(C^* - I) = \det C \prod_{k=1}^{n} (e^{-2\pi i\lambda_k'} - 1).$$

Remark 2.3. Note that the proofs of Propositions 1 and 2 have the following consequence which will be used later. Suppose that for a fixed admissible η, a matrix $Y(t;\eta) = [y_1(t;\eta), \cdots, y_n(t;\eta)]$ is given where each column $y_k(t;\eta)$ is locally given by a convergent series (1.2). Then if there are constants c_{jk} such that (1.3) holds for $1 \leqq j, k \leqq n$, and if $C = [c_{jk}]$ is invertible, then the matrix $Y^*(t;\eta)$ defined by (2.2) must satisfy (1.6), (1.7), and is even uniquely defined by (1.6), (1.7) within the class of matrices which differ from $Y(t;\eta)$ by right-hand constant factors.

Remark 2.4. It follows from Proposition 1 (under assumption (ii)) that $Y(t)$ is a *fundamental* solution matrix if and only if C is invertible. Furthermore, in the case of invertible C, we see from (1.8) and (2.1) that $Y(t)$, if we perform its analytic continuation across the cut $\arg(t-\lambda_k) = \eta$ in the positive sense, *picks up the constant right-hand factor* $C^{-1}(I + C_k^*) = I + \hat{C}_k$, *and* $\hat{C}_k$ *can be seen to have all zero rows except for the* kth *row which is equal to the* kth *row of* C *multiplied by* $(e^{-2\pi i\lambda_k'} - 1)$. In fact, the same formula for the analytic continuation of $Y(t)$ can be seen to hold (using $y_j(t) = y_k(t)c_{kj} + \operatorname{reg}(t-\lambda_k)$) even when C fails to be invertible.

3. Connections between solutions of the associated differential equation

3.1. The geometry of the cuts and the dominance relation. The set of all admissible numbers η is open, hence a union of countably many open intervals, and two numbers $\eta < \tilde{\eta}$ are in the same interval if and only if the following property holds:

For any k, $1 \leqq k \leqq n$, *if we turn the ray* $\arg(t-\lambda_k) = \eta$ (*in the positive sense*) *by an angle of* $\tilde{\eta} - \eta$, *then none of the points* λ_j $(j \neq k,\ 1 \leqq j \leqq n)$ *is crossed*; *i.e., none of the possible choices for* $\arg(\lambda_j - \lambda_k)$ *lies in the interval* $(\eta, \tilde{\eta})$.

Therefore the *critical values*, i.e., the inadmissible numbers η, are the possible values of $\arg(\lambda_j - \lambda_k)$ for all $j \neq k$, $1 \leqq j, k \leqq n$, and if η is a critical value, then $\eta + 2l\pi$ is also critical for every integer l. We choose the following enumeration for the critical values:

Let m be the number of critical values in the interval $(-\pi/2, 3\pi/2]$, number them as

$$\frac{3\pi}{2} \geqq \eta_0 > \eta_1 > \cdots > \eta_{m-1} > -\frac{\pi}{2}, \tag{3.1}$$

and for every integer k, let

$$\eta_{\nu+km} = \eta_\nu - 2k\pi, \qquad \nu = 0, \cdots, m-1. \tag{3.2}$$

Note that m is always an even number and, since the critical values are periodically distributed mod 2π, there are exactly m such critical values in any interval of the form $(\alpha - 2\pi, \alpha]$. Then the set of η_ν for all integers ν is exactly the set of critical values.

Recalling that the *Stokes' directions* for the differential equation (0.1) (see [2, § 2]) are those rays $\arg z = \tau$ where $\operatorname{Re}(\lambda_j - \lambda_k)z$ changes sign as the ray is crossed, we see that the Stokes' directions τ_ν and the critical values η_ν are related by

$$\eta_\nu + \tau_\nu = \frac{3\pi}{2} \quad \text{for every integer } \nu. \tag{3.3}$$

Also, the *dominance relation* $j \prec k$ on τ (see [2, § 2]), which means that

$$\operatorname{Re}(\lambda_j - \lambda_k)z < 0 \quad \text{for } \arg z = \tau, \tag{3.4}$$

has the following interpretation in $\mathscr{P}_\eta$: If (analogously to (3.3)) we set $\eta + \tau = 3\pi/2$, then (3.4) holds if and only if a suitable choice of $\arg(\lambda_j - \lambda_k)$ lies in the interval $(-\pi + \eta, \eta)$. This is also equivalent to saying that the cut from λ_j to ∞ lies on the right-hand side of the cut from λ_k to ∞ looking in the direction η. *Hence for a fixed ν, if we enumerate the eigenvalues $\lambda_1, \cdots, \lambda_n$ such that the* jth *cut (for any value of $\eta \in (\eta_\nu, \eta_{\nu-1})$) lies on the right-hand side of the* kth *cut whenever $j < k$, then the dominance relation $j \prec k$ in $S'_\nu = S(\tau_{\nu-1}, \tau_\nu)$ coincides with the natural ordering of the indices.*

The set of all pairs (j, k), $j \neq k$, having a *change of dominance* from k to j on τ_ν, i.e., $j \prec k$ on $\tau_\nu - \varepsilon$ while $k \prec j$ on $\tau_\nu + \varepsilon$ ($\varepsilon > 0$ and sufficiently small), is denoted by ρ_ν and is called *the position set corresponding to the Stokes' direction* τ_ν (see [2, § 3]). *Then the position set ρ_ν is exactly the set of pairs (j, k) such that for suitably small $\delta > 0$ the* jth *cut is to the right (resp., left) of the* kth *cut in $\mathscr{P}_{\eta_\nu + \delta}$ (resp., $\mathscr{P}_{\eta_\nu - \delta}$).*

We recall from § 1 that for every admissible η we have made fixed selections of the values $y_k(t; \eta)$, or equivalently, of $\log(t - \lambda_k)$ for $t \in \mathscr{P}_\eta$ and $1 \leqq k \leqq n$. For a second admissible value $\tilde{\eta}$ every point $t_0 \in \mathscr{P}_\eta \cap \mathscr{P}_{\tilde{\eta}}$ such that for every k, $1 \leqq k \leqq n$, the selected values of $\log(t_0 - \lambda_k)$ in $\mathscr{P}_\eta$, resp. $\mathscr{P}_{\tilde{\eta}}$, coincide is called a *reference point with respect to* η and $\tilde{\eta}$, or simply a reference point if η and $\tilde{\eta}$ are fixed. That is, a reference point t_0 is such that no possible choice for $\arg(t_0 - \lambda_k)$ lies between η and $\tilde{\eta}$ for any k, $1 \leqq k \leqq n$. We remark that in case $|\eta - \tilde{\eta}| < 2\pi$ then reference points always exist and the set of all reference points with respect to every pair η, $\tilde{\eta}$ is always simply connected. A point λ_j is called *accessible from reference points with respect to* η *and* $\tilde{\eta}$ for a particular choice of cuts in the directions η and $\tilde{\eta}$ (or just simply *accessible* if the cuts are fixed) if points arbitrarily close to λ_j can be connected to reference points by a path which does not cross any of the cuts. This implies that

$$y_k(t; \eta) = y_k(t; \tilde{\eta})$$

for each k such that λ_k is accessible. We now state that $Y^*(t; \eta)$ essentially stays the same as long as η varies in an interval $(\eta_{\nu+1}, \eta_\nu)$ for some integer ν.

PROPOSITION 3. *For a fixed integer ν, take any two values η, $\tilde{\eta}$ such that $\eta_{\nu+1} < \eta < \tilde{\eta} < \eta_\nu$. Then for every reference point t we have*

$$Y^*(t; \eta) = Y^*(t; \tilde{\eta}). \tag{3.5}$$

Proof. Since each point $\lambda_1, \cdots, \lambda_n$ is accessible from a reference point, then $y_k(t; \eta) = y_k(t; \tilde{\eta})$ for each k, $1 \leqq k \leqq n$, hence $C(\eta) = C(\tilde{\eta})$. Therefore from Proposition 1 we see that $Y^*(t; \eta) = Y^*(t; \tilde{\eta})$ for all reference points with respect to η, $\tilde{\eta}$ since that construction produces a unique Y^* which depends only on Y and C.

Remark 3.1. In view of Proposition 3 we may denote $Y(t; \eta)$ (resp. $Y^*(t; \eta)$) for every $\eta \in (\eta_{\nu+1}, \eta_\nu)$ by $Y_\nu(t)$ (resp. $Y^*_\nu(t)$), and we also denote the common value of $C(\eta)$ for all $\eta \in (\eta_{\nu+1}, \eta_\nu)$ by $C_\nu = (c_{jk}^{(\nu)})$, $1 \leqq j, k \leqq n$. Note that (3.5) requires an additional assumption (ii) or (ii)′ which assures the existence of Y^*, while $Y(t; \eta) = Y(t; \tilde{\eta})$ for all $\eta, \tilde{\eta} \in (\eta_{\nu+1}, \eta_\nu)$ holds generally.

3.2. Relations between various matrices $Y^*_\nu(t)$. For admissible values of η separated by a critical value η_ν, the matrices Y^*_ν and $Y^*_{\nu-1}$ generally are not the same and the following proposition describes the connection between them.

PROPOSITION 4. *For every integer ν and every reference point t relative to the numbers η, $\tilde{\eta}$ satisfying*

$$\eta_{\nu+1} < \eta < \eta_\nu < \tilde{\eta} < \eta_{\nu-1}, \tag{3.6}$$

the constant invertible matrix $W_\nu = (w_{jk}^{(\nu)})$ satisfying

$$Y_{\nu-1}^*(t) = Y_\nu^*(t) W_\nu \tag{3.7}$$

is given by

$$w_{jj}^{(\nu)} = 1, \qquad 1 \leqq j \leqq n, \tag{3.8}$$

$$w_{jk}^{(\nu)} = 0, \qquad (j,k) \notin \rho_\nu, \quad 1 \leqq j \neq k \leqq n, \tag{3.9}$$

$$w_{jk}^{(\nu)} = (1 - e^{-2\pi i \lambda_k'}) c_{jk}^{(\nu)} \quad \text{if } (j,k) \in \rho_\nu, \quad 1 \leqq j \neq k \leqq n. \tag{3.10}$$

Proof. By $y_{\nu-1,k}^*(t)$, resp., $y_{\nu,k}^*(t)$ we denote the kth column of $Y_{\nu-1}^*(t)$, resp., $Y_\nu^*(t)$. For some fixed k, $1 \leqq k \leqq n$, let η, $\tilde{\eta}$ be chosen according to (3.6) and note that since $0 < \tilde{\eta} - \eta < 2\pi$ there always exist reference points for which (3.7) holds. For the kth column we have

$$y_{\nu-1,k}^*(t) = y_{\nu,1}^*(t) w_{1k}^{(\nu)} + \cdots + y_{\nu,n}^*(t) w_{nk}^{(\nu)}, \tag{3.11}$$

where we consider the plane $\mathcal{P}_\eta$ together with an additional cut along $\arg(t - \lambda_k) = \tilde{\eta}$. For each point λ_j, $1 \leqq j \leqq n$, which is accessible from reference points relative to these cuts, (3.11) remains valid by means of analytic continuation along an appropriate path and if $j \neq k$ then since the only function in (3.11) which becomes singular at λ_j is $y_{\nu,j}^*(t)$, we conclude that $w_{jk}^{(\nu)} = 0$. Similarly, if $j = k$ (note that λ_k is always accessible) we conclude that $y_{\nu-1,k}^*(t) - y_{\nu,k}^*(t) w_{kk}^{(\nu)} = \operatorname{reg}(t - \lambda_k)$, which in view of (1.3), (1.7) holds if and only if $w_{kk}^{(\nu)} = 1$. (Here we use that at a reference point t we have $y_k(t;\eta) = y_k(t;\tilde{\eta})$.) This leaves us with a discussion of what happens at an inaccessible point λ_j, i.e., a point λ_j for which a possible choice of $\arg(\lambda_j - \lambda_k)$ lies in $(\eta, \tilde{\eta})$. But from our discussion of the dominance relation, this happens if and only if $(j,k) \in \rho_\nu$. Hence the *support of* W_ν, i.e., the offdiagonal positions where W_ν has nonzero entries, is contained in the position set ρ_ν. This establishes (3.8), (3.9). The calculation (3.10) follows as an application of Proposition 5 (see Remark 3.5). It could also be proven here directly by the same method of analytic continuation across the cuts, but we delay this argument until later since the geometry of the cuts is easier to work with in the situation of Proposition 5.

Remark 3.2. Using similar arguments with $Y_\nu^*(t) = Y_{\nu-1}^*(t) W_\nu^{-1}$, it can be shown that the elements of W_ν^{-1} can be expressed in terms of $C_{\nu-1}$ as follows. If we let $W_\nu^{-1} = (w_{jk})$, then $w_{jj} = 1$, $1 \leqq j \leqq n$, $w_{jk} = 0$ for $(j,k) \notin \rho_\nu$, $j \neq k$, and

$$w_{jk} = (1 - e^{2\pi i \lambda_k'}) c_{jk}^{(\nu-1)} \quad \text{for } (j,k) \in \rho_\nu.$$

This calculation also follows from Remark 3.5, as we shall see.

Remark 3.3. If η is any admissible number, then $\mathcal{P}_\eta$ and $\mathcal{P}_{\eta-2\pi}$ consist of the same set of complex numbers and differ only in the different selection of the branches of $\log(t - \lambda_k)$, $1 \leqq k \leqq n$. Hence the matrices $Y(t;\eta)$ and $Y(t;\eta - 2\pi)$ are defined in the same domain of complex numbers t and are obviously related by

$$Y(t;\eta - 2\pi) = Y(t;\eta)\, e^{2\pi i \Lambda'}.$$

Recall that the number m of critical values in any interval of the form $(\alpha - 2\pi, \alpha]$ is even, and define $\mu = m/2$. Then for any $\eta \in (\eta_{\nu+1}, \eta_\nu)$ we see that

$$C_{\nu+m} = e^{-2\pi i \Lambda'} C_\nu\, e^{2\pi i \Lambda'} \quad \text{for every integer } \nu.$$

Moreover, since $Y_\nu^*(t)\, e^{2\pi i\Lambda'}$ and $Y_{\nu+m}^*(t)$ can both be considered as fundamental solution matrices in $\mathscr{P}_{\eta-2\pi}$, the above relation for $Y(t;\eta)$ and $Y(t;\eta-2\pi)$ can be used to check that $Y_\nu^*(t)\, e^{2\pi i\Lambda'}$ has (as a solution in $\mathscr{P}_{\eta-2\pi}$) the properties (1.6), (1.7) which uniquely characterize $Y_{\nu+m}^*(t)$ (according to Lemma 1), hence

$$Y_{\nu+m}^*(t) = Y_\nu^*(t)\, e^{2\pi i\Lambda'}. \tag{3.12}$$

From this it follows that

$$W_{\nu+m} = e^{-2\pi i\Lambda'} W_\nu\, e^{2\pi i\Lambda'} \tag{3.13}$$

for every integer ν, hence *any collection of m consecutive W_ν determines all of them.*

In § 5 we shall see that the connection matrices W_ν equal the normalized connection matrices V_ν (see [2]) of the differential equation $[A(z)]$. In order to relate the elements of some matrix C to the invariants of $[A(z)]$, we will now establish a one-to-one correspondence between a single C_ν and the set $W_{\nu+1}, \cdots, W_{\nu+m}$.

If we define, for an arbitrary but fixed integer ν,

$$C_\nu^+ = (W_{\nu+\mu} \cdots W_{\nu+1})^{-1} \tag{3.14}$$

and

$$C_\nu^- = W_{\nu+m} \cdots W_{\nu+\mu+1}, \tag{3.15}$$

then clearly $W_{\nu+1}, \cdots, W_{\nu+m}$ uniquely determine C_ν^+, C_ν^-. Moreover, C_ν^+, C_ν^- uniquely determine the matrices $W_{\nu+1}, \cdots, W_{\nu+m}$ as we will show. Note that a pair (j, k) changes dominance exactly once within any interval of the form $(\alpha-\pi, \alpha]$ and $(j, k) \in \rho_\nu$ if and only if $(k, j) \in \rho_{\nu+\mu}$. Hence the sets

$$\sigma'_{\nu+1} = \rho_{\nu+1} \cup \cdots \cup \rho_{\nu+\mu}$$

and

$$\sigma'_{\nu+\mu+1} = \rho_{\nu+\mu+1} \cup \cdots \cup \rho_{\nu+m},$$

where $\sigma'_{\nu+1}$, resp. $\sigma'_{\nu+\mu+1}$, denotes the set of pairs (j, k) with $j < k$ in $S'_{\nu+1} = S(\tau_\nu, \tau_{\nu+1})$, resp. $S'_{\nu+\mu+1}$ (compare [2, § 3]), are antisymmetric and transitive sets. Also, the sets $\rho_{\nu+1}, \cdots, \rho_{\nu+m}$ are disjoint, antisymmetric and transitive; hence according to Proposition 3 [2, § 4] the matrices $W_{\nu+1}, \cdots, W_{\nu+\mu}$, resp. $W_{\nu+\mu+1}, \cdots, W_{\nu+m}$ can be uniquely calculated using C_ν^+, resp., C_ν^-. Note that if we order the eigenvalues $\lambda_1, \cdots, \lambda_n$ according to the dominance relation in $S'_{\nu+1}$, then C_ν^+ becomes upper triangular, whereas C_ν^- is lower triangular. In what follows, for simplicity of calculations, we will assume that such an ordering for the eigenvalues has been made.

In order to relate C_ν^+ to C_ν, using (3.14) and the definition (3.7) we obtain

$$Y_{\nu+\mu}^*(t) = Y_\nu^*(t) C_\nu^+ \tag{3.16}$$

for t a reference point with respect to $\eta-\pi$ and η, where η is arbitrary, but fixed, in $(\eta_{\nu+1}, \eta_\nu)$. These reference points include all points lying on the left hand side of all the cuts in the direction η (looking towards ∞). Denoting the elements of C_ν^+ by c_{jk}^+, $1 \leqq j, k \leqq n$, then from (3.16) we obtain

$$y_{\nu+\mu,k}^*(t) = y_{\nu,1}^*(t) c_{1k}^+ + \cdots + y_{\nu,n}^*(t) c_{nk}^+. \tag{3.17}$$

For (3.17) it is sufficient to consider the plane with cuts along $\arg(t-\lambda_j) = \eta$, $1 \leqq j \leqq n$, and an additional cut along $\arg(t-\lambda_k) = \eta-\pi$. Just as in the proof of Proposition 4, if we continue all the functions in (3.17) analytically along a path towards any λ_j $(1 \leqq j \leqq n)$ which is accessible relative to this system of cuts, then $c_{jk}^+ = 0$ if $j \neq k$ and

$c_{kk}^{+}=1$ (note here that λ_k is always accessible). Furthermore, we see that the accessible points λ_j are precisely those for which $j \geqq k$ (according to our ordering of the eigenvalues). In terms of the dominance relation, they are (aside for $j=k$) those j such that $k<j$ in $S'_{\nu+1}$. Now consider an inaccessible point λ_j with $j<k$ (i.e., $j<k$ in $S'_{\nu+1}$). We then perform an analytic continuation of the functions in (3.17) along a path from a reference point which crosses none of the cuts except $\arg(t-\lambda_k)=\eta$, and since the reference points are to the left of this cut, this cut is crossed in the negative sense. The only function in (3.17) which changes when we cross this cut is $y_{\nu,k}^{*}(t)$ and (using Remark 2.1) its analytic continuation is given by

$$y_{\nu,k}^{*}(t)+(e^{2\pi i\lambda'_k}-1)\sum_{l=1}^{n} y_{\nu,l}^{*}(t)c_{lk}^{(\nu)}.$$

Hence the continuation of (3.17) yields

$$y_{\nu+\mu,k}^{*}(t)=\sum_{l=1}^{n} y_{\nu,l}^{*}(t)[c_{lk}^{+}+(e^{2\pi i\lambda'_k}-1)c_{lk}^{(\nu)}]. \tag{3.18}$$

Since each point λ_j, $j<k$, is now accessible, and since the only function in (3.18) which becomes singular at λ_j is $y_{\nu,j}^{*}(t)$, we see that

$$c_{jk}^{+}=(1-e^{2\pi i\lambda'_k})c_{jk}^{(\nu)}, \qquad 1 \leqq j<k \leqq n. \tag{3.19}$$

Likewise, using (3.15) and (3.12) we obtain

$$Y_{\nu+\mu}^{*}(t)=Y_{\nu+m}^{*}(t)C_{\nu}^{-}=Y_{\nu}^{*}(t)\,e^{2\pi i\Lambda'}C_{\nu}^{-}, \tag{3.20}$$

for t a reference point with respect to $\eta-\pi$ and $\eta-2\pi$ (hence lying on the right hand side of the cuts in the direction $\eta-2\pi$). Applying similar arguments as above, we find that $C_{\nu}^{-}=(c_{jk}^{-})$, $1 \leqq j, k \leqq n$, is lower triangular with ones on the diagonal and

$$c_{jk}^{-}=(1-e^{-2\pi i\lambda'_k})c_{jk}^{(\nu+m)}, \qquad 1 \leqq k<j \leqq n.$$

Hence using $C_{\nu+m}=e^{-2\pi i\Lambda'}C_{\nu}\,e^{2\pi i\Lambda'}$, we may calculate c_{jk}^{-} in terms of C_{ν} and we formulate these results as

PROPOSITION 5. *For every integer ν the connection matrices between $Y_{\nu}^{*}(t)$ and $Y_{\nu+\mu}^{*}(t)$ are given by* (3.16) *and* (3.20) (*depending upon in which half-planes they are compared*) *and can be calculated from C_{ν} as follows*:

$$C_{\nu}^{+}=(W_{\nu+\mu}\cdots W_{\nu+1})^{-1}=(c_{jk}^{+}),$$

$$c_{jk}^{+}=\begin{cases}0 & \text{if } k<j \text{ in } S'_{\nu+1},\\ 1 & \text{if } k=j,\ 1 \leqq j \leqq n,\\ (1-e^{2\pi i\lambda'_k})c_{jk}^{(\nu)} & \text{if } j<k \text{ in } S'_{\nu+1};\end{cases} \tag{3.21}$$

$$C_{\nu}^{-}=W_{\nu+m}\cdots W_{\nu+\mu+1}=(c_{jk}^{-}),$$

$$c_{jk}^{-}=\begin{cases}0 & \text{if } j<k \text{ in } S'_{\nu+1},\\ 1 & \text{if } k=j,\ 1 \leqq j \leqq n,\\ e^{2\pi i(\lambda'_k-\lambda'_j)}(1-e^{-2\pi i\lambda'_k})c_{jk}^{(\nu)}, & \text{if } k<j \text{ in } S'_{\nu+1}.\end{cases} \tag{3.22}$$

Furthermore, since the factors in (3.21), (3.22) *are uniquely determined by their product, this establishes a one-to-one correspondence between the elements of the matrix C_{ν} and the connection system $(W_{\nu+1},\cdots,W_{\nu+m})$; moreover, in light of* (3.13), *there is a one-to-one correspondence between a single matrix C_{ν} and the connection system $(W_{\mu})_{\mu=-\infty}^{+\infty}$.*

3.3. Relations between various matrices C_ν.

Remark 3.4. As a consequence of Proposition 5 we see that (from the one-to-one correspondence between an arbitrary C_ν and the system (W_μ)) there is even a one-to-one correspondence between $C_{\nu-1}$ and C_ν. Moreover, the formulas derived in Remark 3.2 enable us to find the explicit formulas expressing $C_{\nu-1}$ in terms of C_ν. Namely, using (3.21), (3.22) and (3.13) for both ν and $\nu-1$, one finds

$$C_\nu^+ W_{\nu+\mu} = W_\nu C_{\nu-1}^+, \tag{3.23}$$

$$e^{2\pi i\Lambda'} C_\nu^- W_{\nu+\mu} = W_\nu e^{2\pi i\Lambda'} C_{\nu-1}^-. \tag{3.24}$$

Furthermore, from (3.21) and (3.22) we also see (with $D = \operatorname{diag}\{1-e^{2\pi i\lambda_1'}, \cdots, 1-e^{2\pi i\lambda_n'}\}$) that

$$C_\nu D = C_\nu^+ - e^{2\pi i\Lambda'} C_\nu^- \quad \text{for every integer } \nu, \tag{3.25}$$

and therefore we see, using (3.23), (3.24), and (3.25) (both with ν and $\nu-1$) that

$$W_\nu C_{\nu-1} D = C_\nu D W_{\nu+\mu} \quad \text{for every integer } \nu. \tag{3.26}$$

Remark 3.5. Using (3.23) and (3.24) we will now see how these equations contain, in particular, the calculations of W_ν in terms of C_ν (3.10), W_ν^{-1} in terms of $C_{\nu-1}$ (Remark 3.2), as well as $W_{\nu+\mu}$ in terms of $C_{\nu-1}$ and $W_{\nu+\mu}^{-1}$ in terms of C_ν. If we label the eigenvalues $\lambda_1, \cdots, \lambda_n$ according to the dominance relation in $S(\tau_\nu, \tau_{\nu+1})$, then as remarked earlier in this section, the cuts in $\mathscr{P}_\eta$ for $\eta \in (\eta_{\nu+1}, \eta_\nu)$ occur from right to left (from the points $\lambda_1, \cdots, \lambda_n$ looking toward ∞ in the direction η). Consider the points $\lambda_1, \cdots, \lambda_n$ in the complex plane together with a line in the direction η_ν through each point. Then (according to the definition of η_ν) at least two points $\lambda_1, \cdots, \lambda_n$ must lie on the same line and we consider the decomposition of the set $\{\lambda_1, \cdots, \lambda_n\}$ into subsets according to whether or not they lie on the same such line. Since there is no critical value in $(\eta_{\nu+1}, \eta_\nu)$ we see (by first considering $\eta < \eta_\nu$ and taken very close to η_ν) that the subset of points which lie on the same line have consecutively numbered indices. Moreover, the position set ρ_ν consists of all pairs (j, k) where $j < k$ and λ_j, λ_k lie on the same line in the direction η_ν. This establishes a blocking of the set of indices $\{1, 2, \cdots, n\}$ in which two indices j, $j+1$ occur in the same block if and only if $\arg(\lambda_{j+1} - \lambda_j) \equiv \eta_\nu \pmod{2\pi}$ (hence one-dimensional blocks correspond to eigenvalues with the property that no other eigenvalue lies on the line determined by it and the direction η_ν). If we compare this labeling and blocking with the corresponding one for $\eta \in (\eta_\nu, \eta_{\nu-1})$, then we see that the single elements keep their same index and within every nontrivial block, the ordering (with respect to the dominance relation) of the indices is exactly the reverse of what it was for $\eta \in (\eta_{\nu+1}, \eta_\nu)$. This explains how the dominance relation changes for consecutive sectors S_ν'.

Let the matrices C_ν^+, $C_{\nu-1}^+$ be partitioned according to this blocking; i.e., the diagonal blocks correspond to the indices occurring in the same block. Recall from Proposition 4 that if W_ν is blocked in the same manner, then since the support of W_ν is in ρ_ν, W_ν is actually diagonally blocked and the diagonal blocks themselves are lower triangular. Likewise, since the support of $W_{\nu+\mu}$ is contained in $\rho_{\nu+\mu}$, the set of opposite pairs of those in ρ_ν, then $W_{\nu+m}$ is also diagonally blocked according to this same blocking and the diagonal blocks are themselves upper triangular. From Proposition 5 we see that if the eigenvalues are ordered according to the dominance relation in $S_{\nu+1}'$ the C_ν^+ is upper triangularly blocked and the diagonal blocks are also upper triangular, while $C_{\nu-1}^*$ is upper triangularly blocked, but the diagonal blocks are lower triangular (comparing the dominance relation in $S_{\nu+1}'$ with that in S_ν' as described above). Thus

comparing the diagonal blocks of each side of (3.23) we see that those on the left-hand side are upper triangular with diagonal equal to I, while those on the right-hand side are lower triangular with the same diagonal I. Hence the factors of the diagonal blocks on both sides are inverses of each other. This implies that the matrix $W_{\nu+\mu}^{-1}$ is calculated in terms of C_ν as follows:

$$W_{\nu+\mu}^{-1} = (w_{jk}), \qquad 1 \leqq j, k \leqq n,$$

$$w_{jj} = 1, \qquad 1 \leqq j \leqq n,$$

$$w_{jk} = 0, \qquad (j, k) \notin \rho_{\nu+\mu},$$

$$w_{jk} = (1 - e^{2\pi i \lambda_k'}) c_{jk}^{(\nu)}, \qquad (j, k) \in \rho_{\nu+\mu}.$$

The calculation of W_ν in terms of the elements of $C_{\nu-1}$ obtained in Remark 3.2 corresponds to a similar argument for the diagonal blocks on the right-hand side of (3.23). Similarly, (3.24) is equivalent to

$$e^{-2\pi i \Lambda'} W_\nu^{-1} e^{2\pi i \Lambda'} C_\nu^- = C_{\nu-1}^- W_{\nu+\mu}^{-1},$$

and arguing in the same manner, one obtains (3.10) from the diagonal blocks on the left-hand side, while from the diagonal blocks on the right-hand side one obtains a formula for the calculation of $W_{\nu+\mu}$ in terms of the elements of $C_{\nu-1}$.

Remark 3.6. These relations also may be used to calculate all the matrices W_ν from a single C_ν by the following inductive procedure. Calculate W_ν and $W_{\nu+\mu}$ from the diagonal blocks of C_ν (with the blocking of indices introduced above associated with η_ν). From (3.26) one can then calculate $C_{\nu-1}$ and blocking it according to blocking associated with $\eta_{\nu-1}$, the diagonal blocks then yield $W_{\nu-1}$ and $W_{\nu+\mu-1}$, etc.

Remark 3.7. As a consequence of the definitions (3.14), (3.15) and the relation (3.13), we see that

$$C_{\nu+\mu}^+ = (C_\nu^-)^{-1}$$

and

$$C_{\nu+\mu}^- = e^{-2\pi i \Lambda'} (C_\nu^+)^{-1} e^{2\pi i \Lambda'}$$

and making use of (3.25) (for $\nu+\mu$) we obtain

$$C_{\nu+\mu} = ((C_\nu^-)^{-1} - (C_\nu^+)^{-1} e^{2\pi i \Lambda'}) D^{-1}. \tag{3.27}$$

Remark 3.8. Using $Y_\nu(t) = Y_\nu^*(t) C_\nu$ (also for $\nu - 1$) and the definition (3.7), one sees (assuming that C_ν is invertible) that

$$Y_{\nu-1}(t) = Y_\nu(t) C_\nu^{-1} W_\nu C_{\nu-1},$$

and using (3.26), we obtain

$$Y_{\nu-1}(t) = Y_\nu(t) D W_{\nu+\mu} D^{-1} \tag{3.28}$$

for reference points t with respect to η, $\tilde{\eta}$ satisfying (3.6). Similarly, using $Y_\nu(t) = Y_\nu^*(t) C_\nu$ (also for $\nu+\mu$), (3.16), (3.27) and (3.25), we obtain

$$Y_{\nu+\mu}(t) = Y_\nu(t) D (C_\nu^-)^{-1} D^{-1} \tag{3.29}$$

for reference points t relative to $\eta - \pi$, η; also using (3.20) we obtain

$$Y_{\nu+\mu}(t) = Y_\nu(t) [D (C_\nu^+)^{-1} e^{2\pi i \Lambda'} D^{-1}] \tag{3.30}$$

for reference points relative to $\eta - \pi$, $\eta - 2\pi$. Since $W_{\nu+\mu}$ can be given (see Remark 3.5) in terms of $C_{\nu-1}$, (3.28) describes the connection formula between consecutive matrices $Y_\nu(t)$ in terms of the elements of $C_{\nu-1}$. The connection formulas for the matrices $Y_\nu(t)$ (analogous to (3.16) and (3.20) for $Y_\nu^*(t)$) are given by (3.29), (3.30). Note that although the formulas (3.28), (3.29), (3.30) were derived under the assumption that C_ν is invertible, the formulas themselves are given by quantities which are always defined (just using assumption (i)). Moreover, all the matrices W_ν may be considered as being defined by (3.21), (3.22) in terms of $C_\nu^\pm$, even though $Y_\nu^*(t)$ may fail to exist. Since the invertibility of all of the C_ν can be brought about by a scalar shift z^γ of $[A(z)]$, and in § 4 we will show that the matrices $Y_\nu(t)$ as well as the quantities in C_ν depend analytically upon the parameter γ when it is varied in a deleted neighborhood of 0, then by analytic extension the connection formulas (3.28), (3.29) and (3.30) can be shown to be valid without the additional assumption that C_ν is invertible.

4. The associated functions corresponding to a general differential equation

4.1. Definition of the associated functions. We will now define associated functions for more general differential equations $[\tilde{A}(z)]$ than the ones treated thus far. Such a class of differential equations are ones which satisfy the following *natural assumptions*:

(a_1) $\tilde{A}(z) = \sum_0^\infty \tilde{A}_\nu z^{-\nu}$ *converges for* $|z| > a$ *and* $\tilde{A}_0$ *has all distinct eigenvalues* (*for a suitable* $a \geqq 0$).

(a_2) *If* $\tilde{F}_0$ *is a constant invertible matrix which diagonalizes* $\tilde{A}_0$, *then* $\operatorname{diag}\{\tilde{F}_0^{-1}\tilde{A}_1\tilde{F}_0\}$ *has no integer entries.*

In discussing how the quantities we shall define correspond to invariants of $[\tilde{A}(z)]$, it is important to have in mind an a priori fixed, but arbitrary, ordering for the eigenvalues $\lambda_1, \cdots, \lambda_n$ of $\tilde{A}_0$ and also to make a fixed, a priori, selection (depending only on $\tilde{A}_0$) for $\tilde{F}_0$ such that

$$\tilde{F}_0^{-1}\tilde{A}_0\tilde{F}_0 = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\} = \Lambda.$$

Under these assumptions, there exists a unique formal fundamental solution matrix for $[\tilde{A}(z)]$ of the form

$$\tilde{H}(z) = \tilde{F}_a(z) z^{\tilde{\Lambda}'} e^{\Lambda z},$$

where $\tilde{\Lambda}' = \operatorname{diag}\{\tilde{F}_0^{-1}\tilde{A}_1\tilde{F}_0\}$ satisfies assumption (i) (§ 1) and $\tilde{F}_a(z) = \sum_0^\infty \tilde{F}_\nu z^{-\nu}$ is a formal power series in z^{-1} which begins with the selected $\tilde{F}_0$.

Every other formal fundamental solution matrix of this type differs from the above $\tilde{H}(z)$ by a constant, invertible, diagonal, right-hand factor which corresponds exactly to the freedom in selecting $\tilde{F}_0$.

Although the above assumptions are natural to make from the point of view of being especially easy to check, there is a slightly more general class of differential equations $[\tilde{A}(z)]$ for which our results also apply and we will say that such differential equations satisfy our *basic assumptions* which we state as follows:

(b) $[\tilde{A}(z)]$ *is a meromorphic differential equation* (*at* ∞); *i.e.*,

$$\tilde{A}(z) = z^{r-1} \sum_0^\infty \tilde{A}_\nu z^{-\nu}$$

converges for $|z| > a$ *for some* $a \geqq 0$, *and is formally meromorphically equivalent* (see [1]) *to a special differential equation* $[A(z)] = \Lambda + A_1 z^{-1}$, *where the entries of* $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$ *are all distinct and* A_1 *satisfies assumption* (i).

Comparing the formal meromorphic invariants (see [1]) we find that $[\tilde{A}(z)]$ satisfies our basic assumptions if and only if it has a formal fundamental solution of the

form

$$\tilde{H}(z) = \tilde{F}_m(z) z^{\tilde{\Lambda}'} e^{\Lambda z}, \tag{4.1}$$

where $\tilde{F}_m(z) = \sum_\nu \tilde{F}_\nu z^{-\nu}$ is a formal meromorphic transformation (note that the sum may be assumed to be taken over all integers ν; however, for negative values of ν only finitely many coefficients $\tilde{F}_\nu$ are nonzero) and $\tilde{\Lambda}'$ has no integer entries. Since the entries of $\tilde{\Lambda}'$ are modulo one equal to $\Lambda' = \operatorname{diag}\{A_1\}$ we see that A_1 satisfies (i) if and only if $\tilde{\Lambda}'$ does. Finally note that since every formal meromorphic transformation $\tilde{F}_m(z)$ can be factored as (apply [1, Lemma 2] to $\tilde{F}_m^{-1}(z)$)

$$\tilde{F}_m(z) = \tilde{T}_m(z)\tilde{F}_a(z),$$

where $\tilde{T}_m(z)$ is an *actual* meromorphic transformation and $\tilde{F}_a(z)$ is a formal analytic transformation, $[\tilde{A}(z)]$ is meromorphically equivalent to an equation having a formal fundamental solution $\tilde{F}_a(z) z^{\tilde{\Lambda}'} e^{\Lambda z}$ and therefore satisfying our "natural assumptions." Hence the class of differential equations satisfying our basic assumptions can equally be characterized as those equations which may have Poincaré rank greater than one, but whose rank can (by means of a meromorphic transformation) be reduced to one and the resulting differential equation then satisfies the natural assumptions (a_1), (a_2). It is, however, theoretically of interest to know that *the results we shall derive depend only on a particular structure of the formal meromorphic invariants* and not on any additional features of the differential equation such as its Poincaré rank (which is not a meromorphic invariant).

Let any fixed differential equation $[\tilde{A}(z)]$ satisfying our basic assumptions be given. We select any formal fundamental solution matrix (4.1) and (analogously to (1.2)) we define the *associated functions corresponding to* $[\tilde{A}(z), \tilde{H}(z)]$ *for each* k, $1 \leqq k \leqq n$, *as*

$$\tilde{y}_k(t; \eta) = \tilde{y}_k(t) = \sum_\nu \tilde{f}_k(\nu)\Gamma(\tilde{\lambda}_k' + 1 - \nu)(t - \lambda_k)^{\nu - (\tilde{\lambda}_k' + 1)}, \tag{4.2}$$

where $\tilde{f}_k(\nu)$ denotes the kth column of $\tilde{F}_\nu$ and we define the nonintegral power as in § 1 by specifiying the branch of $\log(t - \lambda_k)$ for an admissible η. It can be shown (see, for example [3, § 2] or [19, p. 59]) that $\|\tilde{f}_k(\nu)\| \leqq c^\nu \nu!$ for ν sufficiently large, hence the power series in (4.2) converges for $|t - \lambda_k|$ sufficiently small. It is important to observe that $\tilde{y}_k(t)$ does not depend upon a particular *factorization* of a fixed $\tilde{H}(z)$ into $\tilde{F}_m(z) z^{\tilde{\Lambda}'} e^{\Lambda z}$, since all possible factorizations are obtained by replacing $\tilde{\lambda}_k'$ by $\tilde{\lambda}_k' + q_k$ for an arbitrary integer q_k and correspondingly replacing $\sum_\nu \tilde{f}_k(\nu) z^{-\nu}$ by $\sum_\nu \tilde{f}_k(\nu) z^{-\nu - q_k}$. Note that there are always only a finite number of indices $\nu < 0$ for which $\tilde{f}_k(\nu) \neq 0$. We also observe that $\tilde{y}_k(t)$ does depend upon the choice of $\tilde{H}(z)$, however, and when $\tilde{H}(z)$ is replaced by $\tilde{H}(z)D$, $D = \operatorname{diag}\{d_1, \cdots, d_n\}$, $\tilde{y}_k(t)$ is replaced by $\tilde{y}_k(t) d_k$.

4.2. The analytic structure of the associated functions. Our goal is to obtain the complete analytic structure of the associated functions in $\mathcal{P}_\eta$ and also their behavior at ∞. We state this result now as

THEOREM 1. *For any differential equation* $[\tilde{A}(z)]$ *satisfying our basic assumptions* (b), *let* $\tilde{H}(z)$ (*see* (4.1)) *denote any selected formal fundamental solution matrix, let* η *denote any admissible direction, and let* $\tilde{y}_k(t)$, $1 \leqq k \leqq n$, *denote the associated functions corresponding to* $([\tilde{A}(z)], \tilde{H}(z))$ *and* η. *Then the functions* $\tilde{y}_k(t)$, $1 \leqq k \leqq \eta$, *are analytic in* $\mathcal{P}_\eta$, *satisfy*

$$\tilde{y}_k(t) = \tilde{c}_{jk}\tilde{y}_j(t) + \operatorname{reg}(t - \lambda_j), \qquad 1 \leqq j, k \leqq n \tag{4.3}$$

and can be analytically continued along every path which does not contain any of the points $\lambda_1, \cdots, \lambda_n$. *Furthermore, if* S *is any sector in the t-plane of the form*

$$S = \{|t| > R,\ \alpha < \arg t < \beta\}$$

with $0 < \beta - \alpha < 2\pi$ *and* R *so large that none of the points* $\lambda_1, \cdots, \lambda_n$ *lies in* S, *then for any fixed analytic continuation of* $\tilde{y}_k(t)$ *into the whole sector* S, *we obtain*

$$\lim_{t\to\infty} e^{-(a+\varepsilon)|t|}\tilde{y}_k(t) = 0, \qquad t \in S, \tag{4.4}$$

for $\varepsilon > 0$ *arbitrary, where* a *is the radius of convergence of the power series for* $\tilde{A}(z)$.

To prove this, we first realize that $[\tilde{A}(z)]$ is *properly* meromorphically equivalent to a special differential equation $[A(z)]$, $A(z) = A_0 + z^{-1}A_1$. On one hand, according to the foregoing discussion we see that $[\tilde{A}(z)]$ is meromorphically equivalent to an equation satisfying our natural assumptions, and every such equation, according to a result of Birkhoff–Turrittin (see [5] and [18]), is meromorphically equivalent to such an $[A(z)]$. Furthermore, considering the formal meromorphic invariants and applying a constant transformation which puts A_0 into Jordan form, we find $A_0 = \Lambda$, and the entries of $\Lambda' = \operatorname{diag}\{A_1\}$ are (modulo one) congruent to the corresponding entries of $\tilde{\Lambda}'$; hence A_1 satisfies (i). For such a differential equation $[A(z)]$ we know from the previous sections that, corresponding to the unique formal fundamental solution matrix $H(z) = F_b(z)z^{\Lambda'}e^{\Lambda z}$ there exist associated functions $y_k(t)$ which are locally given by (1.2), analytic in $\mathcal{P}_\eta$ for every admissible η and satisfy (1.3). We first formalize the transfer of the behavior of the $y_k(t)$ to the $\tilde{y}_k(t)$ in the following lemma. For every integer ν, by $y_k^{(\nu)}(t)$ we denote the function whose power series expansion is obtained by $|\nu|$-fold termwise differentiation ($\nu \geqq 0$) resp. integration ($\nu < 0$) of the power series expansion of $y_k(t)$. For $\nu = -1$, the function $y_k^{(-1)}(t)$ can be analytically continued as antiderivative of $y_k(t)$, and similarly for $\nu = -2, -3, \cdots$.

LEMMA 2. *Consider any differential equation* $[\tilde{A}(z)]$ *satisfying our basic assumptions and any special differential equation* $[A(z)]$, $A(z) = \Lambda + A_1 z^{-1}$, *where* A_1 *satisfies* (i), *which are equivalent by means of a meromorphic transformation* $T(z)$. *Then for every admissible* η *and fixed* k, $1 \leqq k \leqq n$, *the associated functions* $\tilde{y}_k(t;\eta)$ *and* $y_k(t;\eta)$, *corresponding to the uniquely selected formal fundamental solutions* $\tilde{H}(z)$ *resp.* $H(z)$, *are related as follows*:

If $D = \operatorname{diag}\{d_1, \cdots, d_n\}$ *is such that* $\tilde{H}D = TH$, *and if*

$$T(z) = \sum_\nu T_\nu z^{-\nu},$$

then, for $t \in \mathcal{P}_\eta$, $|t - \lambda_k|$ *sufficiently small and every integer* $d > \operatorname{Re}\lambda_k'$,

$$\tilde{y}_k(t)d_k = \sum_{\nu \leqq d} T_\nu (-1)^\nu y_k^{(-\nu)}(t) + (-1)^d \int_{\lambda_k}^{t} \tilde{T}_d(t-s)\, y_k^{(-d)}(s)\, ds, \tag{4.5}$$

where the path of integration can be taken arbitrarily in $\mathcal{P}_\eta$ *and* $\tilde{T}_d(t)$ *is an entire function given by*

$$\tilde{T}_d(t) = \sum_{\nu=1}^{\infty} T_{\nu+d}\frac{(-1)^\nu}{\Gamma(\nu)}t^{\nu-1}. \tag{4.6}$$

Furthermore, the functions $\tilde{y}_k(t)$ *are analytic in* $\mathcal{P}_\eta$ *and satisfy* (4.3) *with* $\tilde{C} = [\tilde{c}_{jk}]$ *given by*

$$\tilde{C} = DCD^{-1}. \tag{4.7}$$

Proof. Since $T(z)=\sum_\nu T_\nu z^{-\nu}$ transforms $[\tilde{A}(z)]$ into $[A(z)]$,

$$T'(z)=\tilde{A}(z)T(z)-T(z)A(z),$$

which is a system of linear differential equations in the components of $T(z)$. Since the only possible singularities of $T(z)$ can come from singularities of $\tilde{A}(z)$ and $A(z)$, it follows that the series $\sum_\nu T_\nu z^{-\nu}$ converges for $|z|>a$ where a is the radius of convergence of $\tilde{A}(z)$. Therefore

$$\|T_\nu\|\leqq M_\delta(a+\delta)^\nu \quad \text{for all } \nu, \text{ and } \delta>0 \text{ arbitrary.} \tag{4.8}$$

Hence (4.6) converges for all t and defines an entire function.

For fixed k and η, we define, for fixed $d>\operatorname{Re}\lambda_k'$,

$$\hat{y}_k(t)=(-1)^d\int_{\lambda_k}^t \tilde{T}_d(t-s)y_k^{(-d)}(s)\,ds.$$

Then $\hat{y}_k(t)$ is analytic in $\mathscr{P}_\eta$ provided that the path of integration does not cross any cut. Note that the convergence of the integral at λ_k is guaranteed since

$$y_k^{(-d)}(s)=(-1)^d\sum_{\mu=d}^{\infty}\Gamma(\lambda_k'+1-\mu)f_k(\mu-d)(s-\lambda_k)^{\mu-(\lambda_k'+1)}$$

for $|s-\lambda_k|$ small enough. Hence for t close enough to λ_k, if we expand $\tilde{T}_d(t-s)$ and $y_k^{(-d)}(s)$ into power series and interchange summation and integration, we obtain

$$\hat{y}_k(t)=\sum_{\nu=1}^{\infty}T_{\nu+d}\frac{(-1)^\nu}{\Gamma(\nu)}\sum_{\mu=d}^{\infty}\Gamma(\lambda_k'+1-\mu)f_k(\mu-d)\int_{\lambda_k}^t(t-s)^{\nu-1}(s-\lambda_k)^{\mu-(\lambda_k'+1)}\,ds.$$

Making a change of variable $x=(s-\lambda_k)/(t-\lambda_k)$ (note that for t close to λ_k one may integrate along the straight line from λ_k to t), we obtain

$$\int_{\lambda_k}^t(t-s)^{\nu-1}(s-\lambda_k)^{\mu-(\lambda_k'+1)}\,ds=(t-\lambda_k)^{\nu+\mu-\lambda_k'-1}\int_0^1(1-x)^{\nu-1}x^{\mu-\lambda_k'-1}\,dx,$$

and, since the path of integration on the right is on the real axis and the power of x is defined to be the principal value, we get

$$\int_0^1(1-x)^{\nu-1}x^{\mu-\lambda_k'-1}\,dx=\frac{\Gamma(\nu)\Gamma(\mu-\lambda_k')}{\Gamma(\nu+\mu-\lambda_k')}=\frac{\Gamma(\nu)(-1)^\nu\Gamma(\lambda_k'+1-\nu-\mu)}{\Gamma(\lambda_k'+1-\mu)}.$$

This computation yields

$$\hat{y}_k(t)=\sum_{\nu=d+1}^{\infty}T_\nu\sum_{\mu=\nu}^{\infty}\Gamma(\lambda_k'+1-\mu)f_k(\mu-\nu)(t-\lambda_k)^{\mu-\lambda_k'-1};$$

hence

$$\begin{aligned}\sum_{\nu\leqq d}T_\nu(-1)^\nu y_k^{(-\nu)}+\hat{y}_k(t)&=\sum_\nu T_\nu\sum_{\mu=\nu}^{\infty}\Gamma(\lambda_k'+1-\mu)f_k(\mu-\nu)(t-\lambda_k)^{\mu-\lambda_k'-1}\\&=\sum_\mu\Gamma(\lambda_k'+1-\mu)\tilde{f}_k(\mu)(t-\lambda_k)^{\mu-\lambda_k'-1},\end{aligned}$$

where we define

$$\tilde{f}_k(\mu)=\sum_{\nu\leqq\mu}T_\nu f_k(\mu-\nu)$$

(note that this is a finite sum!). Hence the functions on the right-hand side of (4.5) are associated functions corresponding to the formal solution $T(z)H(z)$, and from $\tilde{H}D = TH$ and the fact that the definition of $\tilde{y}_k(t)$ does not depend on the factorization of the formal solution $\tilde{H}$ (hence we may factor $\tilde{H}D = T(z)F_b(z)z^{\Lambda'}e^{\Lambda z}$) it follows that (4.5) holds for $|t-\lambda_k|$ sufficiently small and then in all of $\mathcal{P}_\eta$ if we extend $\tilde{y}_k(t)$ analytically using (4.5).

To prove (4.7), take any pair of indices j, k, $1 \leqq j, k \leqq n$, and define $\tilde{C}$ by (4.7). Then for $d > \operatorname{Re} \lambda_k'$ and $d > \operatorname{Re} \lambda_j'$

$$d_k(\tilde{y}_k(t) - \tilde{y}_j(t)\tilde{c}_{jk})$$
$$= \sum_{\nu \leqq d} T_\nu(-1)^\nu \{y_k^{(-\nu)}(t) - y_j^{(-\nu)}(t)c_{jk}\}$$
$$+ (-1)^d \left\{ \int_{\lambda_k}^{t} \tilde{T}_d(t-s)y_k^{(-d)}(s)\,ds - \int_{\lambda_j}^{t} \tilde{T}_d(t-s)y_j^{(-d)}(s)\,ds\,c_{jk} \right\}.$$

From $y_k(t) = y_j(t)c_{jk} + \operatorname{reg}(t-\lambda_j)$ we conclude for every integer μ that

$$y_k^{(\mu)}(t) = y_j^{(\mu)}(t)c_{jk} + \operatorname{reg}(t-\lambda_j);$$

hence the sum on the right-hand side is regular at λ_j. If we rewrite the sum of the two integrals as

$$\int_{\lambda_k}^{\lambda_j} \tilde{T}_d(t-s)y_j^{(-d)}(s)\,ds\,c_{jk} + \int_{\lambda_k}^{t} \tilde{T}_d(t-s)(y_k^{(-d)}(s) - y_j^{(-d)}(s)c_{jk})\,ds$$

(note that the integrals exist since $y_k^{(-d)}$ and $y_j^{(-d)}$ have integrable singularities at λ_k and λ_j), then the second integral clearly is regular at λ_j (since the integrand is regular there), whereas the first integral even is an entire function of t. So we obtain

$$d_k(\tilde{y}_k(t) - \tilde{y}_j(t)\tilde{c}_{jk}) = \operatorname{reg}(t-\lambda_j), \qquad 1 \leqq j, k \leqq n,$$

which completes the proof of the lemma.

To complete the proof of Theorem 1, we note that (4.5) yields the analytic continuation of $\tilde{y}_k(t)$ along every path avoiding $\lambda_1, \cdots, \lambda_n$ (since $y_k(t)$ can be analytically continued along the path, and the path of integration may be taken to coincide with the selected path). To establish (4.4), note that from the estimate (4.8) it follows by estimating the power series (4.6)

$$\|\tilde{T}_d(t)\| \leqq M_\delta(a+\delta)^{d+1}e^{(a+\delta)|t|}.$$

Since $\|y_k(t)\| \leqq K|t|^\alpha$ for suitable K and α and all $t \in S$ sufficiently large, then possibly by enlarging K and α) the same estimate holds for $y_k^{(-\nu)}(t)$ for the finitely many values of $\nu \leqq d$ for which $T_\nu \neq 0$. Hence estimating (4.5) we find that (4.4) holds (with $\varepsilon = 2\delta$).

4.3. The invariance of $C(\eta)$ and the existence of $\tilde{Y}^*(t)$. Lemma 2 and Theorem 1 have the following consequences which are now listed for later use.

Remark 4.1. The proof of Lemma 2 (with d sufficiently large and a replaced by a sufficiently large constant) applies equally well when $T(z)$ is a meromorphic transformation from *any* $[A(z)]$ to *any* $[\tilde{A}(z)]$ both satisfying our basic assumptions. Hence it follows that for fixed admissible η the diagonal similarity class of C, i.e., $\{DCD^{-1}\}$, is a meromorphic invariant of $[A(z)]$. If

$$T(z) = I + \sum_1^\infty T_\nu z^{-\nu}$$

is a Birkhoff transformation and if both equations have rank one, then the selected formal fundamental solution matrices $H(z)=(F_0+F_1z^{-1}+\cdots)z^{\Lambda'}e^{\Lambda z}$ (resp. $\tilde{H}(z)=(\tilde{F}_0+\tilde{F}_1z^{-1}+\cdots)z^{\tilde{\Lambda}'}e^{\Lambda z}$) both start with the same a priori selected $F_0=\tilde{F}_0$; hence we find $D=I$. So C is a Birkhoff invariant of $[A(z)]$. Formula (4.5) in this case under the assumption $\operatorname{Re}\lambda_k'<0$ becomes (with $d=0$)

$$\tilde{y}_k(t)=y_k(t)+\int_{\lambda_k}^{t}\tilde{T}_0(t-s)y_k(s)\,ds.$$

Our goal in the next sections is to relate the Birkhoff invariant C to the normalized connection system (i.e., the Stokes' multipliers corresponding to the system of normal solutions (see [2]), which are also Birkhoff invariants).

Remark 4.2. In the situation of Theorem 1, if the matrix $\tilde{C}=(\tilde{c}_{jk})$, $1\leqq j,\ k\leqq n$, is invertible, then according to Remark 2.3 the matrix $\tilde{Y}^*(t)=\tilde{Y}(t)\tilde{C}^{-1}$ satisfies, for $k=1,\cdots,n$ (if $\tilde{y}_k^*(t)$ denotes the kth column of $\tilde{Y}^*(t)$),

$$\tilde{y}_k^*(t)=\tilde{y}_k(t)+\operatorname{reg}(t-\lambda_k),$$

$$\tilde{y}_k^*(t)=\operatorname{reg}(t-\lambda_j),\qquad j\neq k.$$

Moreover, the analytic continuation of $\tilde{y}_k(t)$ can be performed across any cut by means of (4.5) provided that the path of integration is deformed continuously with respect to t such that it does not cross any one of the singularities λ_j. Utilizing this, one immediately obtains analogues of Lemma 1 and Remarks 2.1, and 2.4 in § 2, as well as Propositions 3, 4 and 5 and Remarks 3.1, 3.2, 3.3, 3.4 and 3.5 in § 3 under the assumption that $\tilde{C}$ is invertible.

These formulas determine explicitly the analytic continuation of the matrices $\tilde{Y}$ and $\tilde{Y}^*$ across the cuts in any fixed direction η onto their full Riemann surface. Thus the direction η provides us with a particular realization of the abstract Riemann surface, and its independence of η is explained by the connection formulas.

Remark 4.3. Recall from Proposition 2 that C corresponding to a special differential equation is invertible if and only if (ii′) holds. Moreover, from (4.7) we see that $\tilde{C}$ is invertible if and only if C is. If it happens that $\tilde{C}$ is not invertible for some differential equation $[\tilde{A}(z)]$, then a scalar shift z^{γ} takes $[\tilde{A}(z)]$ into $[\tilde{A}(z)-\gamma z^{-1}I]$ and any meromorphically equivalent special differential equation $[A(z)]$ into $[A(z)-\gamma z^{-1}I]$. Since the eigenvalues of A_1 for all meromorphically equivalent special differential equations are congruent modulo one, then in order to make C, hence $\tilde{C}$, invertible, it is only necessary to select γ so that $A_1-\gamma I$ has no eigenvalues which are integers. This occurs for all but a discrete set of γ. The assumption (ii′) is equivalent to the property that the corresponding special differential equation has no single-valued solutions and since this is meromorphically invariant, we see that $\tilde{C}$ is invertible if and only if the general differential equation has no nontrivial single-valued solutions.

4.4. The influence of a scalar shift on the associated functions. For the purpose of later use, we now will establish how the associated functions change with respect to a scalar shift z^{γ}. In case γ is an integer, this is just a very simple meromorphic transformation, and the statements in the following Lemma could also be derived from Lemma 2. Hence the interesting case is where γ is not an integer, but we state the lemma generally.

LEMMA 2′. *Suppose that both* $[A(z)]$ *and* $[A(z)-\gamma z^{-1}I]$ *satisfy our basic assumptions for a suitable complex* γ. *If* $H(z)=F_m(z)z^{\Lambda'}e^{\Lambda z}$ *is a selected formal fundamental solution of* $[A(z)]$, *and if we select* $\tilde{H}(z)=F_m(z)z^{\Lambda'-\gamma I}e^{\Lambda z}$ *as a formal fundamental solution for* $[A(z)-\gamma z^{-1}I]$, *then for every integer* ν *and* $k=1,\cdots,n$, *the associated*

functions $y_{\nu,k}(t)$ *and* $\tilde{y}_{\nu,k}(t)$ *are related by*

$$\tilde{y}_{\nu,k}^{(-d_1-d_2)}(t)=\frac{\sin \pi\lambda'_k}{\Gamma(d_1+\gamma)\sin \pi(\lambda'_k-\gamma)}\int_{\lambda_k}^{t}(t-s)^{d_1+\gamma-1}y_{\nu,k}^{(-d_2)}(s)\,ds, \tag{4.9}$$

where d_1, d_2 *are sufficiently large integers, the path of integration is taken in* $\mathscr{P}_\eta$ *(for fixed* $\eta\in(\eta_{\nu+1},\eta_\nu)$*) and the power of* $(t-s)$ *is defined continuously along the path of integration such that at* $s=\lambda_k$ *it coincides with our usual definition. Furthermore, the corresponding matrices* C_ν *(resp.* $\tilde{C}_\nu$*) are related by*

$$(e^{2\pi i\lambda'_k}-1)c_{jk}^{(\nu)}=(e^{2\pi i(\lambda'_k-\gamma)}-1)\tilde{c}_{jk}^{(\nu)}\quad \text{if } j<k \text{ in } S'_{\nu+1}, \tag{4.10}$$

$$(e^{2\pi i\lambda'_k}-1)c_{jk}^{(\nu)}=(e^{2\pi i(\lambda'_k-\gamma)}-1)\,e^{2\pi i\gamma}\tilde{c}_{jk}^{(\nu)}\quad \text{if } k<j \text{ in } S'_{\nu+1}. \tag{4.11}$$

Proof. Similarly as in the proof of Lemma 2, one finds for $|t-\lambda_k|$ small, using the power series expansion of $y_{\nu,k}^{(-d_2)}(s)$ and taking d_1, d_2 so large that the integral converges at both t and λ_k:

$$\int_{\lambda_k}^{t}(t-s)^{d_1+\gamma-1}y_{\nu,k}^{(-d_2)}(s)\,ds=(-1)^{d_2}\sum_{\mu}f_k(\mu-d_2)\Gamma(\lambda'_k+1-\mu)\int_{\lambda_k}^{t}(t-s)^{d_1+\gamma-1}(s-\lambda_k)^{\mu-\lambda'_k-1}\,ds.$$

Using the same change of variable $x=(s-\lambda_k)/(t-\lambda_k)$ and integrating on a straight line which yields $\arg(t-s)=\arg(t-\lambda_k)=\arg(s-\lambda_k)$, we find, as in the proof of Lemma 2,

$$\int_{\lambda_k}^{t}(t-s)^{d_1+\gamma-1}(s-\lambda_k)^{\mu-\lambda'_k-1}\,ds=(t-\lambda_k)^{d_1+\gamma+\mu-\lambda'_k-1}\frac{\Gamma(d_1+\gamma)\Gamma(\mu-\lambda'_k)}{\Gamma(d_1+\gamma+\mu-\lambda'_k)},$$

and using well-known identities for the gamma-function, we find

$$\begin{aligned}&\frac{\sin \pi\lambda'_k}{\Gamma(d_1+\gamma)\sin \pi(\lambda'_k-\gamma)}\int_{\lambda_k}^{t}(t-s)^{d_1+\gamma-1}y_{\nu,k}^{(-d_2)}(s)\,ds\\&\quad=(-1)^{d_1+d_2}\sum_{\mu}f_k(\mu-d_1-d_2)\Gamma(\lambda'_k-\gamma+1-\mu)(t-\lambda_k)^{\mu-(\lambda'_k-\gamma)-1}\\&\quad=\tilde{y}_{\nu,k}^{(-d_1-d_2)}(t),\end{aligned}$$

which proves (4.9).

In order to prove (4.10), (4.11), we first note that differentiation (resp., termwise integration) of the associated functions corresponds to a transformation z^γ, where γ is an integer, and by Lemma 2 this does not influence the matrices C_ν (resp. $\tilde{C}_\nu$); hence, by applying such integer shifts to both differential equations, we may assume that $d_1=d_2=0$. Now let $k\neq j$ be fixed. If $t_0\neq\lambda_j$ is any point on the cut $\arg(t-\lambda_j)=\eta$, and if by $\tilde{y}_{\nu,k}^{+}(t_0)$(resp. $\tilde{y}_{\nu,k}^{-}(t_0)$) we denote the boundary values of $\tilde{y}_{\nu,k}(t)$ as $t\to t_0$ from the left (resp. from the right) (i.e., as $|t|\to|t_0|$, $\arg t\to(\eta-2\pi)+0$ (resp. $\arg t\to\eta-0$)), then $\tilde{c}_{jk}^{(\nu)}$ can be characterized as the unique number for which

$$\begin{aligned}\tilde{y}_{\nu,k}^{+}(t_0)-\tilde{y}_{\nu,k}^{-}(t_0)&=\{\tilde{y}_{\nu,j}^{+}(t_0)-\tilde{y}_{\nu,j}^{-}(t_0)\}\tilde{c}_{jk}^{(\nu)}\\&=\tilde{y}_{\nu,j}^{+}(t_0)(1-e^{-2\pi i(\lambda'_j-\gamma)})\tilde{c}_{jk}^{(\nu)}\end{aligned} \tag{4.12}$$

(with analogous interpretations of $\tilde{y}_{\nu,j}^{+}(t_0)$ and $\tilde{y}_{\nu,j}^{-}(t_0)$). From (4.9) we see that

$$\tilde{y}_{\nu,k}^{\pm}(t_0)=\frac{\sin \pi\lambda'_k}{\Gamma(\gamma)\sin \pi(\lambda'_k-\gamma)}\left\{\int_{\lambda_k}^{\lambda_j}(t_0-s)^{\gamma-1}y_{\nu,k}(s)\,ds+\int_{\lambda_j}^{t_0}(t_0-s)^{\gamma-1}y_{\nu,k}^{\pm}(s)\,ds\right\},$$

where the second integral is taken along the cut, and in both cases the definition of $\arg(t_0-s)$ is the same, but depends on whether the kth cut is to the left of the jth cut or vice versa. The first possibility occurs if and only if $j<k$ in $S'_{\nu+1}$, and we then have to take $\arg(t_0-s)=\eta$, whereas for $k<j$ in $S'_{\nu+1}$ (i.e., whenever the kth cut is to the right of the jth cut) we take $\arg(t_0-s)=\eta-2\pi$. Hence we find in both cases, using

$$y^+_{\nu,k}(s)-y^-_{\nu,k}(s)=y^+_{\nu,j}(s)(1-e^{-2\pi i\lambda'_j})c^{(\nu)}_{jk},$$

$$\tilde{y}^+_{\nu,k}(t_0)-\tilde{y}^-_{\nu,k}(t_0)=\frac{\sin\pi\lambda'_k(1-e^{-2\pi i\lambda'_j})c^{(\nu)}_{jk}}{\Gamma(\gamma)\sin\pi(\lambda'_k-\gamma)}\int_{\lambda_j}^{t_0}(t_0-s)^{\gamma-1}y^+_{\nu,j}(s)\,ds.$$

Since in the case $k<j$ in the integral $\arg(t-s)$ is taken to have the "correct" value $\eta-2\pi$, we find

$$\frac{\sin\pi\lambda'_j}{\Gamma(\gamma)\sin\pi(\lambda'_j-\gamma)}\int_{\lambda_j}^{t_0}(t_0-s)^{\gamma-1}y^+_{\nu,j}(s)\,ds=\tilde{y}^+_{\nu,j}(t_0),$$

whereas for $j<k$, $\arg(t-s)$ has value η; hence

$$\frac{\sin\pi\lambda'_j}{\Gamma(\gamma)\sin\pi(\lambda'_j-\gamma)}\int_{\lambda_j}^{t_0}(t_0-s)^{\gamma-1}y^+_{\nu,j}(s)\,ds=e^{2\gamma\pi i}\tilde{y}^+_{\nu,j}(t_0).$$

Hence we find

$$\tilde{y}^+_{\nu,k}(t_0)-\tilde{y}^-_{\nu,k}(t_0)=\frac{\sin\pi\lambda'_k\sin\pi(\lambda'_j-\gamma)}{\sin\pi(\lambda'_k-\gamma)\sin\pi\lambda'_j}(1-e^{-2\pi i\lambda'_j})c^{(\nu)}_{jk}\tilde{y}^+_{\nu,j}(t_0)\quad\text{if } k<j \text{ in } S'_{\nu+1},$$

$$\tilde{y}^+_{\nu,k}(t_0)-\tilde{y}^-_{\nu,k}(t_0)=\frac{\sin\pi\lambda'_k\sin\pi(\lambda'_j-\gamma)}{\sin\pi(\lambda'_k-\gamma)\sin\pi\lambda'_j}(1-e^{-2\pi i\lambda'_j})c^{(\nu)}_{jk}e^{2\gamma\pi i}\tilde{y}^+_{\nu,j}(t_0)\quad\text{if } j<k \text{ in } S'_{\nu+1},$$

Since

$$\frac{\sin\pi\lambda'_k\sin\pi(\lambda'_j-\gamma)}{\sin\pi(\lambda'_k-\gamma)\sin\pi\lambda'_j}=\frac{(e^{2\pi i\lambda'_k}-1)(1-e^{-2\pi i(\lambda'_j-\gamma)})}{(e^{2\pi i(\lambda'_k-\gamma)}-1)(1-e^{-2\pi i\lambda'_j})}e^{-2\pi i\gamma},$$

using (4.12) we find that (4.10) and (4.11) follow.

Remark 4.4. From Propositions 4 and 5 it is now easily seen that the matrices W_ν and C^+_ν, C^-_ν (for all integers ν) are invariant with respect to a shift z^γ. This will be of importance later.

5. Representations of the normal solutions using Laplace integrals and convergent factorial series

5.1. Laplace integrals. Let $[A(z)]$ be a differentia! equation satisfying our basic assumptions, let $H(z)=F_m(z)z^{\Lambda'}e^{\Lambda z}$ denote a fixed formal fundamental solution matrix and let the Stokes' directions τ_ν be defined by (3.3) for all integers ν. We recall from the general theory of invariants (see [2, § 5]) that the *system of normal solutions* $X_\nu(z)$, $\nu=0,\pm1,\pm2,\cdots$ corresonding to the pair $([A(z)],H(z))$ is characterized by the properties that for each integer

(Ia) $$X_\nu(z)\cong H(z)\quad\text{as } z\to\infty,\qquad z\in S_\nu=S(\tau_{\nu-1},\tau_{\nu+1})$$

and for $z\in S'_\nu=S(\tau_{\nu-1},\tau_\nu)$,

(Ib) $$V_\nu=X_\nu^{-1}(z)X_{\nu-1}(z)\in\mathcal{U}(\rho_\nu);$$

i.e., the support of V_ν is contained in the position set ρ_ν. In the special case of distinct eigenvalues of A_0 (our natural assumptions) or more generally for our basic assump-

tions since the eigenvalues of Λ are distinct, it can be shown (see [4]) that *each single* normal solution matrix $X_\nu(z)$ is also characterized by the property that

(II) $$X_\nu(z) \cong H(z) \quad \text{as } z \to \infty, \qquad z \in S(\tau_\nu - \pi, \tau_{\nu+1}).$$

A sector of this type $S(\tau_\nu - \pi, \tau_{\nu+1})$ is referred to as an *enlarged half-plane.*

It has been long established (see [5], [9], [10], [11], [16], [17]) (in varying degrees of generality) that solutions of differential equations of this type have convergent Laplace integral expansions in certain half-planes, the formal series may be summed as convergent generalized factorial series in the same half-planes, and the solutions have the asymptotic expansion $H(z)$ in sectors which have various angular openings (up to enlarged half-planes). The contours of integration which have been used include both standard rays and also loops which are asymptotic to certain rays.

In this section we will show that the normal solutions can be expressed as convergent Laplace integrals of the matrices of associated functions $Y_\nu(t)$ and $Y_\nu^*(t)$ using loop contours. We also show the normal solutions can be expressed in terms of convergent factorial series expansions and in the process we show that the associated functions $y_k(t)$ and some corresponding functions $\psi_k(t)$ (both of which are defined locally) can be analytically continued into certain natural domains by explicit methods, thus producing solutions which can be considered as *effectively calculable.* As compared with the classical theorems, we wish to emphasize the following points which distinguish our results.

(1) The normal solutions are identified by selecting the appropriate matrix of associated functions $Y_\nu(t)$ (or $Y_\nu^*(t)$) and particular contours of integration.

(2) The half-planes of convergence for both the Laplace integrals and the factorial series are explicitly given and may be considered as generally the optimal such domains of convergence.

(3) The normalized connection system (V_ν), i.e., the Stokes' multipliers corresponding to the normal solutions, are shown to be all calculated explicitly from any one of the matrices C_ν; hence the matrix C_ν not only determines the analytic continuation of the associated functions $y_k(t)$, but that of the normal solutions of the differential equation $[A(z)]$ as well.

In reference to point (2), the particular information which makes it possible to give the optimal half-planes of convergence is the precise knowledge of the growth of $y_k(t)$ at ∞ (see (4.4)). This, as well as the complete analytic description of $y_k(t)$ in Lemma 2, distinguishes our methods from the classical ones. Wasow [19, pp. 338–339] has pointed out that the classical methods do not yield this information because they do not take full advantage of the analyticity of $A(z)$ in a neighborhood of ∞.

Note that, while the associated functions were considered in a t-plane with certain cuts and choices of logarithms, it is convenient to think of the variable z to vary on the Riemann surface of the logarithm described by two parameters $|z|$ and arg z (compare [1, § 1]). Hence non-integer powers of z will always have a clear meaning, since they are single valued functions on the Riemann surface, if we define

$$z^\alpha = e^{\alpha(\log|z| + i \arg z)}.$$

LEMMA 3. *Let $[A(z)]$ satisfy our basic assumptions and let $H(z) = F_m(z) z^{\Lambda'} e^{\Lambda z}$ be a fixed selected formal fundamental solution. Then for each fixed k and admissible η, we define*

(5.1) $$x_k(z; \eta) = \frac{1}{2\pi i} \int_{\gamma_k(\eta)} e^{zt} y_k(t; \eta)\, dt,$$

where $y_k(t;\eta)$ *is the* kth *function associated with* $[A(z)]$ *and* $H(z)$, *and the path of integration* $\gamma_k(\eta)$ *is the contour in* $\mathcal{P}_\eta$ *from* ∞ *along the left-hand side of the cut* $\arg(t-\lambda_k)=\eta$, *around* λ_k *in the positive sense, and back to* ∞ *along the right-hand side of the cut. Then, if* a *denotes the radius of convergence of* $A(z)$, *the integral* (5.1) *converges for*

$$z\in\mathcal{S}(\eta)=\left\{\operatorname{Re}(z\,e^{i\eta})<-a;\ \frac{\pi}{2}-\eta<\arg z<\frac{3\pi}{2}-\eta\right\}$$

and represents an analytic function satisfying

$$x_k(z;\eta)\cong f_k(z)z^{\lambda_k'}e^{\lambda_k z}\quad \text{as } z\to\infty,\quad z\in\mathcal{S}(\eta) \tag{5.2}$$

with $f_k(z)$ *denoting the* kth *column of* $F_m(z)$.

Proof. The convergence of the integral for $z\in\mathcal{S}(\eta)$ is an immediate consequence of (4.4). To obtain the asymptotic expansion (5.2), we first realize that for $z\in\mathcal{S}(\eta)$

$$\frac{1}{2\pi i}\int_{\gamma_k(\eta)}(t-\lambda_k)^{-\alpha}e^{zt}\,dt=\frac{z^{\alpha-1}e^{\lambda_k z}}{\Gamma(\alpha)}$$

for any complex α which is not a negative integer or zero, where we define the power $z^{\alpha-1}$ according to the selection of $\arg z$ as in (5.2); this formula is, e.g., given in [8, p. 226] for $\eta=\pi$ and in general is found to hold by analytic continuation (by means of rotating the path of integration). Hence the right-hand side of (5.2) is obtained by termwise integration of the expansion of $y_k(t)$ at $t=\lambda_k$, and (5.2) can therefore be proven in a standard manner by expressing $y_k(t)$ as a finite part of the expansion plus an error term whose Laplace integral can be easily estimated.

5.2. Representation of the normal solutions as Laplace integrals. Now let any fixed integer ν be given and we claim that $x_k(z;\eta)$ *does not depend on* η *as long as* η *varies in* $(\eta_{\nu+1},\eta_\nu)$. To show this, consider two values $\eta,\tilde\eta\in(\eta_{\nu+1},\eta_\nu)$ and a reference point t_0 with respect to $\eta,\tilde\eta$ which lies on $\gamma_k(\eta)$. If we now turn the contour $\gamma_k(\eta)$ such that it comes to be in direction $\tilde\eta$ (and still passes through t_0), then it is easily seen that the integral does not change its value (for $z\in\mathcal{S}(\eta)\cap\mathcal{S}(\tilde\eta)$) if we define the integrand by keeping its values fixed at $t=t_0$ and extending it analytically along the new path of integration. But in doing so, we find that the new integrand is $e^{zt}y_k(t;\tilde\eta)$ whereas the rotated path of integration can be taken as $\gamma_k(\tilde\eta)$. Hence

$$x_k(z;\eta)=x_k(z;\tilde\eta)\quad \text{for } z\in\mathcal{S}(\eta)\cap\mathcal{S}(\tilde\eta) \text{ if } \eta,\tilde\eta\in(\eta_{\nu+1},\eta_\nu).$$

Therefore, if we combine the vectors $x_1(z;\eta),\cdots,x_n(z;\eta)$ into a matrix $X_\nu(z)$, then $X_\nu(z)$ is an analytic function for

$$z\in\mathcal{S}_\nu=\bigcup_{\eta_{\nu+1}<\eta<\eta_\nu}\mathcal{S}(\eta),$$

whose columns in every half-plane $\mathcal{S}(\eta)$, $\eta_{\nu+1}<\eta<\eta_\nu$, can be represented as Laplace integrals. Furthermore, since a closed subsector of $\mathcal{S}_\nu$ always can be covered by finitely many (in fact two) half planes $\mathcal{S}(\eta)$, $\eta\in(\eta_{\nu+1},\eta_\nu)$, we find that

$$X_\nu(z)\cong H(z)\quad \text{in } \mathcal{S}_\nu \text{ for every integer } \nu.$$

So far, we did not say that $X_\nu(z)$ is a solution matrix of our given differential equation $[A(z)]$. This however is one of the consequences of

THEOREM 2. *Let* $[A(z)]$ *be any differential equation satisfying our basic assumptions, and let* $H(z)=F_m(z)z^{\Lambda'}e^{\Lambda z}$ *be a selected formal fundamental solution of* $[A(z)]$.

Then for every integer ν, the matrix $X_\nu(z)$, defined by

$$X_\nu(z)=\frac{1}{2\pi i}\int_{\gamma_1(\eta),\cdots,\gamma_n(\eta)} e^{zt}Y_\nu(t)\,dt \quad \text{for } z\in\mathcal{S}(\eta), \tag{5.3}$$

where η can be taken arbitrarily from the interval $(\eta_{\nu+1}, \eta_\nu)$ and the path of integration for the kth column of $Y_\nu(t)$ is taken along $\gamma_k(\eta)$, is the νth normal solution corresponding to $([A(z)], H(z))$. Moreover, the normalized connection matrix V_ν is equal to W_ν, hence the normalized connection system $(V_1, \cdots, V_m)$ is in one-to-one correspondence and can be explicitly calculated in terms of C_ν for any integer ν.

For a proof of Theorem 2 it is technically convenient to first treat the case when the matrices $Y_\nu^*(t)$ (for all integers ν) exist. Even theoretically this is an especially handsome case since then we will be able to choose a common path of integration in (5.3) instead of having different paths for the different columns of $X_\nu(t)$ (integrating $Y_\nu^*(t)$ instead of $Y_\nu(t)$). To that extent, we make the following additional assumption on $[A(z)]$:

(iii) *Let $[A(z)]$ have no single-valued vector solution.* According to the discussion in Remark 4.3, this is equivalent to the property that C_ν is invertible for every integer ν, hence by means of Remark 4.2 assumption (iii) implies the existence of $Y_\nu^*(t)$ for every integer ν. Under this additional assumption we have

THEOREM 2'. *In the situation of Theorem 2, let $[A(z)]$ additionally satisfy (iii). Then for every integer ν we have, for $X_\nu(z)$ defined by* (5.3),

$$X_\nu(z)=\frac{1}{2\pi i}\int_{\gamma(\eta)} e^{zt}Y_\nu^*(t)\,dt \quad \text{for } z\in\mathcal{S}(\eta), \tag{5.4}$$

where η can be taken arbitrarily from the interval $(\eta_{\nu+1}, \eta_\nu)$ and the path of integration $\gamma(\eta)$ is taken as the contour from ∞ along a parallel to direction arg $t=\eta$, *sufficiently far out on the left, around all the singularities $\lambda_1, \cdots, \lambda_n$ in the positive sense and back to ∞ along a parallel to* arg $t=\eta$, *sufficiently far out on the right. Furthermore, for every integer ν we find*

$$X_{\nu-1}(z)=X_\nu(z)W_\nu \quad \text{for } z\in\mathcal{S}_{\nu-1}\cap\mathcal{S}_\nu, \tag{5.5}$$

where W_ν is defined as in Proposition 4.

Proof of Theorem 2'. Take any fixed ν, k and η and let $y_{\nu,k}^*(t)$ denote the kth column of $Y_\nu^*(t)$. Since $e^{zt}(y_{\nu,k}^*(t)-y_k(t;\eta)$ is an analytic function in the closed region containing λ_k and bounded by $\gamma_k(\eta)$, we see by using Cauchy's theorem that $y_k(t;\eta)$ in (5.1) may be replaced by $y_{\nu,k}^*(t)$ and from the same theorem (since $y_{\nu,k}^*(t)$ stays regular at λ_j for $j\neq k$) we conclude that the path of integration $\gamma_k(\eta)$ can be replaced by $\gamma(\eta)$. This proves (5.4).

To obtain (5.5), we consider any $\eta\in(\eta_{\nu+1}, \eta_\nu)$ and $\tilde{\eta}\in(\eta_\nu, \eta_{\nu-1})$. Then if t_0 is a reference point on $\gamma(\eta)$ with respect to η and $\tilde{\eta}$, we may continuously deform $\gamma(\eta)$ (by keeping t_0 fixed) such that its two rays finally become parallel to arg $t=\tilde{\eta}$, and we may do this without crossing any point $\lambda_1, \cdots, \lambda_n$. The new contour finally obtained by this deformation of $\gamma(\eta)$ may then be taken as $\gamma(\tilde{\eta})$, and if we keep the same value for $Y_\nu^*(t)$ at $t=t_0$ and define it along $\gamma(\tilde{\eta})$ by means of analytic continuation, then with this temporary interpretation of $Y_\nu^*(t)$ along $\gamma(\tilde{\eta})$

$$\frac{1}{2\pi i}\int_{\gamma(\tilde{\eta})} e^{zt}Y_\nu^*(t)\,dt \quad \text{for } z\in\mathcal{S}(\tilde{\eta})$$

gives the analytic continuation of $X_\nu(z)$ into $\mathcal{S}(\tilde{\eta})$. Since $Y_{\nu-1}^*(t)=Y_\nu^*(t)W_\nu$ at $t=t_0$ and therefore (by means of analytic continuation) along $\gamma(\tilde{\eta})$ (for this temporary

interpretation of $Y_\nu^*(t)$) we obtain

$$X_{\nu-1}(z)=\frac{1}{2\pi i}\int_{\gamma(\tilde{\eta})} e^{zt}Y_{\nu-1}^*(t)\,dt=\frac{1}{2\pi i}\int_{\gamma(\tilde{\eta})} e^{zt}Y_\nu^*(t)\,dt\,W_\nu$$

for $z\in\mathcal{S}(\tilde{\eta})$; hence (5.5) follows.

Proof of Theorem 2. We first prove Theorem 2 under the additional assumption (iii) in part (α), and then in part (β) we will remove this extra assumption.

(α) As explained earlier in this section, the normal solutions can be characterized by (Ia), (Ib) or by (II). If we already knew that $X_\nu(z)$ defined by (5.3) is a solution matrix of $[A(z)]$, then using Theorem 2′ we would be finished in this case since every closed subsector of $S(\tau_\nu-\pi,\tau_{\nu+1})$ is contained in $\mathcal{S}_\nu$ (at least for $|\tau|$ sufficiently large). So it remains to show that X_ν are solutions of $[A(z)]$. Of course, using $X_\nu(z)\cong H(z)$ in $\mathcal{S}_\nu$ and the fact that $X_\nu(z)$ is invertible for $|z|$ sufficiently large in $\mathcal{S}_\nu$, it is easy to see that $X_\nu'(z)X_\nu^{-1}(z)\cong A(z)$ in $\mathcal{S}_\nu$. But this alone is not sufficient to guarantee that $X_\nu(z)$ is a solution of $[A(z)]$. In order to bring this about we use that the $X_\nu(z)$ satisfy the *closing condition*

$$X_{\nu+m}(z\,e^{2\pi i})=X_\nu(z)\,e^{2\pi i\Lambda'}\quad\text{for } z\in\mathcal{S}_\nu \text{ and every integer } \nu.$$

(This is easy to show using $Y_{\nu+m}^*(t)=Y_\nu^*(t)\,e^{2\pi i\Lambda'}$ and the fact that $\gamma(\eta+2\pi)$ may be taken equal to $\gamma(\eta)$.) Hence the logarithmic derivatives $X_\nu'(z)X_\nu^{-1}(z)$, $1\leqq\nu\leqq m$, in a full neighborhood of ∞ combine (by (5.5)) to yield (see [12, p. 161]) the single-valued analytic function $A(z)$.

(β) In the general situation of Theorem 2 for a suitable selection of γ, the transformation $x=z^\gamma\tilde{x}$ takes $[A(z)]$ into $[A(z)-\gamma z^{-1}I]$, which satisfies (iii) (and our basic assumptions). From Remark 4.4 we recall that the quantities W_ν do not change. Furthermore, we see by partial integration that

$$x_k(t;\eta)=\frac{(-1)^d z^d}{2\pi i}\int_{\gamma_k(\eta)} e^{zt}y_k^{(-d)}(t;\eta)\,dt,$$

and for d large enough, the singularity of $y_k^{(-d)}(t;\eta)$ at λ_k becomes integrable. Hence in this case, we may take the path of integration to be completely on the cut $\arg(t-\lambda_k)=\eta$, from ∞ to λ_k on the left border and back to ∞ on the right border. By comparing the different values of $y_k^{(-d)}(t;\eta)$ on the two borders, we find

$$x_k(z;\eta)=\frac{(-1)^d z^d}{2\pi i}(1-e^{2\pi i\lambda_k'})\int_{\lambda_k}^{\infty(\eta)} e^{zt}y_k^{(-d)}(t;\eta)\,dt,$$

where we integrate on the right border of the cut.

If now $\tilde{y}_k(t;\eta)$ is the kth function associated to $[A(z)-\gamma z^{-1}I]$ (with $\tilde{H}(z)=F_m(z)z^{\Lambda'-\gamma I}e^{\Lambda z}$ as the selected formal solution), then we conclude, from Lemma 2′

$$y_k^{(-d_1-d_2)}(t;\eta)=\frac{\sin\pi(\lambda_k'-\gamma)}{\Gamma(d_1-\gamma)\sin\pi\lambda_k'}\int_{\lambda_k}^{t}(t-s)^{d_1-\gamma-1}\tilde{y}_{\nu,k}^{(-d_2)}(s)\,ds$$

for d_1, d_2 large enough. Taking $d=d_1+d_2$ and interchanging the order of integration we find

$$x_k(t;\eta)=\frac{(-1)^{d_1+d_2+1}z^{d_1+d_2}}{\pi}e^{\pi i\lambda_k'}\frac{\sin\pi(\lambda_k'-\gamma)}{\Gamma(d_1-\gamma)}\int_{\lambda_k}^{\infty(\eta)} e^{zs}\tilde{y}_{\nu,k}^{(-d_2)}(s)\,ds\int_0^{\infty(\eta)} u^{d_1-\gamma-1}e^{zu}\,du,$$

and since

$$\int_{\lambda_k}^{\infty(\eta)} e^{zs}\tilde{y}_{\nu,k}^{(-d_s)}(s)\,ds = \frac{2\pi i}{(-1)^{d_2}z^{d_2}}(1-e^{2\pi i(\lambda_k'-\gamma)})^{-1}\tilde{x}_k(z;\eta),$$

$$\int_0^{\infty(\eta)} u^{d_1-\gamma-1}e^{zu}\,du = z^{\gamma-d_1}e^{(d_1-\gamma)\pi i}\Gamma(d_1-\gamma), \qquad z\in\mathcal{S}(\eta),$$

we find

$$z^{\gamma}\tilde{x}_k(z;\eta) = x_k(z;\eta).$$

This shows that even in the general case the matrix (5.3) is a solution of the differential equation, namely the νth normal solution (since it has the correct asymptotic in an enlarged half-plane). Furthermore, since the scalar shift changed neither W_ν nor V_ν, we see that also in general $V_\nu = W_\nu$, which completes the proof of Theorem 2.

Remark 5.1. An independent proof that $X_\nu(z)$ satisfies the differential equation $[A(z)]$ may be obtained by showing that kth column of $Y(t;\eta)$ satisfies

$$ty_k(t) = \sum_{\mu\leqq d} A_\mu(-1)^\mu y_k^{(-\mu)}(t) + (-1)^d\int_{\lambda_k}^{t}\tilde{A}_d(t-s)y_k^{(-d)}(s)\,ds,$$

with

$$\tilde{A}_d(t) = \sum_{\mu=1}^{\infty} A_{\mu+d}\frac{(-1)^\mu}{\Gamma(\mu)}t^{\mu-1},$$

if one writes

$$A(z) = \sum_{\mu} A_\mu z^{-\mu},$$

(compare [9, p. 389]).

5.3. Representation of the normal solutions as factorial series. To establish the factorial series representation of the normal solutions, we first discuss the properties of a function $\psi_k(u)$ $(k=1,\cdots,n)$ which in a sense is the associated function in case $\lambda_k' = 0$.

LEMMA 4. *Let $[A(z)]$ satisfy our natural assumptions and $H(z) = F_a(z)z^{\Lambda'}e^{\Lambda z}$ be a selected formal fundamental solution. Then for every fixed k, the function $\psi_k(u)$ which is locally defined by*

$$\psi_k(u) = \sum_1^{\infty} f_k(\nu)\frac{(-1)^\nu}{\Gamma(\nu)}u^{\nu-1} \tag{5.6}$$

has the following properties:

(i) *For every admissible η, $\psi_k(u)$ is analytic in an open infinite strip containing the ray* $\arg u = \eta$ *and not containing any of the points $\lambda_j - \lambda_k (j\neq k)$.*

(ii) *In the strip described in* (i), *$\psi_k(u)$ satisfies*

$$\lim_{u\to\infty} e^{-(a+\varepsilon)|u|}\psi_k(u) = 0 \tag{5.7}$$

for $\varepsilon > 0$ arbitrary and a being the radius of convergence of the expansion of $A(z)$.

(iii) *If $\eta\in(\eta_{\nu+1},\eta_\nu)$ for any integer ν, then for the k*th *column of $X_\nu(z)$ we find*

$$x_{\nu,k}(z) = e^{\lambda_k z}z^{\lambda_k'}\left\{f_k(0) + \int_0^{\infty(\eta)}\psi_k(u)\,e^{zu}\,du\right\} \quad \text{for } z\in\mathcal{S}(\eta). \tag{5.8}$$

Proof. By a calculation analogous to that in the proof of Lemma 2', one finds that for sufficiently large d_1, d_2

$$\psi_k^{(-d_1-d_2)}(u)+f_k(0)\frac{u^{d_1+d_2-1}}{\Gamma(d_1+d_2)}=\frac{-\sin \pi\lambda_k'}{\pi\Gamma(\lambda_k'+d_1)}\int_0^u (u-s)^{\lambda_k'+d_1-1}y_{\nu,k}^{(-d_2)}(s+\lambda_k)\,ds. \tag{5.9}$$

Since $\psi_k(u)$ is locally analytic at $u=0$, and by means of (5.9) can be analytically continued to every point of the strip described in (i), we conclude that (i) is satisfied. Furthermore, (ii) follows from estimating (5.9) and using estimation (4.4) for $y_{\nu,k}(s+\lambda_k)$; finally (iii) is obtained by insertion of (5.9) into the right-hand side of (5.8) after partial integration (compare part β of the proof of Theorem 2).

Using the properties of $\psi_k(u)$, it follows (see [19, pp. 324–327]) that $x_{\nu,k}(z)$ can be represented as

$$x_{\nu,k}(z)=e^{\lambda_k z}z^{\lambda_k'}\Big\{f_k(0)+\sum_{l=0}^{\infty}\frac{b_k(l,w)l!}{z(z+w)\cdots(z+lw)}\Big\}, \tag{5.10}$$

and the generalized factorial series on the right converges absolutely for $\operatorname{Re}(z\,e^{i\eta})<-a$ if we take $\arg w=-\eta-\pi$, and $|w|\geqq\pi/(2\min_{j\neq k}|\operatorname{Im}(\lambda_j-\lambda_k)\,e^{-i\eta}|)$. This representation is obtained by re-expanding $\psi_k(u)=\sum_{l=0}^{\infty}b_k(l,w)w^{-l}(1-e^{wu})^l$ and integrating termwise. The coefficients $b_k(l,w)$ can either be obtained from this re-expansion or by the calculation of the formal power series expansion in z^{-1} of the factorial series. Both ways show the existence of a lower triangular (infinite) matrix $M=(m_{lj})$ $(l,j=1,2,\cdots)$ which is independent of w and relates the coefficients $b_k(l,w)$ to the coefficients $f_k(j)$ by means of (see also [19, pp. 329–330])

$$b_k(l-1,w)w^{-l}=\sum_{j=1}^{l}m_{lj}f_k(j)w^{-j},\qquad l=1,2,\cdots, \tag{5.11}$$

which shows that $b_k(l,w)$ is a polynomial in w of degree at most l and, since it can be shown that $m_{ll}=1/(l-1)!$, we see that for $w=0$ the representation (5.10) formally becomes the asymptotic expansion of $x_{\nu,k}(z)$. We formalize this result as

THEOREM 3. *Let $[A(z)]$ satisfy our natural assumptions. Then for every ν and k, $x_{\nu,k}(z)$ can be represented by means of a convergent generalized factorial series*

$$x_{\nu,k}(z)=e^{\lambda_k z}z^{\lambda_k'}\Big\{f_k(0)+\sum_{l=0}^{\infty}\frac{b_k(l,w)l!}{z(z+w)\cdots(z+lw)}\Big\}\quad \text{for } z\in\mathcal{S}(\eta),$$

with $\eta\in(\eta_{\nu+1},\eta_\nu)$, $\arg w=-\eta-\pi$, $|w|\min_{j\neq k}|\operatorname{Im}(\lambda_j-\lambda_k)\,e^{-i\eta}|\geqq\pi/2$ and $b_k(l,w)$ is a polynomial in w of degree at most l whose coefficients can be explicitly found by expanding the factorial series as an asymptotic power series in z^{-1}.

Remark 5.2. We wish to state that Theorem 3 holds equally well when some of the λ_k' are integers; this can be seen by using a scalar shift z^γ to obtain noninteger λ_k'.

Remark 5.3. Doetsch [9, pp. 386–396] has treated the Laplace integral and factorial series expansion for "general" nth order scalar differential equations having Poincaré rank one and all distinct eigenvalues. (Because the eigenvalues are distinct, it can be shown that $[A(z)]$ satisfying our natural assumptions is analytically equivalent to such a scalar differential equation, so there is no loss of generality in treating such nth order scalar differential equations.) By solving (using successive approximations) an integral equation, (scalar) functions $Y(t)=\sum_{\nu=0}^{\infty}k_\nu t^{\nu+d-1}$ with similar properties to our (vectors) $y_k(t)$ are constructed, where d corresponds to our λ_k'. In order to apply the theorem of Nörlund to obtain the factorial series expansion, it is required to have a function which is analytic at $t=0$ as well as in the semi-infinite strip, hence d should be a

positive integer. This could have been done either by a prenormalization of the differential equation using an appropriate scalar factor z^{γ}, or by transforming $Y(t)$ using a convolution of our type (5.9).

6. Invariants of the differential equation. We have shown (see [2] and [12] for the general case and [4] for the special case of distinct eigenvalues) for a differential equation satisfying our natural assumption (a_1) (but not requiring (a_2)) that a *complete system of Birkhoff invariants is given by*

$$(A_0; \Lambda'; (V_\nu), \nu = 1, 2, \cdots, m). \tag{6.1}$$

Moreover, the system of invariants (6.1) is *free* in the following sense: If A_0 is any constant matrix having n distinct eigenvalues, if Λ' is any constant diagonal matrix, and (depending upon the geometry of the eigenvalues of A_0 which determines the support of the matrices V_ν) if the entries of V_ν are arbitrarily prescribed aside from their diagonal and zero positions, then there exists a differential equation $[A(z)]$ satisfying (a_1) which has exactly these Birkhoff invariants.

In Theorem 2 (§ 5) we have shown that the connection system (V_ν) is in one-to-one correspondence with the elements of C_ν for any integer ν. Hence the connection system (V_ν) may be replaced by any single matrix C_ν to obtain the following:

THEOREM 4. *Let $[A(z)]$ be a differential equation satisfying our natural assumptions, let $H(z)$ denote any admissible formal fundamental solution matrix, let $y_k(t; \eta)$, $1 \leqq k \leqq n$, denote the associated functions corresponding to $([A(z)], H(z))$ and let $C_\nu = C(\eta)$ for any $\eta \in (\eta_{\nu+1}, \eta_\nu)$ and any integer ν (see* (1.3)*). Then the collection*

$$(A_0; \Lambda'; C_\nu) \tag{6.2}$$

forms a complete system of Birkhoff invariants for $[A(z)]$. Moreover, if A_0 is any constant matrix with all distinct eigenvalues, Λ' is any constant diagonal matrix with no integer entries, and C_ν is any constant matrix with diagonal equal to I, then there exists a differential equation $[A(z)]$ having these Birkhoff invariants.

Remark 6.1. As noted in § 4, Theorem 1 shows the Birkhoff invariance of the quantities in C_ν (for any integer ν). The completeness and freedom of the invariants (6.2), on the other hand, is shown via their one-to-one correspondence to the invariants (6.1). It appears to be easier to establish these properties working with the invariants (6.1) rather than with the associated functions.

Remark 6.2. As an application of the invariants (6.1), we showed [4] how the vanishing of certain blocks in the matrices V_ν corresponds to the reducibility of the differential equation $[A(z)]$. In particular we showed (see [4, Corollary to Thm. IV]) that $[A(z)]$ has a convergent solution vector i.e., a column of the formal series $F(z)$ in $H(z) = F(z)z^{\Lambda'} e^{\Lambda z}$ converges if and only if the corresponding columns of all the matrices in the connection system (V_ν), $\nu = 1, 2, \cdots, m$ vanish aside from the diagonal element, or equivalently, that the matrices $C_\nu^{\pm}$ both have all zero elements in the kth column aside from the diagonal element. According to Proposition 5 one sees that this is equivalent to C_ν having zero in all offdiagonal positions of its kth column. This property can also be explained directly from the structure of the associated functions, as we now see.

PROPOSITION 6. *In the situation of Theorem* 4, *let $[A(z)]$ have formal fundamental solution matrix $F(z)z^{\Lambda'} e^{\Lambda z}$ and let C_ν be the matrix of constants corresponding to the associated functions for any integer ν. Then the* kth *column of $F(z)$ converges for $|z|$ sufficiently large if and only if the* kth *column of C_ν vanishes aside from the diagonal element.*

Proof. Consider the associated function defined locally by

$$y_k(t) = \sum_0^\infty f_k(\nu)\Gamma(\lambda_k' + 1 - \nu)(t - \lambda_k)^{\nu - \lambda_k' - 1}.$$

If the formal series $\sum_0^\infty f_k(\nu)z^{-\nu}$ converges for $|z|$ sufficiently large, then

$$(t-\lambda_k)^{\lambda_k'+1}y_k(t) \tag{6.3}$$

is an entire function; hence the associated constants $c_{jk} = 0$ for all $j \neq k$.

Conversely, if $c_{jk} = 0$ for all $j \neq k$, then (6.3) represents an entire function. The associated functions corresponding to standard differential equations have polynomial growth at ∞ hence for a standard differential equation (6.3) must be a polynomial, i.e., the formal series $\sum_0^\infty f_k(\nu)z^{-\nu}$ is a polynomial in z^{+1}. Using the result of Birkhoff–Turrittin (see § 4.1), the kth column of each formal series corresponding to a general differential equation is the product of the meromorphic transformation at ∞ times a column vector which is a polynomial in z^{-1}, hence the kth column of $F(z)$ converges.

Postscript. After this paper was finished we obtained a copy of the dissertation of R. Schäfke [20] where he investigates connections between the solutions of the standard equation (0.1) and those of

$$\frac{dy}{dt} = (\Lambda - tI)^{-1}(A_1 - \rho I)y, \tag{*}$$

ρ being a complex parameter. Apart from normalizations, he also obtained the connection formula between Y and Y^* and various continuation formulas for Y^* and Y (cf. §§ 1, 2). Furthermore, he discussed how the Stokes' multipliers V_ν are related to the connection coefficients C_ν under the additional assumption that no three λ_j lie on the line. (Compare this with the general case treated in Proposition 5, Theorem 2.) Our discussion of the special differential equation (1.1) and the construction of Y^* should be considered a preliminary step for the treatment of the general equation (0.3) and its associated functions. Our proofs differ in that they extend immediately to the general case, and the assumptions we make are necessary and sufficient for our results (cf. Propositions 1, 2). On the other hand, (*) is somewhat more general than (1.1) and Schäfke has various interesting results concerning the dependency upon ρ.

REFERENCES

[1] W. BALSER, W. B. JURKAT AND D. A. LUTZ, *A general theory of invariants for meromorphic differential equations: Part I, Formal invariants*, Funkcialaj Ekvacioj, 22 (1979), pp. 197–221.

[2] ———, *A general theory of invariants for meromorphic differential equations: Part II, Proper invariants*, Funkcialaj Ekvacioj, 22 (1979), pp. 257–283.

[3] ———, *A general theory of invariants for meromorphic differential equations: Part III, Applications*, Houston J. Math., 6 (1980), pp. 149–189.

[4] ———, *Birkhoff invariants and Stokes' multipliers for meromorphic linear differential equations*, J. Math. Anal. Appl., 71 (1979), pp. 48–94.

[5] G. D. BIRKHOFF, *Singular points of ordinary linear differential equations*, Trans. Amer. Math. Soc. 10 (1909), pp. 436–470.

[6] ———, *The generalized Riemann problem for linear differential equations and the allied problems for linear difference and q-difference equations*, Proc. Amer. Acad. Arts and Sci., 49 (1913), pp. 531–568.

[7] ———, *On a simple type of irregular singular point*, Trans. Amer. Math. Soc., 14 (1913), pp. 462–476.

[8, 9] G. DOETSCH, *Handbuch der Laplace-Transformation*, vol. I, II, Birkhäuser-Verlag, Basel, 1972.

[10] J. HORN, *Integration linearer Differentialgleichungen durch Laplacesche Integrale und Fakultätenreihen*, Jahresber. d. Deutsch. Math. Ver., 24 (1915), pp. 309–329.

[11] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956, Chapter XIX.
[12] W. B. JURKAT, *Meromorphe Differentialgleichungen*, Lecture Notes in Mathematics, 637, Springer-Verlag, Berlin–Heidelberg–New York, 1978.
[13] W. B. JURKAT, D. A. LUTZ AND A. PEYERIMHOFF, *Birkhoff invariants and effective calculations for meromorphic linear differential equations*; I, J. Math. Anal. Appl., 53 (1976), pp. 438–470.
[14] K. OKUBO, *A global representation of a fundamental set of solutions and a Stokes' phenomenon for a system of linear ordinary differential equations*, J. Math. Soc. Japan, 15 (1963), pp. 268–288.
[15] ———, *Connection problems for systems of linear differential equations*, in Japan–U.S. Seminar on Ordinary Differential and Functional Equations, Lecture Notes in Mathematics, 243, Springer-Verlag, Berlin–Heidelberg–New York, 1971.
[16] W. J. TRJITZINSKY, *Laplace integrals and factorial series in the theory of linear differential and linear difference equations*, Trans. Amer. Math. Soc., 37 (1935), pp. 80–146.
[17] H. L. TURRITTIN, *Convergent solutions of ordinary linear homogeneous differential equations in the neighborhood of an irregular singular point*, Acta Math. 93 (1955), pp. 27–66.
[18] ———, *Reduction of ordinary differential equations to the Birkhoff canonical form*, Trans. Amer. Math. Soc., 107 (1963), pp. 485–507.
[19] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.
[20] R. SCHÄFKE, *Über das globale analytische Verhalten der Lösungen der über die Laplacetransformation zusammenhängenden Differentialgleichungen* $tx' = (A + tB)x$ *und* $(s - B)v' = (\rho - A)v$, Doctoral Dissertation, University of Essen, West Germany, April 1979.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0005 $01.00/0

VECTOR LIFTING AND FACTORABLE DIFFERENTIAL OPERATORS*

L. R. BRAGG†

Abstract. Let $\Phi(\alpha)$ be a $1 \times n$ vector function of a real parameter with components $\phi_i(\alpha) \in C^{n-1}$ in α and let $\Psi(\alpha) = e^{ED_\alpha}\Phi(\alpha) = \sum_{j=0}^{n-1} (j!)^{-1} E^j D_\alpha^j \Phi(\alpha)$, in which E denotes the $n \times n$ matrix with 1's on the superdiagonal and 0's elsewhere. We say that the operator e^{ED_α} lifts the vector $\Phi(\alpha)$ into the vector $\Psi(\alpha)$ with respect to α. If the components of $\Phi(\alpha)$ are nontrivial and independent, solutions of some linear differential equation involving the parameter α, then the first component of $\Psi(\alpha)$ is the solution of an associated iterated or factorable differential equation. Vector liftings with respect to parameters (and operators) are employed throughout this paper to construct solution representations for a variety of factorable linear abstract and partial differential equations.

1. Introduction. Iterated differential operators appear frequently in the theory and application of partial differential equations. Two of the more familiar examples of equations having repeated operators are the biharmonic equation $\Delta_n^2 u = 0$ and its polyharmonic generalization $\Delta_n^p u = 0$, $p \geqq 3$, in which Δ_n denotes the Laplacian operator in n variables. Extensive studies of the structure of solutions of these equations were carried out by M. Picone [9]. In his investigations of the solution structure of the Cauchy problem for the singular (or exceptional) Euler–Poisson–Darboux (EPD) equation, A. Weinstein [10] established the basic role assumed by polyharmonic data. He also made systematic investigations into the solution structure of the iterated equation of generalized axially symmetric potential theory (GASPT) [11] and, later, the iterated wave and EPD equations [12]. In this, Weinstein noted that if L_k is an operator of the form $L_k = \Delta_n + (D_y^2 + k/yD_y)$ (or $\Delta_n - (D_y^2 + k/yD_y)$), with k real, then the general solution of $L_\alpha \cdot L_\beta W = 0$, $\beta \neq \alpha - 2$, is given by $W = U^{(\beta)} + U^{(\alpha-2)}$, in which $U^{(k)}$ denotes a general solution of $L_k U = 0$. The reader is referred to [12] for further details about the method employed along with pertinent references. J. C. Burns [5], [6] has also treated questions that pertain to the iterated GASPT equation.

In this paper, we give an alternative method for constructing solutions of linear evolution type problems with underlying equations that involve iterated differential operators and other types of products of differential operators. This approach draws upon notions about functions of matrices [8] and their uses in solving systems of ordinary differential equations. We start with a solvable "scalar" equation (heat, wave, etc.) that involves one or more basic operators and/or parameters. A matrix translation operator is then employed to "lift" a vector of solutions associated with the scalar equation, into a second vector that satisfies a matrix-vector equation which involves a Jordan block. The first component of this lifted vector then satisfies a scalar equation that involves an iterated differential operator or a product of differential operators. Initial data for a variety of these factorable equations can be incorporated into the vector to be lifted. Depending upon the circumstances, the lifting operation is carried out with respect to either a parameter or an operator. For the purposes of this paper, lifting with respect to an operator will be handled in precisely the same way as lifting with respect to a parameter. A completely rigorous treatment of operator lifting would appear to require further developments in the perturbation theory of operators. Nevertheless, the results obtained by the formal procedure will be seen to be valid.

In § 2, we present the essential features associated with the lifting operation. We illustrate this approach through the treatment of a problem involving an abstract heat

* Received by the editors September 24, 1980.
† Department of Mathematical Sciences, Oakland University, Rochester, Michigan 48063.

equation. In this, the lifting is carried out on a vector determined by a semi-group of operators with respect to the infinitesimal generator of that semi-group. Sections 3 and 4 will be concerned with problems involving abstract versions of the iterated EPD and GASPT equations. For the iterated heat and EPD problems, we relate the initial data to the entries in the vectors to be lifted. In the case of the GASPT equation, we apply a transmutation operator that relates solutions of the GASPT equation to solutions of the heat equation. The remaining two sections treat examples in which the lifting is carried out with respect to a real parameter. The first of these involves the radial heat equation and the second involves the EPD equation. Questions on lifting that require further developments in operator theory will be noted.

2. Lifting of vectors. Let M_n denote the set of $n \times n$ matrices over the reals with identity E_0 and let E_1 be the matrix with 1's on the superdiagonal and 0's elsewhere. Let $E_j = E_1^j$, $j = 0, 1, 2, \cdots$. Then $E_i \cdot E_j = E_{i+j}$ and $E_j = \mathbf{0}$, the zero matrix, if $j \geqq n$. It is clear that if P and Q are any two matrices of the form $\sum_{j=0}^{n-1} \alpha_j E_j$ with α_j real, then $P \cdot Q = Q \cdot P$.

Next, let $\Phi(x)$ be a $1 \times n$ vector function in which all components $\in C^{n-1}$. Then we have the following symbolic version for Taylor's expansion:

$$e^{yE_1 D_x} \cdot \Phi(x) = \sum_{j=0}^{n-1} \frac{y^j E_j}{j!} D_x^j \Phi(x). \tag{2.1}$$

If $\Phi(x)$ has analytic components and y is real, it follows that

$$\sum_{j=0}^{n-1} \frac{y^j E_j}{j!} D_x^j \Phi(x) = \Phi(xE_0 + yE_1); \tag{2.2}$$

and that the right-hand side of (2.2) is simply the vector function Φ with matrix argument xE_0 translated by yE_1. Let $F(x, y)$ denote the vector function appearing in the right member of (2.1). We refer to (2.1) by saying that *the operator* $e^{yE_1 D_x}$ *or* $\sum_{j=0}^{n-1} (y^j E_j / j!) D_x^j$ *lifts the vector* $\Phi(x)$ *into the vector* $F(x, y)$ *with respect to* x. We refer to $F(x, y)$ as the *lifted vector.* In this lifting, there are a variety of possible choices for y, the choice being dependent upon the type of problem under consideration. One can similarly define lifting operators $e^{yE_k D_x}$, $k = 2, \cdots, n-1$, by

$$e^{yE_k D_x} \Phi(x) = \sum_{j=0}^{[(n-1)/k]} \frac{y^j E_j \cdot k}{j!} D_x^j \Phi(x), \tag{2.3}$$

but we will not use them in the discussions to follow.

We now make use of the lifting procedure associated with (2.1) in connection with an abstract heat problem. For this purpose, let X be a Banach space and let A be the infinitesimal generator of a holomorphic semigroup in X. We assume that the domain $\mathcal{D}(A^r)$ is dense in X for r an arbitrarily large positive integer. Then the "scalar" version of the abstract heat problem is given by

$$u'(t) - Au(t) = 0, \qquad t > 0, \tag{2.4a}$$

$$u(0+) = \phi, \qquad \phi \in \mathcal{D}(A^r), \tag{2.4b}$$

in which (2.4b) is understood to mean that $\|u(t) - \phi\| \xrightarrow[t \to 0+]{} 0$. The solution of this problem can be written in the form

$$u(t) = T_A(t) \cdot \phi, \tag{2.5}$$

in which $T_A(t)$ is the semigroup of operators generated by A. Next, let $\phi_k \in \mathcal{D}(A^r)$,

$k = 1, 2, \cdots, n$, and let Φ denote the $1 \times n$ vector with kth component ϕ_k. Then the vector $U(t) = T_A(t) \cdot \Phi = (T_A(t)\varphi_k)$, is the solution of the vector problem

$$U'(t) - AU(t) = 0, \qquad t > 0, \tag{2.6a}$$

$$U(0+) = \Phi, \tag{2.6b}$$

where (2.6b) means that $\max_{1 \leqq k \leqq n} \|U_k(t) - \phi_k\| \xrightarrow[t \to 0+]{} 0$ with $U_k(t)$ the kth component of $U(t)(U_k(t) = T_A(t)\phi_k)$.

To apply the lifting procedure with respect to A, we make use of the relation

$$D_A T_A(t)\phi = tT_A(t)\phi, \qquad \phi \in \mathcal{D}(A^r), \tag{2.7}$$

which is an operator generalization of $D_x e^{xt} = te^{xt}$ for x, a scalar parameter. Now, we define the lifted vector $\mathcal{U}(t)$ of $U(t)$ by

$$\begin{aligned} \mathcal{U}(t) &= e^{E_1 D_A} \cdot U(t) = \sum_{j=0}^{n-1} \frac{E_j}{j!} D_A^j U(t) \\ &= \sum_{j=0}^{n-1} \frac{t^j E_j U(t)}{j!}, \end{aligned} \tag{2.8}$$

the last equality following by repeated applications of (2.7). This lifted vector $\mathcal{U}(t)$ is a solution of the matrix-vector problem

$$\mathcal{U}'(t) - (AE_0 + E_1)\mathcal{U}(t) = \mathbf{0}, \qquad t > 0, \tag{2.9a}$$

$$\mathcal{U}(0+) = \Phi. \tag{2.9b}$$

Moreover, the first component $\mathcal{U}_1(t)$ of $\mathcal{U}(t)$ is given by

$$\mathcal{U}_1(t) = \sum_{k=1}^{n} \frac{t^{k-1}}{(k-1)!} T_A(t)\phi_k, \tag{2.10}$$

which is clearly a solution of the iterated abstract heat equation

$$(D_t - A)^n \cdot Z(t) = 0, \qquad t > 0. \tag{2.11}$$

If a solution $Z(t)$ of (2.11) is required to satisfy the conditions $Z^{(k)}(0+) = \psi_k$, $k = 0, 1, \cdots, n-1$ with $\psi_k \in D(A^r)$, then the reader can verify that the ϕ_k in (2.6) can be selected as follows:

$$\phi_{k+1} = \sum_{j=0}^{k} (-1)^j \binom{k}{j} A^j \psi_{k-j}, \qquad k = 0, 1, \cdots, n-1. \tag{2.12}$$

Remark. The relation (2.7) can be motivated as follows. Let λ be real and let I be an identity operator in X. Then

$$\begin{aligned} D_A T_A(t)\phi &= \lim_{\lambda \to 0+} \frac{T_{A+\lambda I}(t)\phi - T_A(t)\phi}{\lambda} \\ &= \lim_{\lambda \to 0+} \frac{e^{\lambda t} T_A(t)\phi - T_A(t)\phi}{\lambda} \\ &= \lim_{\lambda \to 0+} \frac{e^{\lambda t} - 1}{\lambda} T_A(t)\phi = tT_A(t)\phi. \end{aligned}$$

The literature on perturbations of operators does not appear to provide a general development of (2.7). Further research on operators is indicated for derivations of the type used in this and the following section.

3. The iterated EPD problem. Using the Banach space setting of § 2, suppose that $A = B^2$, in which B is the generator of a continuous group in X. If $a \geqq 0$ and $\phi \in \mathcal{D}(A^r)$, then the solution $w^a(t, \phi)$ of the following scalar version of the abstract EPD problem

$$w''(t) + \frac{a}{t} w'(t) = Aw(t), \qquad t > 0, \tag{3.1a}$$

$$\|w(t) - \phi\| \xrightarrow[t \to 0+]{} 0, \qquad \|w'(t)\| \xrightarrow[t \to 0+]{} 0, \tag{3.1b}$$

can be written in the symbolic form

$$w^a(t, \phi) = F(a, A, t) \cdot \phi, \tag{3.2}$$

in which the operator $F(a, A, t)$ has the formal representation

$$F(a, A, t) = 2^{(a-1)/2} \Gamma\left(\frac{a+1}{2}\right) (tA^{\frac{1}{2}})^{(1-a)/2} I_{(a-1)/2}(tA^{\frac{1}{2}}), \tag{3.3}$$

with I_ν denoting one of the modified Bessel functions [4, p. 264]. As in § 2, let $\phi, \cdots, \phi_n \in \mathcal{D}(A^r)$ and let Φ be the $1 \times n$ vector with kth component ϕ_k. Then the vector problem

$$\tilde{W}''(t) + \frac{a}{t} \tilde{W}'(t) = A\tilde{W}(t), \qquad t > 0, \tag{3.4a}$$

$$\operatorname*{Max}_{1 \leqq k \leqq n} \|\tilde{W}_k(t) - \phi_k\| \xrightarrow[t \to 0+]{} 0, \qquad \operatorname*{Max}_{1 \leqq k \leqq n} \|\tilde{W}'_k(t)\| \xrightarrow[t \to 0+]{} 0, \tag{3.4b}$$

has the solution

$$W(t, \Phi) = F(a, A, t) \cdot \Phi = (F(a, A, t)\phi_k) = (w^a(t, \varphi_k)). \tag{3.5}$$

The lifted matrix-vector problem

$$\begin{gathered} \mathcal{W}''(t) + \frac{a}{t} \mathcal{W}'(t) = (AE_0 + E_1)\mathcal{W}(t), \qquad t > 0, \\ \operatorname*{Max}_{1 \leqq k \leqq n} \|\mathcal{W}_k(t) - \varphi_k\| \xrightarrow[t \to 0+]{} 0, \qquad \operatorname*{Max}_{1 \leqq k \leqq n} \|\mathcal{W}'_k(t)\| \xrightarrow[t \to 0+]{} 0, \end{gathered} \tag{3.6}$$

with $\mathcal{W}_k(t)$ the kth component of $\mathcal{W}(t)$, has its solution vector given by

$$\begin{aligned} \mathcal{W}(t, \Phi) &= e^{E_1 D_A} \tilde{W}(t, \Phi) \\ &= \sum_{j=0}^{n-1} \frac{E_j}{j!} D_A^j F(a, A, t)\Phi. \end{aligned} \tag{3.7}$$

But from the fact that $(z^{-1}D_z)^m\{z^{-\upsilon} I_\upsilon(z)\} = z^{-(\upsilon+m)} I_{\upsilon+m}(z)$ [7, p. 67], it follows formally that

$$D_A^j F(a, A, t) = \frac{\Gamma\left(\frac{a+1}{2}\right) t^{2j}}{2^{2j} \Gamma\left(j + \frac{a+1}{2}\right)} F(a + 2j, A, t). \tag{3.8}$$

Using this in the right-hand side of (3.7), we get

$$\mathcal{W}(t,\Phi)=\sum_{j=0}^{n-1}\frac{\Gamma\left(\frac{a+1}{2}\right)t^{2j}}{2^{2j}j!\,\Gamma\left(\frac{a+1}{2}+j\right)}E_j\cdot F(a+2j,A,t)\Phi$$

$$=\sum_{j=0}^{n-1}\frac{\Gamma\left(\frac{a+1}{2}\right)t^{2j}}{2^{2j}j!\,\Gamma\left(\frac{a+1}{2}+j\right)}E_j(w^{a+2j}(t,\varphi_k)). \tag{3.9}$$

The first component $\mathcal{W}_1(t)$ of the vector $\mathcal{W}(t,\Phi)$ is given by

$$\mathcal{W}_1(t)=\sum_{j=0}^{n-1}\frac{\Gamma\left(\frac{a+1}{2}\right)}{2^{2j}j!\,\Gamma\left(\frac{a+1}{2}+j\right)}t^{2j}w^{a+2j}(t,\phi_{j+1}) \tag{3.10}$$

and clearly satisfies the iterated abstract EPD equation

$$\left(D_t^2+\frac{a}{t}D_t-A\right)^n\cdot\mathcal{W}^*(t)=0,\qquad t>0. \tag{3.11}$$

For r sufficiently large, Taylor's expansion yields

$$w^k(t,\varphi)=\varphi+\frac{1}{2(k+1)}t^2A\varphi+\frac{1}{8(k+1)(k+3)}t^4A^2\varphi$$

$$+\frac{1}{2\cdot4\cdot8(k+1)(k+3)(k+5)}t^6A^3\varphi+\cdots.$$

Using this in (3.10) with the various choices for k there, one can show that if the solution of (3.11) is required to satisfy the conditions

$$\mathcal{W}^{*(2j)}(0+)=\psi_j,\qquad j=0,1,\cdots,n-1,\quad \psi_j\in\mathcal{D}(A^r),$$
$$\mathcal{W}^{*(2j+1)}(0+)=0,\qquad j=0,1,\cdots,n-1, \tag{3.12}$$

then the components of Φ are given in terms of the ψ_j by

$$\phi_{k+1}=\frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left[\prod_{l=1}^{k-j}\left(\frac{a+2l-1}{2l-1}\right)\right]A^j\cdot\psi_{k-j}, \tag{3.13}$$

in which the product is assigned the value 1 when $j=k$.

4. The iterated GASPT equation. Taking $A=B^2$ as in § 3, it follows that if $\phi\in\mathcal{D}(A^r)$ and $a<1$, then the solution of the scalar abstract GASPT problem

$$\left(D_y^2+\frac{a}{y}D_y+A\right)v(y)=0,\qquad y>0,$$
$$\|v(y)-\phi\|\xrightarrow[y\to0+]{}0, \tag{4.1}$$

is given by the transmutation

$$v^a(y,\varphi)=\frac{y^{1-a}}{\Gamma\left(\frac{1-a}{2}\right)}\int_0^\infty e^{-\sigma y^2}\sigma^{-(a+1)/2}\left\{T_A\left(\frac{1}{4\sigma}\right)\varphi\right\}d\sigma \tag{4.2}$$

[2, p. 333]. Then a solution of the lifted matrix-vector problem

$$\begin{aligned}&\left(D_y^2+\frac{a}{y}D_y\right)\mathcal{V}(y)+(AE_0+E_1)\mathcal{V}(y)=0, \qquad y>0,\\ &\mathcal{V}(0+)=\Phi=(\phi_k), \qquad \phi_k\in\mathcal{D}(A^r), \quad k=1,2,\cdots,n\end{aligned} \tag{4.3}$$

can be obtained from the vector solution (2.9) via (4.2). Thus

$$\begin{aligned}\mathcal{V}(y)&=\sum_{j=0}^{n-1}\frac{E_j}{2^{2j}j!}\frac{y^{1-a}}{\Gamma\left(\frac{1-a}{2}\right)}\int_0^\infty e^{-\sigma y^2}\sigma^{-(a+1)/2-j}\left(T_A\left(\frac{1}{4\sigma}\right)\Phi\right)d\sigma\\ &=\sum_{j=0}^{n-1}\frac{y^{2j}}{2^{2j}j!}\frac{\Gamma\left(\frac{1-a-2j}{2}\right)}{\Gamma\left(\frac{1-a}{2}\right)}E_jV^{a+2j}(y,\Phi),\end{aligned} \tag{4.4}$$

in which $V^{a+2j}(y,\Phi)$ is the vector function obtained by replacing ϕ in (4.2) by Φ and a by $a+2j$. Reading off the first component of $\mathcal{V}(y)$ from (4.4), we see that the iterated equation $(D_y^2+a/yD_y+A)^nZ(y)=0$ is satisfied by

$$Z(y)=\sum_{j=0}^{n-1}\frac{\Gamma\left(\frac{1-a-2j}{2}\right)}{\Gamma\left(\frac{1-a}{2}\right)}\frac{y^{2j}}{2^{2j}j!}v^{a+2j}(y,\phi_{j+1}),$$

provided that $a+2n-2<1$.

5. The radial heat equation. We now consider the radial heat problem

$$\begin{aligned}&u_t(r,t)=u_{rr}(r,t)+\frac{2a+1}{r}u_r(r,t), \qquad t>0, \quad a>-\tfrac{1}{2},\\ &u(r,0+)=\phi(r),\end{aligned} \tag{5.1}$$

with $\phi(r)$ bounded and continuous. We next carry out a lifting on a 2-vector associated with (5.1) corresponding to the choice $a=0$.

The scalar problem (5.1) has the solution

$$u(r,t)=\int_0^\infty P_a(r,\xi,t)\phi(\xi)\,d\xi, \tag{5.2}$$

in which

$$P_a(r,\xi,t)=\frac{1}{2t}\xi\,e^{-(r^2+\xi^2)/4t}\left\{\left(\frac{\xi}{r}\right)^a I_a\left(\frac{r\xi}{2t}\right)\right\}, \tag{5.3}$$

with I_a a modified Bessel function see [1]. (The notation here is slightly different to avoid confusion with one of the types of Bessel functions.) Now, let $U(r,t)$ be the

solution of (5.1) with the scalar data function $\varphi(r)$ replaced by the 2-vector $\Phi(r) = (\phi_k(r))$. Let

$$\mathcal{U}(r, t) = e^{E_1 D_a/2} U(r, t). \tag{5.4}$$

Then $\mathcal{U}(r, t)$ is a solution of the lifted matrix-vector problem

$$\begin{aligned} \mathcal{U}_t(r, t) &= D_r^2 \mathcal{U}(r, t) + \frac{1}{r}\{(2a+1)E_0 + E_1\} D_r \mathcal{U}(r, t) \\ \mathcal{U}(r, 0) &= \Phi(r). \end{aligned} \tag{5.5}$$

From (5.3) and [7, p. 71], it follows that

$$D_a\left\{\left(\frac{\xi}{r}\right)^a I_a\left(\frac{r\xi}{2t}\right)\right\}_{a=0} = I_0\left(\frac{r\xi}{2t}\right) \ln\left(\frac{\xi}{r}\right) - K_0\left(\frac{r\xi}{2t}\right), \tag{5.6}$$

in which K_0 denotes one of the modified Bessel functions. Applying (2.1) in (5.4), it follows from (5.6) that the solution of (5.5) that corresponds to $a = 0$ is given by

$$\begin{aligned} \mathcal{U}(r, t) = &\int_0^\infty P_0(r, \xi, t)\Phi(\xi)\, d\xi \\ &+ \frac{1}{2} E_1 \int_0^\infty P_0(r, \xi, t) \ln\left(\frac{\xi}{t}\right)\Phi(\xi)\, d\xi \\ &- \frac{1}{2} E_1 \int_0^\infty \frac{1}{2t} e^{-(r^2+\xi^2)/4t} \cdot \xi K_0\left(\frac{r\xi}{2t}\right)\Phi(\xi)\, d\xi. \end{aligned} \tag{5.7}$$

The first component $\mathcal{U}_1(r, t)$ of $\mathcal{U}(r, t)$ is given by

$$\begin{aligned} \mathcal{U}_1(r, t) = &\int_0^\infty P_0(r, \xi, t)\left\{\phi_1(\xi) + \frac{1}{2}\phi_2(\xi) \ln\left(\frac{\xi}{r}\right)\right\} d\xi \\ &- \frac{1}{4}\int_0^\infty \frac{1}{t} e^{-(r^2+\xi^2)/4t} \cdot \xi K_0\left(\frac{r\xi}{2t}\right)\phi_2(\xi)\, d\xi, \qquad t > 0. \end{aligned} \tag{5.8}$$

It is not difficult to show, from (5.5) with $a = 0$, that $\mathcal{U}_1(r, t)$ satisfies the equation

$$\left(D_t - D_r^2 - \frac{3}{r} D_r\right) \cdot \left(D_t - D_r^2 - \frac{1}{r} D_r\right) \mathcal{U}_1(r, t) = 0, \qquad t > 0. \tag{5.9}$$

6. The EPD equation revisited. In § 3, we carried out a lifting with respect to the operator A in a vector problem associated with (3.1). We now carry out a lifting with respect to the parameter a for an analogous vector problem for the case $n = 2$. Although we can make use of the formal operator (3.3) for this, we take a slightly different but, nevertheless, formal approach.

With the change of variables $\xi = t^2/4$, the problem (3.1) transforms into the hypergeometric problem

$$\left[D_\xi\left(\xi D_\xi + \frac{a-1}{2}\right) - A\right]\tilde{u}(\xi) = 0, \qquad \tilde{u}(0+) = \phi, \tag{6.1}$$

where $\tilde{u}(\xi) = u(2\sqrt{\xi})$. The solution of (6.1) is given formally in terms of a hypergeometric operator by

$$\tilde{u}(\xi, \phi) = {}_0F_1\left(\cdot; \frac{a+1}{2}; \xi A\right) \cdot \phi. \tag{6.2}$$

As before, we consider the vector problem obtained by replacing ϕ in (6.1) by the vector

$$\Phi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \qquad \varphi_i \in \mathcal{D}(A^r).$$

If we denote its solution vector by $\tilde{U}(\xi)$, then the components of $\tilde{U}(\xi)$ are $\tilde{u}_i(\xi) = \tilde{u}(\xi, \phi_i)$, $i = 1, 2$. Now, let $\tilde{\mathcal{U}}(\xi)$ denote the vector solution of the following lifted problem (with respect to a)

$$\begin{aligned} &\left[D_\xi\left(\left(\xi D_\xi + \frac{a-1}{2}\right)E_0 + \frac{1}{2}E_1\right) - A\right]\tilde{\mathcal{U}}(\xi) = 0, \\ &\tilde{\mathcal{U}}(0+) = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}. \end{aligned} \tag{6.3}$$

This solution function is given by

$$\begin{aligned} \tilde{\mathcal{U}}(\xi) &= e^{E_1 D_a}\left\{{}_0F_1\left(\cdot; \frac{a+1}{2}; \xi A\right)\cdot\Phi\right\} \\ &= (E_0 + E_1 D_a)\,{}_0F_1\left(\cdot; \frac{a+1}{2}; \xi A\right)\Phi. \end{aligned} \tag{6.4}$$

Since

$$\frac{\partial}{\partial a}\frac{1}{\left(\frac{a+1}{2}\right)_j} = -\frac{1}{\left(\frac{a+1}{2}\right)_j}\sum_{k=1}^{j}\frac{1}{(a+2k-1)},$$

and

$$\sum_{k=1}^{j}\frac{1}{(a+2k-1)} = \int_0^1 \frac{\sigma^a(1-\sigma^{2j})}{1-\sigma^2}\,d\sigma,$$

it follows formally that

$$\begin{aligned} &D_A\,{}_0F_1\left(\cdot; \frac{a+1}{2}; \xi A\right)\Phi \\ &= -\int_0^1 \frac{\sigma^{a-1}}{1-\sigma^2}\left[{}_0F_1\left(\cdot; \frac{a+1}{2}; \xi A\right)\varphi - {}_0F_1\left(\cdot; \frac{a+1}{2}; \sigma^2\xi A\right)\Phi\right]d\sigma. \end{aligned} \tag{6.5}$$

Using this in (6.4) and taking the first component $\tilde{\mathcal{U}}_1(\xi)$ of $\tilde{\mathcal{U}}(\xi)$, we get

$$\tilde{\mathcal{U}}_1(\xi) = \tilde{u}_1(\xi) - \int_0^1 \frac{\sigma^a}{1-\sigma^2}[\tilde{u}_2(\xi) - \tilde{u}_2(\sigma^2\xi)]\,d\sigma. \tag{6.6}$$

Splitting (6.3) into component equations, one can show that $\tilde{\mathcal{U}}_1(\xi)$ satisfies the factored equation

$$\left[D_\xi\left(\xi D_\xi + \frac{a+1}{2}\right) - A\right]\left[D_\xi\left(\xi D_\xi + \frac{a-1}{2}\right) - A\right]\tilde{\mathcal{U}}_1(\xi) = 0. \tag{6.7}$$

If we return to the original variable t, it follows that the function

$$v(t) = \tilde{u}_1\left(\frac{t^2}{4}\right) - \int_0^1 \frac{\sigma^a}{1-\sigma^2}\left\{\tilde{u}_2\left(\frac{t^2}{4}\right) - \tilde{u}_2\left(\frac{t^2\sigma^2}{4}\right)\right\}d\sigma, \tag{6.8}$$

is a solution of the factored equation

$$\left(D_t^2+\frac{a+2}{t}D_t-A\right)\left(D_t^2+\frac{a}{t}D_t-A\right)v(t)=0. \tag{6.9}$$

REFERENCES

[1] L. R. BRAGG, *The radial heat polynomials and related functions*, Trans. Amer. Math. Soc., 119 (1965), pp. 270–290.

[2] ———, *Hypergeometric operator series and related partial differential equations*, Trans. Amer. Math. Sci., 143 (1969), pp. 319–336.

[3] ———, *Related non-homogeneous partial differential equations*, J. Appl. Anal., 4 (1974), pp. 161–189.

[4] L. R. BRAGG AND J. W. DETTMAN, *An operator calculus for related partial differential equations*, J. Math. Anal. Appl., 22 (1968), pp. 261–271.

[5] J. C. BURNS, *The iterated equation of generalized axially symmetric potential theory II, general solutions of Weinstein's type*, J. Austral. Math. Soc., 7 (1967), pp. 277–289.

[6] ———, *The iterated equation of generalized axially symmetric potential theory V, generalized Weinstein correspondence principle*, J. Austral. Math. Soc., 11 (1970), pp. 129–141.

[7] W. MAGNUS, F. OBERHETTINGER, AND R. SONI, *Formulas and Theorems For Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[8] L. MIRSKY, *An Introduction to Linear Algebra*, Clarendon Press, Oxford, 1955.

[9] M. PICONE, *Nuovi indirizzi di ricerca nella teoria e nel calcolo della soluzioni di talune equazioni lineari alli derivate parziali della Fisica-mathematica*, Ann. Scoula Norm. Sup. Pisa, Serie II, V, (1936).

[10] A. WEINSTEIN, *Sur le problème de Cauchy pour l'équation de Poisson et l'équation des ondes*, C.R. Acad. Sci. Paris, 234 (1952), pp. 2584–2585.

[11] ———, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–28.

[12] ———, *On a class of partial differential equations of even order*, Ann. mat. pura appl., IV (1955), pp. 245–254.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0006 $01.00/0

IMPLICIT DEGENERATE EVOLUTION EQUATIONS AND APPLICATIONS*

EMMANUELE DI BENEDETTO† AND R. E. SHOWALTER‡

Abstract. The initial-value problem is studied for evolution equations in Hilbert space of the general form

$$\frac{d}{dt}\mathcal{A}(u)+\mathcal{B}(u)\ni f,$$

where $\mathcal{A}$ and $\mathcal{B}$ are maximal monotone operators. Existence of a solution is proved when $\mathcal{A}$ is a subgradient and either $\mathcal{A}$ is strongly monotone or $\mathcal{B}$ is coercive; existence is established also in the case where $\mathcal{A}$ is strongly monotone and $\mathcal{B}$ is subgradient. Uniqueness is proved when one of $\mathcal{A}$ or $\mathcal{B}$ is continuous self-adjoint and the sum is strictly monotone; examples of nonuniqueness are given. Applications are indicated for various classes of degenerate nonlinear partial differential equations or systems of mixed elliptic-parabolic-pseudo-parabolic types and problems with nonlocal nonlinearity.

1. Introduction. Let $\mathcal{A}$ and $\mathcal{B}$ be maximal monotone operators from a Hilbert space V to its dual V^*. Such operators are in general multi-valued and their basic properties will be recalled below. We shall consider initial-value problems of the form

$$\frac{d}{dt}\mathcal{A}(u)+\mathcal{B}(u)\ni f, \qquad \mathcal{A}u(0)\ni v_0, \tag{1.1}$$

where $f\in L^2(0, T; V^*)$ and $v_0\in V^*$ are given. It is assumed throughout our work that $\mathcal{A}$ is a compact operator from V to V^*. In applications to partial differential equations this assumption limits the order of the operator $\mathcal{A}$ to be strictly lower than that of $\mathcal{B}$. Both operators will be required to satisfy boundedness conditions, and one or the other is assumed to be a subgradient.

The objective of this work is to prove existence of a solution of (1.1) when $\mathcal{A}$ and $\mathcal{B}$ are possibly degenerate. Observe that we must in general assume some condition of coercivity on the pair of operators. To see this, we note that if one of them is identically zero then (1.1) is equivalent to a one-parameter family of "stationary" problems of the form $M(u(t))\ni F(t)$, where M is maximal monotone. But if M is, e.g., a subgradient in a space of finite dimension, it is surjective only if it is coercive. Thus it is appropriate to assume that at least one of $\mathcal{A}$ or $\mathcal{B}$ is coercive. In accord with this remark our work will proceed as follows. First we replace $\mathcal{A}$ by the coercive operator $\mathcal{A}+\varepsilon\mathcal{R}$, where $\varepsilon>0$ and $\mathcal{R}: V\to V^*$ is the Riesz isomorphism determined by the scalar product on V, and we solve the initial-value problem for the "regularized" equation

$$\frac{d}{dt}(\mathcal{A}+\varepsilon\mathcal{R})(u_\varepsilon)+\mathcal{B}(u_\varepsilon)\ni f. \tag{1.2}$$

Here we may take $\varepsilon=1$ with no loss of generality and we make no coercivity assumptions on either $\mathcal{A}$ or $\mathcal{B}$. Next we assume $\mathcal{B}$ is coercive and let $\varepsilon\to 0^+$ in order to recover (1.1) with (possibly) degenerate $\mathcal{A}$. Since $\mathcal{R}$ is of the same order as $\mathcal{B}$ this

* Received by the editors September 27, 1979, and in revised form January 12, 1981.This research was sponsored in part by the United States Army under contracts DAAG29-75-C-0024 and DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grants MCS78-09525 A01 and MCS75-07870 A01.

† Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

‡ Department of Mathematics RLM 8.100, The University of Texas at Austin, Austin, Texas 78712.

regularization is analogous to the Yoshida approximation. The operator $\mathscr{A}$ is assumed to be a subgradient in the above. Finally, we show the initial-value problem can be solved for (1.2) when $\mathscr{B}$ (but not necessarily $\mathscr{A}$) is a subgradient.

We mention some related work on equations of the form in (1.1). The theory of such implicit evolution equations divides historically into three cases. The first and certainly the easiest is where $\mathscr{B}\circ\mathscr{A}^{-1}$ is Lipschitz or monotone in some space [6], [23]. The second is that one of the operators is (linear) self-adjoint, and this case includes the majority of the applications to problems where singular or degenerate behavior arises due to spatial coefficients or geometry [2], [25]. These situations are described in the book [9] to which we refer for details and a very extensive bibliography. The third case is that wherein both operators are possibly nonlinear. This considerably more difficult case has been investigated by Grange and Mignot [12] and more recently by Barbu [4]. In both of these studies a compactness assumption similar to ours is made. Our boundedness assumptions are more restrictive than those in the papers above, but they assume f is smooth and that both operators are subgradients. By not requiring that $\mathscr{B}$ be a subgradient in (1.1) we obtain a significantly larger class of applications to partial differential equations, especially to systems.

Our work is organized as follows. In § 2 we recall certain information on maximal monotone operators and then state our results on the existence of solutions of the initial-value problems (1.1) and for (1.2). The proofs are given in §§ 3 and 4. Section 5 contains elementary examples of how nonuniqueness occurs, and we show there that uniqueness holds in the situation where one of the operators is self-adjoint. Section 6 is concerned with the structure and construction of maximal monotone operators between Hilbert spaces which characterize certain partial differential equations and associated boundary conditions. These operators are used to present in § 7 a collection of initial-boundary-value problems for partial differential equations which illustrate the applications of our results to the existence theory of such problems.

2. Preliminaries and main results. We begin by reviewing information on maximal monotone operators. Refer to [1], [3], [11] for additional related material and proofs. Then we shall state our existence theorems for the Cauchy problem (1.1).

Let V be a real Hilbert space and A a subset of the product $V\times V$. We regard A as a function from V to 2^V, the set of subsets of V, or as a multi-valued mapping or operator from V into V; thus, $f\in A(u)$ means $[u,f]\in A$. We define the domain $D(A)=\{u\in V: Au \text{ nonempty}\}$, range $R_g(A)=\cup\{Au: u\in V\}$ and inverse $A^{-1}(u)=\{v\in V: u\in A(v)\}$ of A as indicated. The operator A is *monotone* if $(f_1-f_2, u_1-u_2)_V\geqq 0$ whenever $[u_j,f_j]\in A$ for $j=1,2$. This is equivalent to $(I+\lambda A)^{-1}$ being a contraction for every $\lambda>0$. We call A *maximal monotone* if it is maximal in the sense of inclusion of graphs. Then we have a monotone A maximal monotone if and only if $R_g(I+\lambda A)=V$ for some (hence, all) $\lambda>0$. If A is maximal monotone we can define its *resolvent* $J_\lambda\equiv(I+\lambda A)^{-1}$, a contraction defined on all V, and its *Yoshida approximation* $A_\lambda=\lambda^{-1}(I-J_\lambda)$, a monotone Lipschitz function defined on all V. For $u\in V$ we have $A_\lambda(u)\in A(J_\lambda(u))$. We denote weak convergence of x_n to x by $x_n\rightharpoonup x$.

LEMMA 2.1. *Let A be maximal monotone, $[x_n, y_n]\in A$ for $n\geqq 1$, $x_n\rightharpoonup x$, $y_n\rightharpoonup y$ and* $\liminf (y_n, x_n)_V\leqq(y, x)_V$. *Then $[x, y]\in A$. If in addition* $\limsup (y_n, x_n)_V\leqq(y, x)_V$, *then $(y_n, x_n)_V\to(y, x)_V$. We observe that A induces on $L^2(0, T; V)$ a maximal monotone operator (denoted also by A) defined by $v\in A(u)$ if and only if $v(t)\in A(u(t))$ for a.e. $t\in[0, T]$.*

A special class of maximal monotone operators arises as follows. If $\varphi: V\to(-\infty,\infty]$ is a proper, convex and lower semicontinuous function, we define the *subgradient*

$\partial\varphi \subset V \times V$ by

$$\partial\varphi(x) = \{z \in V: \varphi(y) - \varphi(x) \geqq (z, y - x) \text{ for all } y \in V\}.$$

The operator $\partial\varphi$ is maximal monotone. Furthermore it is useful to consider the convex *conjugate* of φ defined by

$$\varphi^*(z) \equiv \sup\{(z, y)_V - \varphi(y), y \in V\}.$$

The following are equivalent: $z \in \partial\varphi(x)$, $x \in \partial\varphi^*(z)$, and $\varphi(x) + \varphi^*(z) = (x, z)_V$; thus $\partial\varphi^*$ is the inverse of $\partial\varphi$. We mention the following *chain rule* [1]. Let $H^1(0, T; V)$ denote the space of absolutely continuous V-valued functions on $[0, T]$ whose derivatives belong to $L^2(0, T; V)$.

LEMMA 2.2. *If* $u \in H^1(0, T; V)$, $v \in L^2(0, T; V)$ *and* $[u(t); v(t)] \in \partial\varphi$ *for a.e.* $t \in [0, T]$, *then the function* $t \to \varphi(u(t))$ *is absolutely continuous on* $[0, T]$ *and*

$$\frac{d}{dt}\varphi(u(t)) = (w, u'(t)_V), \quad \text{all } w \in \partial\varphi(u(t)),$$

for a.e. $t \in [0, T]$.

There is a version of a monotone operator from V to its dual space V^* which is equivalent to the above through the Riesz map $\mathscr{R}: V \to V^*$. Thus, $\mathscr{A} \subset V \times V^*$ is monotone if and only if $A \equiv \mathscr{R}^{-1} \circ \mathscr{A}$ is monotone in $V \times V$ and maximal monotone if and only if $R_g(\mathscr{R} + \mathscr{A}) = V^*$ in addition. We shall use these two equivalent notions interchangeably. Our applications to partial differential equations will lead to operators on $V \times V^*$. Also the subgradient is naturally constructed in the $W - W^*$ duality of a Banach (or topological vector) space W. Finally we cite the following chain rule.

LEMMA 2.3. *Let* V *and* W *be locally convex spaces with duals* V^* *and* W^*. *Let* $\Lambda: V \to W$ *be continuous and linear with dual* $\Lambda^*: W^* \to V^*$. *If* $\varphi: W \to (-\infty, \infty]$ *is proper, convex and lower semicontinuous then so also is* $\varphi \circ \Lambda: V \to (-\infty, \infty]$, *and if* φ *is continuous at some point of* $R_g(\Lambda)$ *we have* [11]

$$\partial(\varphi \circ \Lambda) = \Lambda^* \circ \partial\varphi \circ \Lambda.$$

Our results on the existence of solutions of the Cauchy problem (1.1) are stated as follows.

THEOREM 1. *Let* W *be a reflexive Banach space and* V *a Hilbert space which is dense and embedded compactly in* W. *Denote the injection by* $i: V \to W$ *and the dual (restriction) operator by* i^* $W^* \to V^*$. *Assume the following*:

[A_1] *The real-valued* φ *is proper, convex and lower semicontinuous on* W, *continuous at some point of* V, *and* $\partial\varphi \circ i: V \to W^*$ *is bounded.*

[B_1] *The operator* $\mathscr{B}: V \to V^*$ *is maximal monotone and bounded. Define* $\mathscr{A} \equiv i^* \circ \partial\varphi \circ i$. *Then for each given* $f \in L^2(0, T; V^*)$ *and* $[u_0, v_0] \in \mathscr{A}$ *there exists a triple* $u \in H^1(0, T; V)$, $v \in H^1(0, T; V^*)$, *and* $w \in L^2(0, T; V^*)$ *such that*

$$\frac{d}{dt}(\mathscr{R}u(t) + v(t)) + w(t) = f(t), \tag{2.1a}$$

$$v(t) \in \mathscr{A}(u(t)), \ w(t) \in \mathscr{B}(u(t)), \quad a.e.\ t \in [0, T], \tag{2.1b}$$

$$\mathscr{R}u(0) + v(0) = \mathscr{R}u_0 + v_0. \tag{2.1c}$$

THEOREM 2. *In addition to the above, assume*:

[A_2] $\partial\varphi \circ i: L^2(0, T; V) \to L^2(0, T; W^*)$ *is bounded.*

[B_2] *$\mathcal{B}: L^2(0, T; V) \to L^2(0, T; V^*)$ is bounded and coercive, i.e.,*

$$\lim_{\substack{\|u\|_{L^2(0,T;V)} \to +\infty \\ [u,v] \in \mathcal{B}}} \frac{\int_0^T v(t)(u(t))\, dt}{\|u\|_{L^2(0,T;V)}} = +\infty.$$

Then for each given $f \in L^2(0, T; V^*)$ and $v_0 \in R_g(\mathcal{A})$ there exists a triple $u \in L^2(0, T; V)$, $v \in H^1(0, T; V^*)$, $w \in L^2(0, T; V^*)$ such that

$$\frac{d}{dt} v(t) + w(t) = f(t), \tag{2.2a}$$

$$v(t) \in \mathcal{A}(u(t)),\ w(t) \in \mathcal{B}(u(t)), \quad a.e.\ t \in [0, T], \tag{2.2b}$$

$$v(0) = v_0. \tag{2.2c}$$

Remarks. From Lemma 2.3 it follows that $\mathcal{A} = \partial(\varphi|_V)$ where $\varphi|_V = \varphi \circ i$ is the restriction of φ to V. Since $\mathcal{A}: V \to V^*$ is bounded it follows that $D(\mathcal{A}) = V$; hence,

$$V \subset D(\partial\varphi) \subset \text{dom}\,(\varphi) \subset W,$$

and φ is continuous on the space V. Also, since $\varphi(0) < \infty$ we may assume with no loss of generality that $\varphi(0) \geqq 0$ and thus $\varphi^*(z) \geqq 0$ for all $z \in V$.

From the compactness of $i^*: W^* \to V^*$ it follows that $\mathcal{A}: V \to V^*$ is compact, i.e., maps bounded sets into relatively compact sets.

Since $\mathcal{B}$ is bounded and maximal monotone we have $D(\mathcal{B}) = V$. It is important for our applications that we have made no assumptions which directly relate $\mathcal{A}$ and $\mathcal{B}$. Specifically, we do not compare $\mathcal{A}(x)$ and $\mathcal{B}(x)$ in angle or in norm.

Finally, we give a variation on Theorem 1 in which only the second operator $\mathcal{B}$ is a subgradient. The compactness assumption on $\mathcal{A}$ is retained.

THEOREM 3. *Let the spaces V and W be given as before. Assume the following:*

[A_3] *The operator $\mathcal{A}: V \to V^*$ is maximal monotone with $R_g(\mathcal{A}) \subset W^*$ and $\mathcal{A}: V \to W^*$ is bounded.*

[B_3] *The real-valued ψ is proper, convex and lower semicontinuous on V and $\mathcal{B} \equiv \partial\psi: V \to V^*$ is bounded.*

Then for given $f \in L^2(0, T; V^*)$ and $[u_0, v_0] \in \mathcal{A}$ there exists a triple $u \in H^1(0, T; V)$, $v \in H^1(0, T; V^*)$ and $w \in L^2(0, T; V^*)$ satisfying (2.1).

3. Proofs of Theorem 1 and Theorem 3. These proofs are very similar; let us consider first Theorem 1. We formulate (2.1) in the space V. Set $A = \mathcal{R}^{-1} \circ \mathcal{A}$, $B = \mathcal{R}^{-1} \circ \mathcal{B}$, etc., and consider the equivalent equation

$$\frac{d}{dt}(u(t) + v(t)) + w(t) = f(t), \tag{3.1a}$$

$$v(t) \in A(u(t)), \qquad w(t) \in B(u(t)), \quad \text{a.e. } t \in [0, T]. \tag{3.1b}$$

Let $\lambda > 0$ and consider the approximation of (3.1) by

$$\frac{d}{dt}(u_\lambda(t) + v_\lambda(t)) + B_\lambda(u_\lambda(t)) = f(t), \tag{3.2a}$$

$$v_\lambda(t) \in A(u_\lambda(t)), \qquad t \in [0, T]. \tag{3.2b}$$

Since $(I + A)^{-1}$ and B_λ are both Lipschitz continuous from V to V, (3.2) has a unique absolutely continuous solution u_λ with $u_\lambda(0) + v_\lambda(0) = u_0 + v_0$. Since $(I + A)^{-1}$ is a function, we have $u_\lambda(0) = u_0$ and $v_\lambda(0) = v_0$.

We derive a priori estimates on u_λ. Take the scalar product in V of (3.2a) with $u_\lambda(t)$ and note

$$(v_\lambda'(t), u_\lambda(t))_V = \frac{d}{dt}\varphi^*(v_\lambda)$$

by Lemma 2.2, where φ^* is the conjugate of $\varphi|_V$ in V. Integrating the resulting identity gives

$$\frac{1}{2}\|u_\lambda(t)\|_V^2 + \varphi^*(v_\lambda(t))$$

$$\leqq \frac{1}{2}\|u_0\|_V^2 + \varphi^*(v_0) + \int_0^t (\|f(s)\|_V + \|B_\lambda(0)\|_V)\|u_\lambda(s)\|_V\, ds, \qquad 0 \leqq t \leqq T.$$

Since $\{B_\lambda(0)\}$ is bounded by the fact that $0 \in D(B)$, $\varphi^* \geqq 0$ and $f \in L^2(0, T; V)$, we have proved the first part of the following lemma.

LEMMA 3.1. *The following are bounded independent of* $\lambda > 0$:

(a) $\|u_\lambda\|_{L^\infty(0,T;V)}$, $\|\mathscr{R}v_\lambda\|_{L^\infty(0,T;W^*)}$, $\|J_\lambda(u_\lambda)\|_{L^\infty(0,T;V)}$, $\|B_\lambda(u_\lambda)\|_{L^\infty(0,T;V)}$,

(b) $\|u_\lambda'\|_{L^2(0,T;V)}$, $\|v_\lambda'\|_{L^2(0,T;V)}$.

Proof. The second and third terms of (a) are bounded because the operators $\mathscr{A}: V \to W^*$ and $J_\lambda \equiv (I + \lambda B)^{-1}: V \to V$ are bounded. Since $B_\lambda(u_\lambda) \in B(J_\lambda(u_\lambda))$ and B is bounded, the last term in (a) is bounded.

To obtain (b) we take the scalar product of (3.2a) by $u_\lambda'(t)$, note that $(v_\lambda'(t), u_\lambda'(t))_V \geqq 0$ by (3.2.b) and the monotonicity of A, and thereby obtain

$$\|u_\lambda'(t)\|_V^2 \leqq (\|f(t)\|_V + \|B_\lambda(u_\lambda(t)\|_V)\|u_\lambda'(t)\|_V,$$

so we bound the first term in (b). The second follows from (3.2.a).

Note that we have $\{\mathscr{R}v_\lambda\}$ bounded in $L^2(0, T; W^*)$ and $\{\mathscr{R}v_\lambda'\}$ bounded in $L^2(0, T; V^*)$. Since W^* is compact in V^* it follows from [17, p. 58] that $\{\mathscr{R}v_\lambda\}$ is (strongly) relatively compact in $L^2(0, T; V^*)$. From this observation and Lemma 3.1 it follows that we may pass to a subsequence, again denoted by u_λ, v_λ, for which we have

(3.3a) $$u_\lambda \rightharpoonup u, \quad B_\lambda(u_\lambda) \rightharpoonup w, \quad u_\lambda' \rightharpoonup u',$$

(3.3b) $$v_\lambda \to v \text{ (strongly)}, \quad v_\lambda' \rightharpoonup v' \text{ in } L^2(0, T; V),$$

(3.3c) $$u_\lambda(t) \rightharpoonup u(t) \quad \text{and} \quad v_\lambda(t) \to v(t), \quad \text{all } t \in [0, T].$$

Since $u_\lambda - J_\lambda(u_\lambda) = \lambda B_\lambda(u_\lambda) \to 0$ there follows

(3.3d) $$J_\lambda(u_\lambda) \rightharpoonup u \quad \text{in } L^2(0, T; V).$$

It remains to show that u, v, w satisfy (3.1) and the initial condition. First we use (3.3a) and (3.3b) and Lemma 2.1 to obtain $v \in A(u)$. Next we take the scalar product of (3.2a) with any $x \in V$ and integrate to get

$$(u_\lambda(t) + v_\lambda(t), x)_V + \int_0^t (B_\lambda(u_\lambda(s)), x)_V\, ds = \int_0^t (f(s), x)_V\, ds + (u_0 + v_0, x)_V.$$

Taking the limit as $\lambda \to 0$ gives (since x is arbitrary)

$$u(t) + v(t) + \int_0^t (w - f)\, ds = u_0 + v_0, \qquad 0 \leqq t \leqq T.$$

From this identity we obtain (3.1a) and $u(0)+v(0)=u_0+v_0$; since $v(0)\in A(u(0))$ and $(I+A)^{-1}$ is a function we have $u(0)=u_0$. In order to show $w\in B(u)$, and thereby finish the proof of Theorem 1, it suffices by Lemma 2.1 to show

$$\limsup_{\lambda\to 0}(B_\lambda(u_\lambda), J_\lambda(u_\lambda)-u)_{L^2(0,T;V)}\leqq 0.$$

We note further that

$$\begin{aligned}(B_\lambda(u_\lambda), J_\lambda(u_\lambda)) &= (B_\lambda(u_\lambda), J_\lambda(u_\lambda)-u_\lambda)+(B_\lambda(u_\lambda), u_\lambda)\\ &= -\lambda(B_\lambda(u_\lambda), B_\lambda(u_\lambda))+(B_\lambda(u_\lambda), u_\lambda)\end{aligned}$$

so it suffices to show

$$\limsup_{\lambda\to 0}(B_\lambda(u_\lambda), u_\lambda-u)_{L^2(0,T;V)}\leqq 0. \tag{3.4}$$

By (3.2a) it follows (3.4) is equivalent to

$$\liminf_{\lambda\to 0}(u_\lambda'+v_\lambda', u_\lambda-u)_{L^2(0,T;V)}\geqq 0. \tag{3.5}$$

Define $\psi(x)=\frac{1}{2}\|x\|_V^2+\varphi(x)$, $x\in V$ so that $\partial\psi=I+\partial\varphi$. From (3.2b) and Lemma 2.2 we obtain

$$(u_\lambda'(t)+v_\lambda'(t), u_\lambda(t))_V=\frac{d}{dt}\psi^*(u_\lambda(t)+v_\lambda(t)),$$

and integrating yields

$$(u_\lambda'+v_\lambda', u_\lambda)_{L^2(0,T;V)}=\psi^*(u_\lambda(T)+v_\lambda(T))-\psi^*(u_0+v_0).$$

Similarly we have from (3.1a)

$$(u'+v', u)_{L^2(0,T;V)}=\psi^*(u(T)+v(T))-\psi^*(u_0+v_0).$$

By (3.3c) and weak lower semicontinuity of ψ^* we have

$$\psi^*(u(T)+v(T))\leqq\liminf_{\lambda\to 0}\psi^*(u_\lambda(T)+v_\lambda(T)),$$

and our preceding calculations show that this is equivalent to (3.5).

Remark 3.1. From Lemma 2.1 we find that

$$(B_\lambda(u_\lambda), J_\lambda(u_\lambda))_{L^2(0,T;V)}\to(w, u)_{L^2(0,T;V)}.$$

If we also have B (or $\mathcal{B}$) *strongly monotone* then we can take the limit in the estimate

$$(B_\lambda(u_\lambda)-w, J_\lambda(u_\lambda)-u)_{L^2(0,T;V)}\geqq c\|J_\lambda(u_\lambda)-u\|^2_{L^2(0,T;V)}$$

to conclude that $\{J_\lambda(u_\lambda\}$ and $\{u_\lambda\}$ converge strongly to u in $L^2(0, T; V)$.

Remark 3.2. It is clear that we actually have $v(t)\in A(u(t))$ for every $t\in[0, T]$.

The proof of Theorem 3 closely follows the preceding pattern. That is, formulate (2.1) as the equivalent initial value problem for (3.1) and approximate this by (3.2) with $u_\lambda(0)+v_\lambda(0)=u_0+v_0$ for each $\lambda>0$.

To derive a priori bounds we take the scalar product of (3.2a) with $u_\lambda'(t)$ and integrate to obtain

$$\int_0^T\|u_\lambda'\|_V^2+\psi_\lambda(u_\lambda(T))\leqq\psi_\lambda(u_0)+\int_0^T(f(t), u_\lambda'(t))_V\,dt. \tag{3.6}$$

Here ψ_λ is the Yoshida approximation of ψ. We may assume ψ is nonnegative and the same holds for ψ_λ, so we have the first part of the following.

LEMMA 3.2. *The following are bounded independent of* $\lambda > 0$:

(a) $$\|u_\lambda\|_{L^\infty(0,T;V)}, \quad \|u'_\lambda\|_{L^2(0,T;V)}, \quad \|\mathscr{R}v_\lambda\|_{L^\infty(0,T;W^*)},$$
$$\|\boldsymbol{J}_\lambda(u_\lambda)\|_{L^\infty(0,T;V)}, \quad \|B_\lambda(u_\lambda)\|_{L^\infty(0,T;V)},$$

(b) $$\|v_\lambda\|_{L^\infty(0,T;V)}, \qquad \|v'_\lambda\|_{L^2(0,T;V)}.$$

Proof. The bound on the first two terms in (a) follow from (3.6) and the remaining terms in (a) are bounded by $[A_3]$ and $[B_3]$. Next we take the scalar product of (3.2a) with $v'_\lambda(t)$, and obtain (b) as was done in Lemma 3.1.

We may pass to a subsequence satisfying (3.3) and we obtain as in Theorem 1 the triple u, v, w satisfying the equation (3.1a) and initial condition and $v(t) \in Au(t)$, $t \in [0, T]$. It remains to show $w \in B(u)$ and this is equivalent to showing (cf. (3.5))

$$\liminf_{\lambda \to 0} (u'_\lambda + v'_\lambda, u_\lambda)_{L^2(0,T;V)} \geqq (u' + v', u)_{L^2(0,T;V)}. \tag{3.7}$$

Since $u'_\lambda \in L^2(0, T; V)$ we may integrate by parts to compute

$$(u'_\lambda + v'_\lambda, u_\lambda)_{L^2(0,T;V)} = \tfrac{1}{2}(\|u_\lambda(T)\|_V^2 - \|u_0\|_V^2) - (v_\lambda, u'_\lambda)_{L^2(0,T;V)} + (v_\lambda(T), u_\lambda(T))_V - (v_0, u_0) \tag{3.8a}$$

and similarly, since $u' \in L^2(0, T; V)$,

$$(u' + v', u)_{L^2(0,T;V)} = \tfrac{1}{2}(\|u(T)\|_V^2 - \|u_0\|_V^2) - (v, u')_{L^2(0,T;V)} + (v(T), u(T))_V - (v_0, u_0)_V. \tag{3.8b}$$

Finally we observe that (3.7) follows immediately from (3.3) and (3.8).

Remark 3.2. If in addition B is strongly monotone, then $\{u_\lambda\}$ converges strongly to u in $L^2(0, T; V)$.

4. Proof of Theorem 2. Choose $u_0 \in A^{-1}(v_0)$. For each $\lambda > 0$ let $u_\lambda, v_\lambda \in H^1(0, T; V)$, $w_\lambda \in L^2(0, T; V)$ satisfy

$$\lambda u'_\lambda(t) + v'_\lambda(t) + w_\lambda(t) = f(t), \tag{4.1a}$$

$$v_\lambda(t) \in \boldsymbol{A}(u_\lambda(t)), \qquad w_\lambda(t) \in \boldsymbol{B}(u_\lambda(t)), \qquad \text{a.e. } t \in [0, T], \tag{4.1b}$$

$$\lambda u_\lambda(0) + v_\lambda(0) = \lambda u_0 + v_0. \tag{4.1c}$$

The problem (4.1) has a solution by Theorem 1, and our plan is to show that we may take the limit as $\lambda \to 0$ in (4.1) to obtain a solution u, $w \in L^2(0, T; V)$, $v \in H^1(0, T; V)$ of

$$v'(t) + w(t) = f(t) \tag{4.2a}$$

$$v(t) \in \boldsymbol{A}(u(t)), \qquad w(t) \in \boldsymbol{B}(u(t)), \quad \text{a.e. } t \in [0, T], \tag{4.2b}$$

$$v(0) = v_0. \tag{4.2c}$$

With our notation $A = \mathscr{R}^{-1} \circ \mathscr{A}$, etc., (4.2) is equivalent to (2.2).

We proceed to derive a priori estimates. Consider first the initial condition. Since $(\lambda I + A)^{-1}$ is a function it follows from (4.1c) that

$$u_\lambda(0) = u_0, \qquad v_\lambda(0) = v_0, \qquad \lambda > 0. \tag{4.3}$$

LEMMA 4.1. *The following are bounded independent of* $\lambda > 0$:

(a) $$\|u_\lambda\|_{L^2(0,T;V)}, \qquad \lambda^{1/2}\|u_\lambda\|_{L^\infty(0,T;V)},$$

(b) $$\|w_\lambda\|_{L^2(0,T;V)}, \qquad \|\mathcal{R}v_\lambda\|_{L^2(0,T;W^*)}.$$

Proof. Take the scalar product of (4.1a) with $u_\lambda(t)$ and integrate to obtain

$$\begin{aligned} &\frac{\lambda}{2}\|u_\lambda(t)\|_V^2 + \varphi^*(v_\lambda(t)) + \int_0^t (w_\lambda, u_\lambda)_V \\ &\leqq \frac{\lambda}{2}\|u_0\|_V^2 + \varphi^*(v_0) + \int_0^t (f, u_\lambda)_V, \qquad 0 \leqq t \leqq T. \end{aligned} \tag{4.4}$$

We drop the second (nonnegative) term in (4.4) and note by the monotonicity of B that $(w_\lambda, u_\lambda)_V \geqq (\xi, u_\lambda)_V$ for some $\xi \in B(0)$. Thus (4.4) gives

$$\int_0^T (w_\lambda, u_\lambda)_V \leqq \|f\|_{L^2(0,T;V)}\|u_\lambda\|_{L^2(0,T;V)} + C,$$

and the coercivity of B implies the boundedness of the first term in (a). The second now follows from (4.4) and now part (b) follows from our assumptions $[A_1]$ and $[B_1]$.

LEMMA 4.2. *The following are bounded independent of* $\lambda > 0$:

$$\|v_\lambda'\|_{L^2(0,T;V)}, \qquad \|\lambda u_\lambda'\|_{L^2(0,T;V)}.$$

Proof. Take the scalar product of (4.1a) with $v_\lambda'(t)$. Since $(u_\lambda'(t), v_\lambda'(t))_V \geqq 0$ by the monotonicity of A, we obtain

$$\|v_\lambda'(t)\|_V^2 \leqq (\|f(t)\|_V + \|w_\lambda(t)\|_V)\|v_\lambda'(t)\|_V,$$

from which the first bound is immediate. To obtain the second we take the scalar product of (4.1a) with $u_\lambda'(t)$ and drop the nonnegative term $(u_\lambda'(t), v_\lambda'(t))_V$. This gives

$$\lambda\|u_\lambda'(t)\|_V^2 \leqq (\|f(t)\|_V + \|w_\lambda(t)\|_V)\|u_\lambda'(t)\|_V,$$

and hence the desired bound.

We have now shown that $\{\mathcal{R}v_\lambda\}$ is bounded in $L^2(0, T; W^*)$ and that $\{\mathcal{R}v_\lambda'\}$ is bounded in $L^2(0, T; V^*)$. Since W^* is compact in V^* it follows that $\{\mathcal{R}v_\lambda\}$ is strongly compact in $L^2(0, T; V^*)$. From this observation, Lemma 4.1 and Lemma 4.2 it follows we may pass to a subsequence (which we denote again by $\{u_\lambda\}, \{v_\lambda\}, \{w_\lambda\}$) for which in $L^2(0, T; V)$ we have

$$u_\lambda \rightharpoonup u, \quad w_\lambda \rightharpoonup w, \quad v_\lambda \rightharpoonup v, \quad v_\lambda' \rightharpoonup v'.$$

Note that $\lambda u_\lambda \to 0$ and it follows that $\lambda u_\lambda' \rightharpoonup 0$ by standard arguments. Furthermore, we may assume $v_\lambda(t) \to v(t)$ in V for all $t \in [0, T]$ by equicontinuity of $\{v_\lambda\}$, and similarly $\lambda u_\lambda(t) \to 0$ in V for all $t \in [0, T]$.

It remains to show that the triple u, v, w obtained above constitutes a solution of (4.2). Let $x \in V$, take the scalar product of x with (4.1a) and integrate to obtain

$$(\lambda u_\lambda(t) + v_\lambda(t), x)_V + \int_0^t (w_\lambda(s), x)_V = \int_0^t (f(s), x)_V\, ds + (\lambda u_0 + v_0, x)_V.$$

Since weak convergence in $L^2(0, T; V)$ implies weak convergence in $L^2(0, t; V)$ letting $\lambda \to 0$ gives that

$$(v(t), x)_V + \int_0^t (w(s), x)_V\, ds = \int_0^t (f(s), x)_V\, ds + v_0, \qquad x \in V, \quad t \in [0, T].$$

That is,

$$v(t)+\int_0^t w(s)\,ds=\int_0^t f(s)\,ds+v_0,\quad \text{a.e. } t\in[0,T],$$

and this implies (4.2a) and (4.2c). From Lemma 2.1 there follows $v\in A(u)$ so it remains only to establish $w\in B(u)$. For this it suffices by Lemma 2.1 to show

$$\limsup_{\lambda\to 0}(w_\lambda, u_\lambda)_{L^2(0,T;V)}\leqq(w,u)_{L^2(0,T;V)}. \tag{4.5}$$

In order to prove (4.5) we first note by (4.1a) and (4.2a) that it is equivalent to

$$\liminf_{\lambda\to 0}(\lambda u_\lambda'+v_\lambda', u_\lambda)_{L^2(0,T;V)}\geqq(v',u)_{L^2(0,T;V)}. \tag{4.6}$$

Since $u_\lambda(t)\in A^{-1}(v_\lambda(t))=\partial\varphi^*(v_\lambda(t))$ a.e. on $[0,T]$, where φ^* is the conjugate of $\varphi|_V$, we obtain from Lemma 2.2

$$\begin{aligned}(\lambda u_\lambda'+v_\lambda', u_\lambda)_{L^2(0,T;V)}&=\frac{\lambda}{2}\|u_\lambda(T)\|_V^2+\varphi^*(v_\lambda(T))-\frac{\lambda}{2}\|u_0\|_V^2-\varphi^*(v_0)\\&\geqq\varphi^*(v_\lambda(T))-\frac{\lambda}{2}\|u_0\|_V^2-\varphi^*(v_0).\end{aligned}$$

Similarly we compute

$$(v',u)_{L^2(0,T;V)}=\varphi^*(v(T))-\varphi^*(v_0).$$

Since $\{v_\lambda\}$ are equi-uniformly-continuous we have $v_\lambda(t)\to v(t)$ at every $t\in[0,T]$, so the lower semicontinuity of φ^* gives

$$\liminf_{\lambda\to 0}\varphi^*(v_\lambda(T))\geqq\varphi^*(v(T)).$$

In view of the preceding computations this is exactly (4.6).

Remark 4.1. If B is strongly monotone then $\{u_\lambda\}$ converges strongly to u in $L^2(0,T;V)$.

5. Remarks on uniqueness. We first present an example which shows that gross nonuniqueness of solutions of (1.1) can occur, even if both operators are strongly monotone subgradients. Moreover the nonuniqueness occurs in each term of the triple u, v, w, not just in the latter two terms selected, respectively, from $A(u)$ and $B(u)$. Next we shall show that uniqueness does hold for (1.1) when at least one of the operators is continuous, linear and symmetric and the sum of the operators is strictly monotone. Our last example shows that symmetry of the linear operator is essential.

Example 1. Let $V=W=R$, the space of real numbers, and define

$$A(s)=B(s)=s+H(s-1),$$

where

$$H(r)=\begin{cases}1, & r>0,\\ [0,1], & r=0,\\ 0, & r<0\end{cases}$$

denotes the Heaviside function and $f\equiv 0$. Consider the initial-value problem (1.1), which takes the form

$$\begin{aligned}&v'(t)+w(t)=0, \qquad v(0)=2,\\ &v(t)-u(t)\in H(u(t)-1), \qquad w(t)-u(t)\in H(u(t)-1).\end{aligned} \tag{5.1}$$

Let g be any maximal monotone graph or continuous function from R to R such that $g(s)=s$ for $s\notin[1,2]$ and $g(s)\subset[1,2]$ for $s\in[1,2]$. Then, if v is a solution of

$$v'(t)+g(v(t))=0, \qquad t\geqq 0, \quad v(0)=2, \tag{5.2}$$

it follows that with $u(t)\equiv A^{-1}(v(t))$ and $w(t)\equiv -v'(t)$ we have a solution of (5.1). This procedure yields an abundance of solutions.

We display some special cases of the above. Pick $c\in[\frac{1}{2},1]$ and define g_c to be the maximal monotone graph such that $g_c(t)=\{c^{-1}\}$, $t\in(1,2)$, and $g_c(t)=\{t\}$, $t\notin[1,2]$. The corresponding solution v_c of (5.2) is given by

$$v_c(t)=2-\frac{t}{c}, \quad 0\leqq t\leqq c, \qquad v_c(t)=e^{c-t}, \quad t\geqq c.$$

With the two functions u_c and w_c given by

$$u_c(t)=1, \qquad w_c(t)=\frac{1}{c} \quad \text{for } 0\leqq t\leqq c,$$

$$u_c(t)=w_c(t)=e^{c-t}, \qquad t\geqq 0,$$

this provides a continuum of solutions of (5.1).

We can give the following elementary sufficient conditions for uniqueness to hold for (1.1) or, equivalently, for (4.2).

THEOREM 4. *Let A and B be monotone operators on a Hilbert space V. Suppose $A+B$ is strictly monotone and that one of A or B is continuous, linear and symmetric. Then for each function $f:[0,T]\to V$ and $v_0\in V$ there is at most one solution u, v, w of* (4.2).

Proof. Suppose A is continuous, linear and symmetric. For $j=1,2$ let u_j, v_j, w_j be a solution of (4.2). Take the scalar product of the difference of (4.2a) with u_1-u_2 to obtain

$$\frac{1}{2}\frac{d}{dt}(A(u_1(t)-u_2(t)),u_1(t)-u_2(t))_V+(w_1(t)-w_2(t),u_1(t)-u_2(t))_V=0.$$

Integrating this identity and using (4.2c) gives

$$\frac{1}{2}(A(u_1(t)-u_2(t)),u_1(t)-u_2(t))_V+\int_0^t(w_1-w_2,u_1-u_2)_V\,ds=0, \qquad 0\leqq t\leqq T,$$

and this implies

$$Au_1(t)=Au_2(t),\ (w_1(t)-w_2(t),u_1(t)-u_2(t))_V=0 \quad \text{a.e. } t\in[0,T].$$

Since $A+B$ is strictly monotone we have $u_1(t)=u_2(t)$; hence $v_1(t)=Au_1(t)=Au_2(t)=v_2(t)$ and, by (4.1a), $w_1(t)=w_2(t)$ a.e. on $[0,T]$.

Suppose now B is continuous, linear and symmetric. Starting with two solutions as above we integrate the corresponding equations (4.2a) to obtain

$$v_j(t)+B(\theta_j(t))=v_0+\int_0^t f, \qquad j=1,2, \tag{5.3}$$

where $\theta_j(t)\equiv\int_0^t u_j$. Taking the difference of (5.3) for $j=1,2$, then the scalar product with $\theta_1'-\theta_2'$ and integrating gives us

$$\int_0^t(v_1-v_2,\theta_1'-\theta_2')_V+\frac{1}{2}(B(\theta_1(t)-\theta_2(t)),\theta_1(t)-\theta_2(t))_V=0. \tag{5.4}$$

Since $v_j(t) \in A(\theta_j'(t))$ a.e., each term is nonnegative. It follows that $B(\theta_1(t) - \theta_2(t)) = 0$ on $[0, T]$, and thus from (5.4) that

$$(v_1(t) - v_2(t), u_1(t) - u_2(t))_V = 0 \quad \text{a.e. } t \in [0, T],$$

so the desired results follows by strict monotonicity of $A + B$.

Finally we cite an example to show that the symmetry condition cannot be eliminated from Theorem 4.

Example 2. Let $H^1(0, 1)$ be the Sobolev space of those absolutely continuous functions on the interval $(0, 1)$ whose first derivatives belong to $L^2(0, 1)$; set $V = \{v \in H^1(0, 1): v(1) = 0\}$ and note that $V \subset L^2(0, 1) \subset V^*$. Define $\mathscr{A}: V \to V^*$ by $\mathscr{A}(v) = -v'$. Clearly $\mathscr{A}$ is linear and we have

$$\mathscr{A}(v)(v) = -\int_0^1 v'v \, ds = \frac{1}{2}|v(0)|^2 \geqq 0,$$

so $\mathscr{A}$ is monotone. Let β be given by

$$\beta(r) = \begin{cases} \dfrac{r}{2}, & r < 0 \text{ or } r > 1, \\ \dfrac{r^2}{2}, & 0 \leqq r \leqq 1, \end{cases}$$

and define $\mathscr{B}: V \to V^*$ by

$$\mathscr{B}(u)(v) = \int_0^1 \beta(u'(s))v'(s) \, ds, \qquad u, v \in V.$$

It is easy to check that $\mathscr{B}$ is a strictly monotone subgradient on V.

Consider the Cauchy problem

$$\frac{d}{dt}\mathscr{A}(u) + \mathscr{B}(u) = 0, \qquad \mathscr{A}u(0) = -1 \tag{5.5}$$

with the above operators. A solution u of (5.5) is a weak solution of the initial-boundary-value problem

$$(-u_x)_t - (\beta(u_x))_x = 0, \qquad 0 < x < 1, \quad 0 < t, \tag{5.6a}$$

$$u_x(0, t) = u(1, t) = 0, \tag{5.6b}$$

$$-u_x(x, 0) = -1, \tag{5.6c}$$

where the subscripts denote partial derivatives. Consider the following two functions:

$$u^{(1)}(x, t) = \begin{cases} \dfrac{x^2 + t^2}{2t} - 1, & 0 < x < t < 1, \\ x - 1, & 0 < t < x < 1, \end{cases}$$

$$u^{(2)}(x, t) = \begin{cases} \dfrac{t}{2} - 1, & 0 < x < \dfrac{t}{2} < \dfrac{1}{2}, \\ x - 1, & 0 < \dfrac{t}{2} < x < 1, \quad t < 1. \end{cases}$$

It is a straightforward computation to check that both $u^{(1)}$ and $u^{(2)}$ are solutions of (5.6), hence, both are solutions of (5.5). Note that the only condition of Theorem 4 not met in this example is the symmetry of $\mathscr{A}$. It shows also that $\mathscr{B}$ being a subgradient is not a satisfactory substitute for $\mathscr{B}$ to be continuous and self-adjoint.

6. Construction of differential operators. We have been discussing evolution equations which contain a pair of nonlinear operators from a Hilbert space V to its dual V^*. In our applications the generalized solutions obtained in our theorems may satisfy natural or variational boundary conditions (e.g., of Neumann type) which are implicit in the functional identity

$$\frac{d}{dt}\mathscr{A}(u(t))+\mathscr{B}(u(t))\ni f(t) \tag{6.1}$$

in V^*. Such boundary conditions are classically recovered by Green's formula so we shall describe an appropriate extension of this formula which requires a minimum of regularity of the generalized solution. The objective is to resolve each term in (6.1) into two parts, a differential operator in distributions over a region Ω, the formal operator, and a constraint on the boundary Γ, the boundary operator. Then we briefly recall basic facts on Sobolev spaces and construct a rather general nonlinear operator $\mathscr{B}$ which will be used in the next section to illustrate theorems in some examples of initial-boundary-value problems.

Assume we are given a linear surjection $\gamma: V \to T$, called a "trace" operator, which is a strict homomorphism onto its range T, called "boundary values" of V. Let V_0 be the kernel of γ and note that the dual operator, $\gamma^*(g) = g \circ \gamma$, is an isomorphism of the dual space T^* onto the annihilator $V_0^\perp$ in V^*. Suppose there is given a continuous seminorm $|\cdot|$ on V for which V_0 is dense in the seminorm space $U \equiv \{V, |\cdot|\}$. Then we naturally identify U^* simultaneously as a subspace of V^* and of V_0^*.

We resolve the operator $\mathscr{A}: V \to 2^{V^*}$ into a formal part in V_0^* and a boundary part in T^*. For each $u \in D[\mathscr{A}]$ set $A_0(u) = \{F|_{V_0}: F \in \mathscr{A}(u)\}$, the set of restrictions to V_0 of functionals in $\mathscr{A}(u)$. Then set $D[\mathscr{A}_0] \equiv \{u \in V: A_0(u) \cap U^* \neq \phi\}$ and define $\mathscr{A}_0: V \to 2^{U^*}$ by $\mathscr{A}_0(u) = A_0(u) \cap U^*$. That is, $\mathscr{A}_0$ is the set of those functionals in $A_0(u)$ which have (unique) continuous extensions in $U^* \subset V^*$. Now let $u \in D[\mathscr{A}_0]$ and $F \in \mathscr{A}(u)$ with $F_0 = F|_{V_0} \in U^*$; hence, $F_0 \in \mathscr{A}_0(u)$. Then in $V_0^\perp$ we have $F - F_0 = \gamma^*(g)$ for a unique $g \in T^*$, so we can define $\partial_{\mathscr{A}}(u) \subset T^*$ to be the set of all such g. Thus, for each $F \in \mathscr{A}(u)$ for which $F_0 = F|_{V_0} \in U^*$, there is a unique $g \in T^*$ for which

$$F(v) = F_0(v) + g(\gamma v), \qquad v \in V,$$

and we indicate this by

$$\mathscr{A}(u) = \mathscr{A}_0(u) + \gamma^*(\partial_{\mathscr{A}}(u)), \qquad u \in D[\mathscr{A}_0]. \tag{6.2}$$

In our applications V_0^* is a space of distributions over Ω and T is the space of boundary values of the Sobolev space V, so (6.2) is the abstract *Green's formula* for the operator $\mathscr{A}$.

In many examples the solutions of (6.1) will have the additional regularity properties described below.

LEMMA 6.1. *Let $v \in H^1(0, T; V^*)$ with $v(t) \in \mathscr{A}(u(t))$ a.e. on $[0, T]$, and set $v_0(t) = v(t)|_{V_0}$ for each $t \in [0, T]$. Let $v_0(t) \in U^*$ and define $g(t) \in T^*$ by $v(t) = v_0(t) + \gamma^*(g(t))$ for $t \in [0, T]$. If $v_0'(t) \in U^*$ a.e. on $[0, T]$, then $g \in H^1(0, T; T^*)$ and*

$$v'(t) = v_0'(t) + \gamma^*(g'(t)), \quad \text{a.e. } t \in [0, T].$$

The preceding situation occurs, for example, in the case of linear symmetric $\mathscr{A}$ and in certain other special cases [2,], [9], [17], [25].

Suppose the operator $\mathscr{A}$ is given as above and let a second operator $\mathscr{B}: V \to 2^{V^*}$ be given. Resolve it likewise into two parts:

$$\mathscr{B}(u) = \mathscr{B}_0(u) + \gamma^*(\partial_{\mathscr{B}}(u)), \qquad u \in D[\mathscr{B}_0]. \tag{6.3}$$

Let there be given $f_0 \in L^2(0, T; U^*)$, $g_0 \in L^2(0, T; T^*)$, $v_0 \in R_g[A_0]$ and $g_0 \in T^*$ with $v_0 + \gamma^*(g_0) \in R_g[\mathscr{A}_0]$. Consider a solution of the Cauchy problem

$$\frac{d}{dt}\mathscr{A}(u(t)) + \mathscr{B}(u(t)) \ni f_0(t) + \gamma^*(g_0(t)), \quad \text{a.e. } t \in [0, T],$$

$$\mathscr{A}(u(0)) \ni v_0 + \gamma^*(g_0),$$

that is, a triple u, v, w for which

$$\begin{aligned} &v(t) \in \mathscr{A}(u(t)), \qquad w(t) \in \mathscr{B}(u(t)), \\ &v'(t) + w(t) = f_0(t) + \gamma^*(g_0(t)), \quad \text{a.e. } t \in [0, T], \\ &v(0) = v_0 + \gamma^*(g_0). \end{aligned} \tag{6.4}$$

By restricting the above functionals to V_0 we obtain

$$\begin{aligned} &v_0(t) \in A_0(u(t)), \qquad w_0(t) \in B_0(u(t)), \\ &v_0'(t) + w_0(t) = f_0(t) \quad \text{in } V_0^*, \quad \text{a.e. } t \in [0, T], \\ &v_0(0) = v_0. \end{aligned} \tag{6.5}$$

If Lemma 6.1 applies, then we obtain $w_0(t) \in U^*$ and the identities (6.2) and (6.3) give

$$\begin{aligned} &g_{\mathscr{A}}(t) \in \partial_{\mathscr{A}}(u(t)), \qquad g_{\mathscr{B}}(t) \in \partial_{\mathscr{B}}(u(t)), \\ &g_{\mathscr{A}}'(t) + g_{\mathscr{B}}(t) = g_0(t) \quad \text{in } T^*, \quad \text{a.e. } t \in [0, T], \\ &g_{\mathscr{A}}(0) = g_0. \end{aligned} \tag{6.6}$$

Thus (6.4) implies (6.5) and, in the situation of Lemma 6.1, also (6.6), so we call a solution of (6.4) a *weak solution* of the pair (6.5), (6.6). The first will give a partial differential equation and the second yields variational boundary conditions in our examples.

Let Ω be a bounded open set in R^n which lies locally on one side of its smooth boundary Γ. $H^1(\Omega)$ is the space of functions φ in $L^2(\Omega)$ for which each of the partial derivatives $D_j\varphi = \partial\varphi/\partial x_j$ belongs to $L^2(\Omega)$, $1 \leqq j \leqq n$. Letting D_0 denote the identity on $L^2(\Omega)$, we can express the norm on $H^1(\Omega)$ by

$$\|\varphi\|_{H^1(\Omega)} = \left(\sum_{j=0}^{n} \|D_j\varphi\|^2_{L^2(\Omega)} \right)^{1/2}.$$

We shall let V be a closed subspace of $H^1(\Omega)$ containing $C_0^\infty(\Omega)$ and let $\gamma: V \to L^2(\Gamma)$ be the indicated restriction to V of the trace map [19]. We let T be the range of γ (a subspace of $H^{1/2}(\Gamma)$) and denote the kernel by $V_0 \equiv H_0^1(\Omega)$. Since Γ is smooth there is a unit outward normal vector $n(s) = [n_1(s), \cdots, n_n(s)]$ at each point $s \in \Gamma$. Note that the test functions $C_0^\infty(\Omega)$ are dense in V_0 so the dual V_0^* is the space of (first order) distributions on Ω. We refer to (19) for information on these Sobolev spaces. Specifically, we shall use the trace operator between Sobolev spaces of fractional order.

We shall construct an operator $\mathcal{B}: V \to 2^{V^*}$ which will occur in many of our examples. For each integer k, $-1 \leqq k \leqq n$, let there be given a continuous, convex function $\psi_k; R \to R$ whose subgradient, $\beta_k \equiv \partial\psi_k$, satisfies

$$|w| \leqq C(|s|+1) \quad \text{if } w \in \beta_k(s), \quad s \in R, \quad -1 \leqq k \leqq n, \tag{6.7}$$

where C is some large constant. Then define $\psi: V \to R$ by

$$\psi(u) = \sum_{k=0}^{n} \int_\Omega \psi_k(D_k u(x))\, dx + \int_\Gamma \psi_{-1}(\gamma(u(s)))\, ds, \qquad u \in V.$$

From the estimates (6.7) it follows that ψ is a sum of continuous convex functions so we can compute its subgradient term by term. Recall that the subgradient F of the convex function $v \to \int_\Omega \varphi_k(v)\, dx$ at $w \in L^2(\Omega)$ is determined by $F(x) \in \beta_k(w(x))$, a.e. $x \in \Omega$. Since $D_k: V \to L^2(\Omega)$ is continuous linear, the subgradient of the convex function $v \to \int_\Omega \varphi_k(D_k v)\, dx$ at $u \in V$ is given by $\{D_k^* F: F \in \beta_k(D_k u)$ a.e.$\}$. See [11, pp. 26–28] and [1, p. 47] for proofs of these facts. These observations show that the subgradient of ψ is

$$\mathcal{B}(u) = \partial\psi(u) = \sum_{k=0}^{n} D_k^* \beta_k(D_k u) + \gamma^* \beta_{-1}(\gamma u), \qquad u \in V. \tag{6.8}$$

To be precise, we have $F \in \mathcal{B}(u)$ if and only if there exists $f_k \in \beta_k(D_k u)$ in $L^2(\Omega)$, $0 \leqq k \leqq n$, and $f_{-1} \in \beta_{-1}(\gamma u)$ in $L^2(\Gamma)$ for which

$$F(v) = \int_\Omega \sum_{k=0}^{n} f_k(x) D_k v(x)\, dx + \int_\Gamma f_{-1}(s) v(s)\, ds, \qquad v \in V.$$

By restricting the above to $v \in V_0 = H_0^1(\Omega)$ we see the formal part is the distribution

$$F|_{V_0} = -\sum_{k=1}^{n} D_k f_k + f_0 \in V_0^*.$$

We denote this by the equality (of sets)

$$B_0(u) = -\sum_{k=1}^{n} D_k \beta_k(D_k u) + \beta_0(u). \tag{6.9}$$

Let us interpret (6.3) with $U^* = L^2(\Omega)$. First, if $D_k f_k \in U^*$ for $1 \leqq k \leqq n$, then by the classical Green's theorem we have, from above,

$$F(v) - F|_{V_0}(v) = \int_\Gamma \left\{ \sum_{k=1}^{n} f_k(s) n_k(s) + f_{-1}(s) \right\} v(s)\, ds, \qquad v \in V.$$

Thus $u \in D(\mathcal{B})$ and we have shown

$$\sum_{k=1}^{n} f_k n_k + f_{-1} \in \partial_{\mathcal{B}}(u) \quad \text{with } f_k \in \beta_k(D_k u).$$

That is, when the terms are as regular as indicated we have

$$\partial_{\mathcal{B}}(u) = \sum_{k=1}^{n} \beta_k(D_k u) n_k + \beta_{-1}(u). \tag{6.10}$$

Furthermore, $\partial_{\mathcal{B}}(u)$ is defined without these regularity assumptions on the individual terms; it is sufficient to have $F|_{V_0} \in U^*$. Finally, we note that from (6.7) it follows that $\mathcal{B}$ satisfies the assumptions $[B_1]$ of Theorem 1 and $[B_3]$ of Theorem 3. It is also bounded from $L^2(0, T; V)$ to $L^2(0, T; V^*)$ and it will satisfy $[B_2]$ of Theorem 2 if, in addition,

there is a pair of numbers K, $c>0$ such that

$$\psi_k(s)\geqq c|s|^2-K,\ s\in R,\ 1\leqq k\leqq n \text{ and one of the following:}$$

(a) the estimate holds for $k=0$, or

(6.11) (b) the estimate holds for $k=-1$, or

(c) $v\in V$ and $v=$ constant imply $v\equiv 0$.

From (6.11) we can show that

$$\psi(v)\geqq c_1\|v\|_V^2-K_1,\qquad v\in V,$$

and this implies the coercivity condition in $[B_2]$.

7. Examples of partial differential equations. We shall describe some examples of initial-boundary-value problems for partial differential equations to illustrate the applications of our results. These examples were chosen merely to suggest a variety of problems that can be resolved by our Theorems, and they are not intended to be best possible in any sense.

(a) *Elliptic-parabolic equations.* For $k=0$ and -1, let $\varphi_k: R\to R$ be convex and continuous with subgradient, $\alpha_k\equiv\partial\varphi_k$, satisfying

$$|w|\leqq C(|s|+1)\quad \text{if } w\in\alpha_k(s),\quad s\in R.$$

Set $W=H^r(\Omega)$, $\frac{1}{2}<r<1$, $V=H^1(\Omega)$, and note that $V\to W$ is compact and $\gamma: W\to L^2(\Gamma)$ is continuous [19]. Thus we can define by

$$\varphi(v)\equiv\int_\Omega\varphi_0(v(x))\,dx+\int_\Gamma\varphi_{-1}(\gamma(v(s)))\,ds,\qquad v\in W,$$

a continuous and convex function $\varphi: W\to R$ with subgradient

$$\mathscr{A}(u)=\partial\varphi(u)=\alpha_0(u)+\gamma^*(\alpha_{-1}(\gamma u)),$$

bounded from W to W^*. That is, $F\in\mathscr{A}(u)$ if and only if there exist $f_0\in\alpha_0(u)$ in $L^2(\Omega)$ and $f_{-1}\in\alpha_{-1}(\gamma(u))$ in $L^2(\Gamma)$ for which

$$F(v)=\int_\Omega f_0(x)v(x)\,dx+\int_\Gamma f_{-1}(s)v(s)\,ds,\qquad v\in V, \tag{7.1}$$

so the formal and boundary parts of $\mathscr{A}$ are given, respectively, by

$$A_0(u)=\alpha_0(u),\qquad \partial_{\mathscr{A}}(u)=\alpha_{-1}(\gamma u). \tag{7.2}$$

From Theorem 2 we obtain the existence of a weak solution of the initial-boundary-value problem

$$\begin{aligned}
&\frac{\partial}{\partial t}A_0(u)+B_0(u)\ni f_0 \quad \text{in } L^2(0,T;H^{-1}(\Omega)),\\
&A_0u(0)\ni v_0,\\
&\frac{\partial}{\partial t}\partial_{\mathscr{A}}(u)+\partial_{\mathscr{B}}(u)\ni g_0 \quad \text{in } L^2(0,T;H^{-1/2}(\Gamma)),\\
&\partial_{\mathscr{A}}u(0)\ni b_0.
\end{aligned} \tag{7.3}$$

This is made precise in the form (6.5) and (6.6), where the operators are specified in (6.9), (6.10) and (7.2).

Remarks. By our choice of $V = H^1(\Omega)$, all boundary conditions in (7.3) are of variational type. Dirichlet-type constraints are obtained by taking subspaces of $H^1(\Omega)$.

We require that f_0 and g_0 be square-summable, with values in $H^{-1}(\Omega)$ and $H^{-1/2}(\Gamma)$ respectively, and we assume (6.11) to obtain coercivity of $\mathcal{B}$. The boundedness assumptions on α_k $(k = 0, -1)$ can be relaxed somewhat by using embedding theorems, e.g., of W into $L^p(\Omega)$.

There is no bound on the degeneracy permitted in the operator $\mathcal{A}$; we include even the (uninteresting) elliptic case $\mathcal{A} \equiv 0$. The case of $A_0 = 0$ leads to an evolution on the boundary subject to an elliptic equation in the interior; such problems arise from diffusion in a medium bounded by material of markedly lower diffusivity [25].

The classical porous-media equation and the weak form of the two-phase Stefan free-boundary problem are included in (7.3). In the latter, the enthalpy is given by $\alpha_0(s) = (1 + cH(s))s + LH(s)$, where $L > 0$ is the latent heat of fusion and $H(\cdot)$ is the Heaviside function [14], [16]. Such problems arise in welding, with the nonlinear term $\beta_0(u)$ representing a source of heat due to electrical resistance.

Note that each solution of (5.1) is also a (spatially independent) solution of (7.3), so there is much nonuniqueness in (7.3).

(b) *Pseudoparabolic equations.* Here we set $V = H_0^1(\Omega)$, so $T = \{0\}$ and all boundary conditions are of Dirichlet type. The operator $\mathcal{A}$ is given as above by (7.2); the operator $\mathcal{B}$ is also given as before but we shall only assume (6.7), not (6.11). On the space V we take the (equivalent) scalar product and corresponding Riesz map

$$\mathcal{R}u(v) = \int_\Omega \sum_{k=1}^{n} D_k u(x) D_k v(x)\, dx, \qquad u, v \in V,$$

so we have $\mathcal{R} = -\Delta_n \equiv -\sum_{k=1}^{n} D_k^2$. Assume $f_0 \in L^2(0, T; H^{-1}(\Omega))$ and $v_0 \in \alpha_0(u_0)$, $u_0 \in H_0^1(\Omega)$ are given. Then either from Theorem 1 or from Theorem 3 we obtain existence of a solution of the problem

$$\begin{aligned}
&u \in H^1(0, T; H_0^1(\Omega)), \qquad u(0) = u_0,\\
&v \in H^1(0, T; H^{-1}(\Omega)), \qquad v(0) = v_0,\\
&w \in L^2(0, T; H^{-1}(\Omega)),\\
&\frac{\partial}{\partial t}(v(t) - \Delta_n u(t)) + w(t) = f_0(t),\\
&v(t) \in A_0(u(t)), \qquad w(t) \in B_0(u(t)).
\end{aligned} \tag{7.4}$$

The operators A_0 and B_0 are given by (7.2) and (6.9) respectively.

Remarks. The partial differential equation in (7.4) is of the form of a nonlinear parabolic plus the term $(\partial/\partial t)\Delta_n u(x, t)$. Such equations are known to arise in various diffusion problems and are called *pseudoparabolic* [9], [15], [28]. Similar problems with variational boundary conditions can be considered; we obtain weak solutions in the form (6.4). However, since $R_g(A_0 + \mathcal{R}) = H^{-1}(\Omega)$, we cannot use Lemma 6.1, in general, to deduce (6.6). This situation occurs even in the linear case [26].

The operator $-\Delta_n$ in (7.4) can be replaced by the Riesz operator of any equivalent scalar product on $H_0^1(\Omega)$. This trivial observation is useful in introducing elliptic linear operators in its place.

We have not made use of the fact that only one of the operators $\mathcal{A}$, $\mathcal{B}$ need be a subgradient. In particular, we are free to add to one of $\mathcal{A}$ or $\mathcal{B}$ any linear combination of first order derivatives. (See Example (d) below.)

Nonuniqueness of solutions of (7.4) follows from that of solutions of (5.1).

In the preceding examples the nonlinearity arises from the *local* dependence on the solution, e.g., from nonlinear functions of the values of u or ∇u at each point of Ω. We next display examples of *global* nonlinearity arising from the "total energy" or the "total flux" in the system. The following preliminary result will be useful.

LEMMA 7.1. *Let $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be continuous, bilinear, symmetric and nonnegative real-valued functions on the Hilbert space V. Then for $\alpha, \beta \in R$, the function*

$$\varphi(u) \equiv \tfrac{1}{2} \max\{a(u,u)+\alpha, b(u,u)+\beta\}, \qquad u \in V$$

is convex, continuous and its subgradient is given by

$$\partial\varphi(u) = \begin{cases} \{A(u)\} & \text{if } a(u,u)+\alpha > b(u,u)+\beta, \\ \{(\lambda A + (1-\lambda)B)(u),\ 0 \leqq \lambda \leqq 1\} & \text{if } a(u,u)+\alpha = b(u,u)+\beta, \\ \{B(u)\}, & \text{if } a(u,u)+\alpha < b(u,u)+\beta, \end{cases}$$

where $Au(v) = a(u,v)$, $Bu(v) = b(u,v)$, $v \in V$.

Proof. We need only to compute $\partial\varphi(u)$. For the first and last cases we compute the Gateaux derivative $\lim_{t\to 0}\{(\varphi(u+tv)-\varphi(u))/t\}$ to obtain the desired results. Now assume $a(u,u)+\alpha = b(u,u)+\beta$. An easy computation gives

$$t^{-1}(\varphi(u+tv)-\varphi(u)) = \max\left\{a(u,v)+\frac{t}{2}a(v,v),\ b(u,v)+\frac{t}{2}b(v,v)\right\}$$

so we have the equivalence of $f \in \partial\varphi(u)$,

$$f(v) \leqq t^{-1}(\varphi(u+tv)-\varphi(u)), \qquad v \in V, \quad t>0,$$

and of

$$f(v) \leqq \max\{a(u,v), b(u,v)\}, \qquad v \in V.$$

This is equivalent to $f = \lambda Au + (1-\lambda)Bu$ for some λ, $0 \leqq \lambda \leqq 1$.

(c) *Energy-dependent elliptic-parabolic equation.* We shall use Theorem 2 with the operator $\mathcal{B}$ given by (6.8), so we assume (6.7) and (6.11). Choose $V = H^1(\Omega)$ so the space of boundary values is $T = H^{1/2}(\Gamma)$. Define on $W \equiv L^2(\Omega)$ the function

$$\varphi(u) = \frac{1}{2}\max\left\{1, \int_\Omega |u(x)|^2\,dx\right\}, \qquad u \in W.$$

The subgradient $\mathcal{A} = \partial\varphi$ is given by Lemma 7.1 and we have $\mathcal{A} = A_0 = \mathcal{A}_0$, $R_g(\mathcal{A}) = L^2(\Omega)$. Finally, let $v_0 \in L^2(\Omega)$, $f_0 \in L^2(0,T;L^2(\Omega))$, $g_0 \in L^2(0,T;H^{-1/2}(\Gamma))$ be given and define

$$f(t)(v) = \int_\Omega f_0(x,t)v(x)\,dx + g_0(t)(\gamma v), \qquad v \in V.$$

Then we obtain a weak solution of

$$\begin{aligned} &\frac{\partial v}{\partial t} + B_0(u) \ni f_0 \quad \text{in } L^2(0,T;H^{-1}(\Omega)), \\ &v(x,0) = v_0(x) \quad \text{in } L^2(\Omega), \\ &\partial_{\mathcal{B}}(u) \ni g_0 \quad \text{in } L^2(0,T;H^{-1/2}(\Gamma)), \end{aligned} \tag{7.5}$$

where v is determined by

$$v \in \begin{cases} \{0\}, & \text{if } \int_\Omega |u|^2\,dx < 1, \\ \{\lambda u : 0 \leqq \lambda \leqq 1\} & \text{if } \int_\Omega |u|^2\,dx = 1, \\ \{u\}, & \text{if } \int_\Omega |u|^2\,dx > 1. \end{cases}$$

Thus, the type of the equation is either elliptic (with parameter t) or parabolic and depends on the total energy $\int_\Omega |u|^2\,dx$.

(d) *A flux-dependent equation.* Take $V = H_0^1(\Omega)$, $W = L^2(\Omega)$ and $T = \{0\}$. Let the convex function φ_0 and its bounded subgradient $\alpha_0 = \partial\varphi_0$ be given as above in Example (a), and define $\mathscr{A} = \alpha_0$ in $L^2(\Omega)$; cf. (7.2). Denoting the gradient of u by $\vec{\nabla} u$, we define the continuous convex

$$\psi(u) = \frac{1}{2}\max\left\{N, \int_\Omega |\vec{\nabla} u(x)|^2\,dx\right\}, \qquad u \in V.$$

Let $\vec{b} \in R^n$ and define $\mathscr{B}: V \to 2^{V^*}$ by

$$\mathscr{B}(u) = \vec{b} \cdot \vec{\nabla} u + \partial\psi(u).$$

Note that $\mathscr{B}$ is maximal monotone, bounded and coercive. Let $v_0 \in R_g(\mathscr{A})$ and $f_0 \in L^2(0, T; H^{-1}(\Omega))$. From Theorem 2 we obtain existence of a solution of the problem

$$\begin{aligned} & u \in L^2(0, T; H_0^1(\Omega)), \quad v \in H^1(0, T; H^{-1}(\Omega)), \\ & \frac{\partial v}{\partial t} + \vec{b} \cdot \vec{\nabla} u - K\left(\int_\Omega |\vec{\nabla} u|^2\,dx\right)\Delta_n u = f_0, \\ & v(x, t) \in \alpha_0(u(x, t)), \qquad v(x, 0) = v_0(x), \end{aligned} \tag{7.6}$$

where the maximal monotone $K: R \to R$ is given by

$$K(s) = \begin{cases} \{0\}, & s < N, \\ [0, 1], & s = N, \\ \{1\}, & s > N. \end{cases}$$

Remarks. In the region where $\int_\Omega |\vec{\nabla} u|^2\,dx < N$ the equation in (7.6) is a *conservation law* of the form

$$\frac{\partial v}{\partial t} + \vec{b}\vec{\nabla} g(v) \ni f_0, \tag{7.7}$$

where the maximal monotone $g: R \to R$ is the inverse to α_0. Thus (7.6) suggests a *penalty method* [18] to approximate solutions of (7.7). We shall develop these observations elsewhere.

In order to consider (7.6) in the form (6.1) it is essential that $\mathscr{B}$ is not required to be a subgradient.

(e) *Elliptic-parabolic systems.* Our final example consists of a pair of equations of the type given above in Example (a) that are (nonlinearly) coupled. For $i = 0, 1$ and

$k=0,-1$, let $\varphi_k^i: R \to R$ be convex and continuous with subgradient, $\alpha_k^i \equiv \partial \varphi_k^i$, satisfying

$$|w| \leqq C(|s|+1) \quad \text{for } w \in \alpha_k^i(s),\ s \in R. \tag{7.8}$$

On the product space $W \equiv H^r(\Omega) \times H^r(\Omega)$, $\frac{1}{2} < r < 1$, we have the continuous trace operator $\gamma([u^1, u^2]) = [\gamma(u^1), \gamma(u^2)]$ which maps W into $L^2(\Gamma) \times L^2(\Gamma)$. Thus we define by

$$\varphi(v) = \sum_{i=1}^{2} \int_\Omega \varphi_0^i(v^i(x))\, dx + \sum_{i=1}^{2} \int_\Gamma \varphi_{-1}^i(\gamma v^i(s))\, ds, \qquad v = [v^1, v^2] \in W,$$

a continuous and convex function whose subgradient is given by

$$\mathscr{A}(u) \equiv \partial \varphi(u) = [\alpha_0^1(u^1) + \gamma^*(\alpha_{-1}^1(\gamma(u^1))),\ \alpha_0^2(u^2) + \gamma^*(\alpha_{-1}^2(\gamma(u^2)))],$$
$$u = [u_1, u_2] \in W.$$

The operator $\mathscr{A}: W \to 2^{W^*}$ is bounded; its formal and boundary parts are given, respectively, by (see (7.2))

$$A_0(u) = [\alpha_0^1(u^1), \alpha_0^2(u^2)], \qquad \partial_{\mathscr{A}}(u) = [\alpha_{-1}^1(\gamma(u^1)), \alpha_{-1}^2(\gamma(u^2))]. \tag{7.9}$$

Hereafter we restrict γ to the product space $V \equiv H^1(\Omega) \times H^1(\Omega)$. Assume we are given a set of continuous and convex functions $\psi_k^i: R \to R$ for $i = 1, 2$, $-1 \leqq k \leqq n$, whose subgradients $\beta_k^i \equiv \partial \psi_k^i$ all satisfy the estimate (6.7). For $i = 1, 2$ we define $\psi^i: H^1(\Omega) \to R$ as in § 6; its subgradient is then given by (see (6.8))

$$\mathscr{B}^i(u^i) = \partial \psi^i(u^i) = \sum_{k=0}^{n} D_k^* \beta_k^i(D_k u^i) + \gamma^* \beta_{-1}^i(\gamma u^i), \qquad u^i \in H^1(\Omega).$$

The formal and boundary parts of $\mathscr{B}^i$ are given by (6.9) and (6.10) for each of $i = 1, 2$. Thus we have two pairs of operators similar to the pair in Example (a). The coupling of the corresponding equations will be attained by a maximal monotone graph $\mu: R \to 2^R$ which is bounded, i.e., (7.8) holds for $w \in \mu(s)$. Then we define a maximal monotone operator M on $R \times R$ by

$$M([s_1, s_2]) = \{[w, -w]: w \in \mu(s_1 - s_2)\}, \qquad [s_1, s_2] \in R \times R.$$

This operator M induces a corresponding operator on $L^2(\Omega) \times L^2(\Omega)$, hence, from V into V^*, which we also denote by M. Finally we define

$$\mathscr{B}([u_1, u_2]) = [\mathscr{B}^1(u_1), \mathscr{B}^2(u_2)] + M(u_1, u_2), \qquad [u_1, u_2] \in V.$$

This $\mathscr{B}$ is the sum of maximal monotone operators, each of which is defined on all of V, so $\mathscr{B}$ is maximal monotone. Similarly $\mathscr{B}$ is bounded, and we note that $\mathscr{B}$ is coercive if both of $\mathscr{B}^1$ and $\mathscr{B}^2$ are coercive.

Assume that we are given the following data:

$$f_0^i \in L^2(0, T; H^{-1}(\Omega)), \qquad g_0^i \in L^2(0, T; H^{-1/2}(\Gamma)), \qquad i = 1, 2,$$
$$[v_0^1, v_0^2] \in R_g(A_0), \qquad (v_{-1}^1, v_{-1}^2] \in R_g(\partial_{\mathscr{A}}).$$

If the functions $\{\beta_k^i: -1 \leqq k \leqq n\}$ satisfy (6.11) for both $i=1$ and $i=2$, then from Theorem 2 it follows that there exists a weak solution of the system

$$\frac{\partial}{\partial t}\alpha_0^1(u^1(x,t))+B_0^1(u^1(x,t))+\mu(u^1(x,t)-u^2(x,t)) \ni f_0^1(x,t),$$

$$\frac{\partial}{\partial t}\alpha_0^2(u^2(x,t))+B_0^2(u^2(x,t))-\mu(u^1(x,t)-u^2(x,t)) \ni f_0^2(x,t)$$

$$\text{in } L^2(0,T;H^{-1}(\Omega)), \tag{7.10}$$

$$\frac{\partial}{\partial t}\alpha_{-1}^i(\gamma u^i(s,t))+\partial_{\mathscr{B}^i}(u^i(s,t)) \ni g_0^i, \qquad i=1,2, \quad \text{in } L^2(0,T;H^{-1/2}(\Gamma)),$$

$$\alpha_0^i(u^i(x,0)) \ni v_0^i(x), \qquad i=1,2, \quad \text{in } L^2(\Omega),$$

$$\alpha_{-1}^i(\gamma u^i(s,0)) \ni v_{-1}^i(s), \qquad i=1,2, \quad \text{in } L^2(\Gamma).$$

Remarks. All of the operators in this system are (possibly) multi-valued, so each of the "equations" should be made precise as was done in our preceding examples. See (6.9), (6.10), (7.2) and (7.9) for related computations.

The only requirement on the α_k^i is that they be maximal monotone graphs in R which satisfy the bound (7.8). Thus much degeneracy is possible in the leading operator given by (7.9). Related Stefan-type free-boundary problems can be so considered.

Interesting examples of the coupling term arise in applications to diffusion problems. These include problems with a *semipermeable* membrane, $\mu(s)=s^+$ (where s^+ denotes s if $s>0$ and 0 otherwise), or those with a *threshold* phenomenon $\mu(s)=(s-\varepsilon)^+-(-s-\varepsilon)^+$. The operator M as given above is a subgradient; this is easily verified by showing it is cyclic monotone [1]. However we may add to M nonsymmetric monotone terms, for example, $[-s_2, s_1]$, and thereby obtain systems of the form (6.1) in which $\mathscr{B}$ is not a subgradient.

Systems of equations of *pseudoparabolic* type can be resolved similarly by Theorem 1. For example, we can choose $V=H_0^1(\Omega)\times H_0^1(\Omega)$ with scalar product on each factor as given in Example (b), and obtain existence of a solution of the problem

$$\frac{\partial}{\partial t}(\alpha_0^1(u^1(x,t))-\Delta_n u^1(x,t))+B_0^1(u^1(x,t))+\mu(u^1(x,t)-u^2(x,t)) \ni f_0^1(x,t),$$

$$\frac{\partial}{\partial t}(\alpha_0^2(u^2(x,t))-\Delta_n u^2(x,t))+B_0^2(u^2(x,t))-\mu(u^1(x,t)-u^2(x,t)) \ni f_0^2(x,t) \tag{7.11}$$

$$\text{in } L^2(0,T;H^{-1}(\Omega)),$$

$$u^j \in H^1(0,T;H_0^1(\Omega)), \quad u^j(x,0)=u_j(x), \quad \alpha_0^j(u^j(x,0)) \ni v_j(x)$$

$$j=1,2, \quad \text{in } L^2(\Omega)$$

where the data are given as above with $v_j \in A_0(u_j)$ for $j=1,2$.

REFERENCES

[1] H. BRÉZIS, *Opérateurs maximaux monotones et semigroupes de contraction dans les espaces d'Hilbert.* Math. Studies 5, North-Holland/American Elsevier, New York, 1973.

[2] ———, *On some degenerate non-linear parabolic equations.* Proc. Amer. Math. Soc. XVIII, 1968.

[3] ———, *Monotonicity methods in Hilbert space and some applications to nonlinear partial differential equations.* Proc. Symposium by the Mathematics Research Center, Madison, Wisconsin, Academic Press, New York, 1971.

[4] V. BARBU, *Existence for non-linear Volterra equations in Hilbert space*, this journal, 10 (1979), pp. 552–569.
[5] F. E. BROWDER, *Non linear operators in Banach space*, Proc. Amer. Math. Soc., XVII, 2, 1968.
[6] B. CALVERT, *The equation* $A(t, u(t))' + B(t, u(t)) = 0$, Math. Proc. Phil. Soc., 79 (1976), pp. 545–562.
[7] J. R. CANNON AND C. D. HILL, *On the movement of a chemical reaction interface*, Indiana Univ. Math. J., 20 (1970), pp. 429–454.
[8] J. R. CANNON, W. T. FORD AND A. V. LAIR, *Quasilinear parabolic systems*, J. Differential Equations, 20 (1976), pp. 441–472.
[9] R. W. CARROL AND R. E. SHOWALTER, *Singular and degenerate Cauchy problems*, Mathematics in Science and Engineering, Vol. 27, Academic Press, New York, 1976.
[10] M. G. CRANDALL AND A. PAZY, *Semigroupes of non-linear contractions and dissipative sets*, J. Funct. Anal., 3 (1969), pp. 376–418.
[11] I. EKELAND AND R. TEMAM, *Analyse convex et problèmes variationnelles.* Dunod Gauthier-Villars, Paris, 1974.
[12] O. GRANGE AND F. MIGNOT, *Sur la resolution d'une equation et d'une inequation paraboliques non-lineares.* J. Funct. Anal., 11 (1972), pp. 77–92.
[13] A. FRIEDMAN, *The Stefan problem in several space variables.* Trans. Amer. Math. Soc., 132 (1968), pp. 51–87.
[14] ———, *Partial Differential Equations of Parabolic Type*, Prentice Hall, Englewood Cliffs, NJ, 1964.
[15] M. E. GURTIN AND P. J. CHEN, *On a theory of heat conduction involving two temperatures.* Z. Angew. Math. Phys., 19 (1968), 614–627.
[16] O. A. LADYZENSKAJIA, V. A. SOLONNIKOV AND N. N. URAL'TZEVA, *Linear and quasi-linear equations of parabolic type*, AMS Translation Monograph 23, American Mathematical Society, Providence, RI, 1968.
[17] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites nonlinéaires.* Dunod, Paris, 1969.
[18] ———, *Sur quelques question d'analyse de mechanique et de contrôle optimal.* Les Presses de l'Universitè de Montreal, Montreal, 1976.
[19] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogenes et applications*, Vol. I, Dunod, Paris, 1962.
[20] G. MINTY, *Monotone (non linear) operators in a Hilbert space.* Duke Math. J., 29 1962, pp. 341–346.
[21] ———, *On the monotonicity of the gradient of a convex function*, Pacific J. Math., 14 (1964), pp. 243–242.
[22] ———, *A theorem on maximal monotone sets in Hilbert space*, J. Math. Anal. Appl., 11 (1965), pp. 437–439.
[23] R. E. SHOWALTER, *Existence and representation theorems for a semilinear Sobolev equation in Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 527–543.
[24] ———, *Degenerate evolution equations and applications*, Indiana Univ. Math. J., 23 (1974), pp. 655–677.
[25] ———, *Nonlinear degenerate evolution equations and partial differential equations of mixed type*, SIAM J. Math. Anal. 6 (1975), pp. 25–42.
[26] ———, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
[27] G. STAMPACCHIA, *On some regular multiple integral problems in the calculus of variations*, Comm. Pure Appl. Math., 16 (1963), pp. 383–421.
[28] VON REINHARD KLUGE AND G. BRUCKNER, *Uber einige klassen nichtlinearer differential-gleichungen und ungleichungen im Hilbert Raum.* Math. Nachr., 64 (1974), pp. 5–32.
[29] K. YOSHIDA, *Functional Analysis*, 4th ed., Springer-Verlag 123 (1974).

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0007 $01.00/0

POSITIVITY OF WEIGHTED WIGNER DISTRIBUTIONS*

A. J. E. M. JANSSEN†

Abstract. In [4] a number of inequalities involving Wigner distributions and their moments are given. The present paper gives theorems on the positivity of weighted Wigner distributions, where the weight function is assumed to be radially symmetric. The main tool is a formula expressing weighted Wigner distributions of a function in terms of its Hermite coefficients and certain integrals involving Laguerre polynomials.

1. Introduction. If $f \in L^2(\mathbb{R})$, $g \in L^2(\mathbb{R})$, then the (mixed) Wigner distribution of f and g is defined by

$$W(x, y; f, g) = \int_{-\infty}^{\infty} e^{-2\pi i y t} f(x + \tfrac{1}{2}t) \overline{g(x - \tfrac{1}{2}t)}\, dt$$

for $x \in \mathbb{R}$, $y \in \mathbb{R}$. If $f = g$, then we call $W(x, y; f, g)$ the *Wigner distribution* of f. Note that $W(x, y; f, g)$ is continuous, bounded and square integrable over $\mathbb{R}^2$. See, e.g., [4, § 2], [5, §§ 12, 13, 14, 15] and [12, § 4] for general information about the Wigner distribution; we like to refer in particular to the first reference, where an interpretation of the Wigner distribution as an energy density function in time and frequency is given. The Wigner distribution can also be defined for tempered distributions, or for generalized functions of the class S^* associated with the space S of smooth functions (an entire function f is called smooth if there are $A > 0$, $B > 0$ such that $f(x+iy) = O(\exp(-\pi A x^2 + \pi B y^2))$; cf. [5, §§ 2, 27] and [10, Appendix 3, § 2]). We shall derive our formulas and inequalities for functions $f \in S$ only (the Wigner distribution of such an f depends smoothly on its two variables; cf. [5, § 7, 13]); in many cases, however, the result considered can be proved to hold in S^* (and in particular in $L^2(\mathbb{R})$) as well by noting that the space S is S^*-dense in S^* (cf. [5, §§ 17, 18]).

An important role is played by the Hermite functions $\psi_n (n = 0, 1, \cdots)$. We take the same normalization as in [4, § 3] and [5, § 27, 6.3], i.e. we have for $x \in \mathbb{C}$, $w \in \mathbb{C}$

$$\exp(\pi x^2 - 2\pi(x - w)^2) = \sum_{n=0}^{\infty} c_n w^n \psi_n(x),$$

$$c_n = (n!)^{-1/2} 2^{-1/4} (4\pi)^{n/2}, \qquad n = 0, 1, \cdots.$$

The Laguerre polynomials $L_n (n = 0, 1, \cdots)$ of zeroth order are given by (cf. [14, 5.1.6])

$$L_n(x) = \sum_{j=0}^{n} \binom{n}{j} (-x)^j / j!, \qquad n \geqq 0.$$

We have (cf. [14, § 5.1])

$$(1 - w)^{-1} \exp(-xw(1 - w)^{-1}) = \sum_{n=0}^{\infty} w^n L_n(x)$$

for $x \geqq 0$, $w \in \mathbb{C}$, $|w| < 1$.

We note that every $f \in S$ has an expansion $\sum_{n=0}^{\infty} (f, \psi_n)\psi_n$, where the series converges in S-sense to f (cf. [5, § 27, 6.3 and § 23]).

* Received by the editors August 12, 1980, and in revised form December 2, 1980.
† California Institute of Technology, Pasadena, California 91125.

2. A formula for weighted Wigner distributions. This section gives a formula for weighted Wigner distributions, where the weight functions are assumed to be radially symmetric. The methods for deriving this formula are essentially the same as the ones in [4, § 4]. The starting point, a formula for the (mixed) Wigner distribution of Hermite functions is from Groenewold (cf. [9, (5.16)]). For the sake of completeness, we include a proof.

2.1. LEMMA. *Let $n = 0, 1, \cdots$, $m = 0, 1, \cdots$, and let c_n, c_m be as in the Introduction. For $x \in \mathbb{R}$, $y \in \mathbb{R}$ we have*

$$c_n c_m W(x, y; \psi_n, \psi_m) = 2^{1/2} \exp(-2\pi|z|^2) \sum_{j=0}^{\min(n,m)} \frac{(-4\pi)^j}{j!} \frac{(4\pi\bar{z})^{n-j}}{(n-j)!} \frac{(4\pi z)^{m-j}}{(m-j)!},$$

where $z = x + iy$.

Proof. Abbreviating "the coefficient of $v^n w^m$ in" by $C_{v^n w^m}$, we get

$$\begin{aligned} &c_n c_m W(x, y; \psi_n, \psi_m) \\ &= C_{v^n w^m} \int_{-\infty}^{\infty} \exp(-2\pi i y t + \pi(x + \tfrac{1}{2}t)^2 \\ &\qquad -2\pi(x + \tfrac{1}{2}t - v)^2 + \pi(x - \tfrac{1}{2}t)^2 - 2\pi(x - \tfrac{1}{2}t - w)^2)\, dt \end{aligned}$$

for $x \in \mathbb{R}$, $y \in \mathbb{R}$. This can be written as

$$2^{1/2} \exp(-2\pi|z|^2) C_{v^n w^m} \exp(-4\pi v w + 4\pi\bar{z}v + 4\pi z w).$$

The lemma follows on expanding $\exp(-4\pi v w + 4\pi\bar{z}v + 4\pi z w)$ as

$$\sum_{j=0}^{\infty} \frac{(-4\pi v w)^j}{j!} \sum_{k=0}^{\infty} \frac{(4\pi\bar{z}v)^k}{k!} \sum_{l=0}^{\infty} \frac{(4\pi z w)^l}{l!}. \qquad \square$$

2.2. Taking $n = m$ in the lemma, we get

$$W(x, y; \psi_n, \psi_n) = (-1)^n 2 \exp(-2\pi|z|^2) L_n(4\pi|z|^2).$$

In [11, formula 17], a similar expression is given for what is called the ambiguity function of ψ_n (cf. also [13], where weighted integrals of ambiguity functions are considered). This is no surprise since ambiguity function and Wigner distribution are related by a double Fourier transformation (cf. [10, Appendix 3, 1.2(v)]).

2.3. THEOREM. *Let $K : [0, \infty) \to \mathbb{C}$ be a measurable function such that $\int_0^\infty |K(x)|^2 \exp(-\varepsilon x)\, dx < \infty$ for all $\varepsilon > 0$. Then we have for $f \in S$, where $\iint$ indicates integration over $\mathbb{R}^2$,*

$$\begin{aligned} &\iint W(x, y; f, f) K(2\pi(x^2 + y^2))\, dx\, dy \\ &\qquad = \sum_{n=0}^{\infty} (-1)^n |(f, \psi_n)|^2 \int_0^\infty e^{-r} K(r) L_n(2r)\, dr. \end{aligned}$$

Proof. The double integral converges absolutely, and so does the series at the right-hand side (cf. [5, § 27, 6.3] and [16, Chapt. VI, § 6]). We have (cf. Introduction)

$$W(x, y; f, f) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} (f, \psi_n) \overline{(f, \psi_m)} W(x, y; \psi_n, \psi_m)$$

for $x \in \mathbb{R}$, $y \in \mathbb{R}$. If we multiply this formula by $K(2\pi(x^2 + y^2))$ and integrate over $\mathbb{R}^2$, then the terms in the series with $n \neq m$ cancel by the lemma (use polar coordinates).

Hence

$$\iint W(x, y; f, f)K(2\pi(x^2+y^2))\,dx\,dy$$

$$= \sum_{n=0}^{\infty} |(f, \psi_n)|^2 \iint W(x, y; \psi_n, \psi_n)K(2\pi(x^2+y^2))\,dx\,dy,$$

and the theorem follows easily from § 2.2. □

3. Applications. We use the theorem to show that certain weighted integrals of Wigner distributions are nonnegative. We consider weight functions with radial symmetry around the point (0, 0), but, as in [4], our results can be extended to weight functions with elliptic symmetry around any point (a, b) in the plane. We are especially interested in nonnegative weight functions.

3.1. The following lemma and examples show that alternating series involving Laguerre polynomials are often nonnegative.

LEMMA. *Assume that* $a_n \downarrow 0$, $\sum_{n=0}^{\infty} a_n^2 \leqq \infty$. *Then* $\sum_{n=0}^{\infty} (-1)^n a_n L_n(x) \geqq 0$ *a.e.*

Proof. From orthonormality of the functions $e^{-x/2}L_n(x)$ on $(0, \infty)$ (cf. [14, (5.1.1)]), we see that $\sum_{n=0}^{\infty} (-1)^n a_n L_n(x)$ converges locally in L^2-sense. Now $S_n(x) = \sum_{k=0}^{n} (-1)^k L_k(x) \geqq 0$ for all $x \geqq 0$, $n = 0, 1, \cdots$ by [14, Problem Section, Problem 100]. We get by partial summation for $N = 0, 1, \cdots$,

$$\sum_{n=0}^{N} (-1)^n a_n L_n(x) = \sum_{n=0}^{N} S_n(x)(a_n - a_{n+1}) + a_{N+1}S_{N+1}(x).$$

It follows from [15, Chapt. IV, § 7, 1b, d] that $S_N(x)$ is bounded in N for every $x \geqq 0$. The lemma follows by letting $N \to \infty$. □

COROLLARY. *If* $K(r) = \sum_{n=0}^{\infty} (-1)^n a_n e^{-r} L_n(2r)$, *where the* a_n's *are as in the lemma, then* $K \geqq 0$, *and we have for* $f \in S$

$$\iint W(x, y; f, f)K(2\pi(x^2+y^2))\,dx\,dy = \frac{1}{2}\sum_{n=0}^{\infty} a_n |(f, \psi_n)|^2.$$

3.2. In many cases the positivity of $\sum_{n=0}^{\infty} (-1)^n a_n L_n(x)$ can also be established by employing identities for Laguerre polynomials. In the list of examples below we assume $f \in S$.

(i) Taking $-1 < w < 0$,

$$K(r) = \sum_{n=0}^{\infty} (-1)^n |w|^n e^{-r} L_n(2r)$$

$$= (1-w)^{-1} \exp(-(1+w)(1-w)^{-1}r),$$

we find that

$$(1-w)^{-1} \iint W(x, y; f, f) \exp\left(-2\pi \frac{1+w}{1-w}(x^2+y^2)\right) dx\,dy = \frac{1}{2}\sum_{n=0}^{\infty} |(f, \psi_n)|^2 |w|^n.$$

Cf. also [4, Thm. 4.2]. We can also use $\int_0^{\infty} e^{-st} L_n(t)\,dt = s^{-n-1}(s-1)^n$ with $0 < s \leqq 1$ (cf. [14, Problem Section, Problem 19]) to derive the same formula.

(ii) We have $\sum_{n=0}^{\infty} w^n L_n(x)/n! = \exp(w) J_0(2i|w|^{1/2}x)$ (cf. [6, § 5]), where J_0 is the Bessel function of first kind and zeroth order (cf. [14, Chapt. I, 1.71]). Note that

$J_0(it) = \sum_{\nu=0}^{\infty} (t/2)^{2\nu}/(\nu!)^2 > 0$ for $t \geqq 0$. By taking the appropriate K we get for $w < 0$

$$e^w \iint W(x, y; f, f) J_0(8\pi i|w|^{1/2}(x^2+y^2)) \exp(-2\pi(x^2+y^2))\, dx\, dy$$

$$= \frac{1}{2} \sum_{n=0}^{\infty} |(f, \psi_n)|^2 \frac{|w|^n}{n!}.$$

(iii) We have $x^\alpha = \sum_{n=0}^{\infty} (-1)^n \Gamma^2(\alpha+1) L_n(x)/n!\Gamma(\alpha-n+1)$ for $\alpha > -1$ (cf. [16, Chapt. VI, § 8]). By taking the appropriate K we get for $\alpha > -1$

$$(4\pi)^\alpha \iint W(x, y; f, f)(x^2+y^2)^\alpha \exp(-2\pi(x^2+y^2))\, dx\, dy$$

$$= \frac{1}{2} \sum_{n=0}^{\infty} |(f, \psi_n)|^2 \Gamma^2(\alpha+1)/n!\Gamma(\alpha-n+1).$$

In case α is an integer, the sum is a finite one, and if α is not an integer, the sum contains non-positive terms (viz. the terms with $n = [\alpha]+2, [\alpha]+4, \cdots$).

(iv) We have $2^n n! \sum_{k=0}^{n} (-1)^k L_k(2x^2+2y^2) = \sum_{k=0}^{n} \binom{n}{k} H_k^2(x) H_{n-k}^2(y)$, where the H_k's are the Hermite polynomials (cf. [14, Problem Section, Problem 100]). We get by taking the appropriate K

$$\frac{2^{-n}}{n!} \iint W(x, y; f, f) \sum_{k=0}^{n} \binom{n}{k} H_k^2(x\sqrt{2\pi}) H_{n-k}^2(y\sqrt{2\pi}) \exp(-2\pi(x^2+y^2))\, dx\, dy$$

$$= \frac{1}{2} \sum_{k=0}^{n} |(f, \psi_k)|^2.$$

(v) We have $(-1)^{k+m+n} \int_0^\infty L_k(x) L_m(x) L_n(x) e^{-x}\, dx \geq 0$ for $k, m, n = 0, 1, \cdots$ (cf. [14, Problem Section, Problem 94]). Let $k = 0, 1, \cdots$, and take $K(r) = e^{-r} L_k^2(2r)$. We get

$$\iint W(x, y; f, f) L_k^2(4\pi(x^2+y^2)) \exp(-2\pi(x^2+y^2))\, dx\, dy$$

$$= \frac{1}{2} \sum_{n=0}^{\infty} |(f, \psi_n)|^2 (-1)^n \int_0^\infty e^{-r} L_k^2(r) L_n(r)\, dr$$

and the terms in the right-hand side series are nonnegative.

(vi) We have

$$L_n^2(x) = 2^{-2n} \sum_{k=0}^{n} \binom{2k}{k} \binom{2(n-k)}{n-k} L_{2k}(2x)$$

(cf. [14, Problem Section, Problem 101]). We derive

$$\iint W(x, y; f, f) L_n^2(2\pi(x^2+y^2)) \exp(-2\pi(x^2+y^2))\, dx\, dy$$

$$= \sum_{k=0}^{n} |(f, \psi_{2k})|^2 \binom{2k}{k} \binom{2(n-k)}{n-k} 2^{-2n}.$$

(vii) In [2] it is claimed and in [7] it is proved that

$$\sum_{k=0}^{n} (-1)^k (\lambda+1)_{n-k} (\lambda+1)_k L_k(x)/k!(n-k)! \geqq 0$$

for $\lambda \geqq 0$, $x \geqq 0$. Here $(a)_0 := 1$, $(a)_k := a(a-1)\cdot\cdots\cdot(a+k-1)$ for $a \in \mathbb{R}$, $k = 1, 2, \cdots$. For example, if we multiply by $n!\lambda^{-n}$ and let $\lambda \to \infty$, we get

$$\iint W(x, y; f, f) \exp(-2\pi(x^2+y^2)) K(4\pi(x^2+y^2))\, dx\, dy = \frac{1}{2}\sum_{k=0}^{n} \binom{n}{k} |(f, \psi_k)|^2,$$

where $K(r) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} L_k(r) \geqq 0$, $r \geqq 0$.

(viii) We have $\int_0^\infty x^\mu e^{-x/2} L_n(x)\, dx = 2^\mu \Gamma(\mu+1) b_n$, where b_n is the coefficient of r^n in $(1-r)^\mu (1+r)^{-\mu-1}$ for $\mu > -1$, $n = 0, 1, \cdots$ (cf. [14, (9.5.20) and (9.5.21)]). It can be shown that $(-1)^n b_n \geqq 0$ for $n = 0, 1, \cdots$ and $\mu \geqq -\frac{1}{2}$. Indeed, if we put $(1+r)^\mu (1-r)^{-\mu-1} = \sum_{n=0}^{\infty} a_n(\mu) r^n$, then a_n is a polynomial in μ, and $a_{2n}(-\frac{1}{2}) = 2^{-2n}\binom{2n}{n}$, $a_{2n+1} = 0$ for $n = 0, 1, \cdots$. For the derivatives of a_n with respect to μ in the point $-\frac{1}{2}$ we have

$$\begin{aligned} a_n^{(k)}\left(-\frac{1}{2}\right) &= C_{r^n}\left(\frac{d}{d\mu}\right)^k (1+r)^\mu (1-r)^{-\mu-1}\Big|_{\mu=-\frac{1}{2}} \\ &= C_{r^n}\left(\log\frac{(1+r)}{(1-r)}\right)^k \Big/ \sqrt{1-r^2} \geqq 0, \end{aligned}$$

where C_{r^n} abbreviates "coefficient of r^n in", as

$$\log\frac{(1+r)}{(1-r)} = 2\sum_{l=0}^{\infty} \frac{r^{2l}}{(2l+1)}, \quad \frac{1}{\sqrt{1-r^2}} = \sum_{m=0}^{\infty} a_m\left(-\frac{1}{2}\right) r^m.$$

It follows that $a_n(\mu) \geqq 0$ for $n = 0, 1, \cdots$, $\mu \geqq -\frac{1}{2}$.

Taking $K(r) = (2r)^\mu$ we get

$$\begin{aligned} &(4\pi)^\mu \iint W(x, y; f, f)(x^2+y^2)^\mu\, dx\, dy \\ &\qquad = \frac{1}{2}\sum_{n=0}^{\infty} (-1)^n |(f, \psi_n)|^2 \int_0^\infty e^{-r/2} r^\mu L_n(r)\, dr \end{aligned}$$

for $\mu > -1$. This is nonnegative for $\mu \geqq -\frac{1}{2}$ (even for noninteger values of μ; compare (iii)). For $\mu = -\frac{1}{2}$ we get

$$(4\pi)^{-1/2} \iint W(x, y; f, f)(x^2+y^2)^{-1/2}\, dx\, dy = \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty} \binom{2n}{n} 2^{-4n} |(f, \psi_{2n})|^2.$$

We note that in [4, Thm. 4.3] the case that μ is a positive integer is considered.

3.3. If f is an even function, then sharper results than the above ones can be obtained. This is no surprise since $W(0, 0; f, f) \geqq 0$ if f is even. Note that for an even f the terms in the series of the theorem in § 2.3 with odd index cancel (as $(f, \psi_{2n+1}) = 0$ for $n = 0, 1, \cdots$). Hence we can use now functions K that have an expansion in Laguerre functions with nonnegative even coefficients only. So assume that $f \in S$ is even.

(i) In § 3.2(ii) we can take $-1 < w < 1$. We thus find that

$$\iint W(x, y; f, f) \exp(-\pi\delta(x^2+y^2))\, dx\, dy \geqq 0$$

for all $\delta \geqq 0$. This follows, of course, also from the formula for $H_{\alpha\alpha}(0, 0; f, f)$ in the proof of [4, Thm. 4.2].

(ii) We have $\int_0^\infty L_k(x)L_m(x)L_n(x)x^{-1/2}\,e^{-3x/2}\,dx \geqq 0$ for $k, m, n = 0, 1, \cdots$ (cf. [1, 3.4]). Hence, for $k = 0, 1, \cdots$,

$$(4\pi)^{-1/2}\iint W(x, y; f, f)(x^2+y^2)^{-1/2}\exp(-4\pi(x^2+y^2))L_k^2(4\pi(x^2+y^2))\,dx\,dy$$

$$= \sum_{n=0}^{\infty} |(f, \psi_{2n})|^2 \int_0^\infty e^{-3r}r^{-1/2}L_k^2(2r)L_{2n}(2r)\,dr \geqq 0,$$

and the integrals in the right-hand side series are nonnegative.

(iii) It follows from a result of Askey and Gasper (cf. [8, (9.7)]) that

$$\int_0^\infty L_k(x)L_m(x)L_n(x)\,e^{-\rho x}\,dx \geqq 0$$

for $\rho \geqq 2$, $k, m, n = 0, 1, \cdots$. Hence, for $k = 0, 1, \cdots$,

$$\iint W(x, y; f, f)\exp(-2\pi(\rho+1)(x^2+y^2))L_k^2(4\pi(x^2+y^2))\,dx\,dy$$

$$= \sum_{n=0}^{\infty} |(f, \psi_{2n})|^2 \int_0^\infty e^{-2\rho r}L_k^2(2r)L_{2n}(2r)\,dr \geqq 0,$$

and the integrals in the right-hand side series are non-negative.

(iv) The coefficient of r^{2n} in $(1-r)^\mu(1+r)^{-\mu-1}$ is positive for all $n = 0, 1, \cdots$ and all $\mu \in \mathbb{R}$. For $\mu \geqq -\frac{1}{2}$ this follows from 3.2 (viii), and for $\mu \leqq -\frac{1}{2}$ we can use the fact that the coefficients of r^{2n} in $(1-r)^\mu(1+r)^{-\mu-1}$ and $(1-r)^\sigma(1+r)^{-\sigma-1}$ are the same if $\sigma = -\mu - 1$. This implies (cf. 3.2 (viii)) that

$$\iint W(x, y; f, f)(x^2+y^2)^\mu\,dx\,dy \geqq 0$$

for all $\mu > -1$.

3.4. We ask ourselves how restrictive the condition $a_n \geqq 0$ for a function $K(r) = \sum_{n=0}^\infty (-1)^n a_n e^{-r}L_n(2r)$ is. It follows from 3.2(iii) and Parseval's formula that

$$L(\alpha) := 2^\alpha \int_0^\infty x^\alpha\, e^{-x}K(x)\,dx = \sum_{n=0}^{\infty} a_n \frac{\Gamma^2(\alpha+1)}{n!\Gamma(\alpha-n+1)}$$

for every $\alpha > -1$. Suppose $K(x) = O(e^{-mx})$ for some $m > 1$. Then $L(\alpha) = O(2^\alpha\alpha!/(m+1)^\alpha)$ if $\alpha \to \infty$. Taking integer values for α (so that the series contains nonnegative terms only), we get $L(\alpha) \geqq \alpha!\alpha(\alpha-1)\cdots(\alpha-n+1)a_n/n!$ for all n. It easily follows that $a_n = 0$ for all n, whence $K \equiv 0$. In particular, a K with $a_n \geqq 0$ cannot have a compact support.

In a similar way one can show from 3.2(iii) that

$$W(x, y; f, f) \neq O(\exp(-\pi A(x^2+y^2))$$

if $f \not\equiv 0$, $A > 2$.

3.5. We finally note a stability property of the class U of all $K: [0, \infty) \to [0, \infty)$ for which $(-1)^n \int_0^\infty e^{-x}L_n(2x)K(x)\,dx \geqq 0$. Assuming proper integrability conditions, we put $\hat{K}(x, y) := K(2\pi(x^2+y^2))$ for $K \in U$, and $\hat{K}_1 * \hat{K}_2$ for the ordinary convolution of $\hat{K}_1$ and $\hat{K}_2$. Now note that $\hat{K}_1 * \hat{K}_2$ depends on $2\pi(x^2+y^2)$ only, and that for every $f \in S$,

if $K_1 \in U$, $K_2 \geqq 0$,

$$\iint W(x, y; f, f)(\hat{K}_1 * \hat{K}_2)(x, y)\, dx\, dy$$

$$= \iint \hat{K}_2(a, b)\left(\iint W(x+a, y+b)\hat{K}_1(x, y)\, dx\, dy\right) da\, db \geqq 0$$

by the remark in the beginning of this section. Theorem 2.3 shows that $\hat{K}_1 * \hat{K}_2 = \hat{K}_3$ for some $K_3 \in U$. An explicit formula for K_3 is

$$K_3(t) = \frac{1}{4\pi} \int_0^\infty K_2(s) \int_0^{2\pi} K_1(t+s-2\sqrt{st}\cos\theta)\, d\theta\, ds.$$

The case $K_1(t) = e^{-t}$ gives: if $K_2 \in U$, then $K_3 \in U$, where

$$K_3(t) = \frac{1}{2} e^{-t} \int_0^\infty e^{-s} K_2(s) J_0(2i\sqrt{st})\, ds.$$

REFERENCES

[1] R. ASKEY, *Certain rational functions whose power series have positive coefficients*, II, this Journal, 5 (1974), pp. 53–57.

[2] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums*, II, American J. Math., 98 (1976), pp. 709–737.

[3] ———, *Certain rational functions whose power series have positive coefficients*, Amer. Math. Monthly, 79 (1972), pp. 327–341.

[4] N. G. DE BRUIJN, *Uncertainty principles in Fourier analysis*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967.

[5] ———, *A theory of generalized functions, with applications to Wigner distribution and Weyl correspondence*, Nieuw Archief voor Wiskunde, 21 (3) (1973), pp. 205–280.

[6] W. N. BAILEY, *On the product of two Laguerre polynomials*, Quart. J. Math., 10 (1939), pp. 60–66.

[7] G. GASPER, *Positive sums of the classical orthogonal polynomials*, this Journal, 8 (1977), pp. 423–447.

[8] ———, *Positivity and special functions*, in Theory and Applications of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975.

[9] H. J. GROENEWOLD, *On the principle of elementary quantum mechanics*, Physica 21 (7) (1946), pp. 405–460.

[10] A. J. E. M. JANSSEN, *Application of the Wigner Distribution to Harmonic Analysis of Generalized Stochastic Processes*, MC-tract 114, Mathematisch Centrum, Amsterdam, 1979.

[11] J. R. KLAUDER, *The design of radar signals having both high range solution and high velocity resolution*, Bell System Tech. J. 39 (1960), pp. 809–820.

[12] J. C. T. POOL, *Mathematical aspects of the Weyl correspondence*, J. Math. Phys. 7 (1966), pp. 66–76.

[13] R. PRICE AND E. M. HOFSTETTER, *Bounds on the volume and height distributions of the ambiguity function*, IEEE Trans. Inform. Theory, 11 (1965), pp. 207–214.

[14] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed. Colloquium Publication, Vol. 23, American Mathematical Society, Providence, Rhode Island, 1975.

[15] G. SANSONE, *Orthogonal Functions*. Rev. Engl. ed. Interscience, New York, 1959.

[16] F. G. TRICOMI, *Vorlesungen über Orthogonalreihen*, 2nd rev. ed., Die Grundlehren der mathematischen Wissenschaften, Vol. 76, Springer-Verlag, New York, 1970.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0008 $01.00/0

THE NONCHARACTERISTIC CAUCHY PROBLEM FOR A CLASS OF EQUATIONS WITH TIME DEPENDENCE. I. PROBLEMS IN ONE SPACE DIMENSION*

JOHN B. BELL†

Abstract. The noncharacteristic Cauchy problem is considered for a general class of operators in one space dimension which are second order in space and first order in time. A weighted energy technique is used to prove uniqueness and logarithmic continuous dependence on the data. The technique is also applied to two problems of higher order. The results are then extended to systems of equations which are coupled in lower order derivative terms.

1. Introduction. Several authors [1]–[4], [6], [8] have considered the noncharacteristic Cauchy problem for the heat equation. These works share the common feature that the results depend on the analyticity properties of the solution as well as special representations of the solution to prove the desired estimates. Ewing and Falk [5] treated a similar, somewhat more general operator using data in addition to the noncharacteristic Cauchy data.

We wish to study a more general class of equations in which we cannot apply analyticity considerations. In particular, we wish to consider the second order operator

$$Lu = a(x,t)u_{xx} + b(t)u_t = F(x,t,u,u_x) = \mathcal{F}(u), \tag{1.1}$$

where $a(x,t) \geqq c > 0$ and

$$|F(x,t,u_1,v_1) - F(x,t,u_2,v_2)| \leqq c(|u_1-u_2| + |v_1-v_2|).$$

(Throughout, c is an explicitly determinable, generic positive constant depending only on coefficients and geometry. Another notational convention will be the omission of the arguments of functions. Furthermore, x and t subscripts denote differentiation.) In addition, we assume that $\partial a/\partial x$ and db/dt are bounded functions.

Let Ω be a space–time region with boundaries

$$(s_1(t), t) \quad \text{for } 0 \leqq t,$$
$$(s_2(t), t) \quad \text{for } 0 \leqq t,$$
$$(x, 0) \quad \text{for } s_1(0) \leqq x \leqq s_2(0),$$

where $s_1(t)$ and $s_2(t)$ are piecewise C^1 curves with $s_1(t) < s_2(t)$ for all t. We let $\Sigma = \{(s_1(t), t) t_0 \leqq t \leqq t_1\}$. Σ is thus a noncharacteristic segment of the boundary of Ω.

We can now pose the problem P1:

P1
$$Lu = \mathcal{F}(u) \quad \text{in } \Omega,$$
$$u = g \quad \text{on } \Sigma,$$
$$\frac{\partial u}{\partial n} = h \quad \text{on } \Sigma.$$

In § 2 we will prove uniqueness and continuous dependence on the data within a restricted class of functions for problem P1.

* Received by the editors February 1, 1980, and in revised form December 16, 1980. This work was supported by a National Science Foundation Graduate Research Fellowship and the NSWC IR Fund.

† Naval Surface Weapons Center-R44, White Oak, Maryland 20910.

We will consider analogous problems for the fourth order operators

$$u_{tt} + ku_{xxxx} = \mathcal{F}, \qquad k > 0 \tag{1.2}$$

and

$$-u_{tt} + ku_{xxxx} = \mathcal{F}, \qquad k > 0 \tag{1.3}$$

in § 3. In § 4 we will extend the results for operators (1.1), (1.2), and (1.3) to weakly coupled systems of equations.

Our approach to the problem will be based on the use of level sets of a function f. We thus introduce

$$D_\alpha = \{(x, t): f(x, t) \leqq \alpha\} \cap \Omega,$$
$$S_\alpha = \partial D_\alpha \cap \Omega,$$
$$\Sigma_\alpha = \partial D_\alpha \cap \partial\Omega.$$

Payne [7] used surfaces of this type in conjunction with logarithmic convexity techniques to treat second order elliptic equations; we will use them with a weighted energy argument. Precisely, we will show that for functions f satisfying explicit criteria set forth in the theorems we can bound certain integrals over D_α from which uniqueness and logarithmic continuous dependence on the data can be deduced.

The fifth section exhibits an example of a suitable family of surfaces and some concluding remarks.

2. Second order problem. In this section we prove the basic inequality from which we deduce uniqueness and continuous dependence estimates for problem P1. We approximate u by a function ϕ in Ω, where ϕ is assumed to have bounded second derivatives in Ω and bounded first derivatives in $\Omega \cup \Sigma$. Letting

$$w = u - \phi,$$

we see that

$$Lw = (\mathcal{F}(u) - \mathcal{F}(\phi)) + (\mathcal{F}(\phi) - L\phi). \tag{2.1}$$

We now substitute $w = e^{\lambda f} v$ in (2.1), yielding

$$Le^{\lambda f} v = [(a(\lambda^2 f_x^2 + \lambda f_{xx}) + bf_t)v + 2\lambda a f_x v_x + a v_{xx} + b v_t] e^{\lambda f}.$$

We next form groups of odd and even terms, where odd and even refer to the number of derivatives of v appearing in the expression. This leads to defining

$$L_o v = 2\lambda a f_x v_x + b v_t,$$
$$L_e v = \lambda^2 (a f_x^2 + a\lambda^{-1} f_{xx} + \lambda^{-1} b f_t) v + a v_{xx}.$$

It then follows that

$$\begin{aligned} a^{-1}(2 \times L_e v \times L_o v) &\leqq a^{-1}(L_e v + L_o v)^2 \\ &= a^{-1} e^{-2\lambda f}[\mathcal{F}(u) - \mathcal{F}(\phi) + (\mathcal{F}(\phi) - L\phi)]^2 \\ &\leqq c\, e^{-2\lambda f}[|\mathcal{F}(u) - \mathcal{F}(\phi)|^2 + |Lu - \mathcal{F}(\phi)|^2]. \end{aligned} \tag{2.2}$$

Expanding (2.2), integrating over D_α, and using the Lipschitz behavior of F yields

$$\begin{aligned}&\iint_{D_\alpha}\{[2\lambda^3 af_x^3+O(\lambda^2)]vv_x+2\lambda af_x v_x v_{xx}+O(\lambda^2)vv_t+bv_{xx}v_t\}\,dx\,dt\\ &\leqq c\iint_{D_\alpha}e^{-2\lambda f}(w^2+w_x^2)\,dx\,dt+c\iint_{D_\alpha}[L\phi-\mathscr{F}(\phi)]^2\,dx\,dt\\ &\leqq c\iint_{D_\alpha}(\lambda^2v^2+v_x^2)\,dx\,dt+c\iint_{D_\alpha}[L\phi-\mathscr{F}(\phi)]^2\,dx\,dt.\end{aligned}\tag{2.3}$$

Integrating by parts we find that

$$\begin{aligned}&\oint_{\partial D_\alpha}[\lambda^3 af_x^3n_x+O(\lambda^2)]v^2\,dS+\oint_{\partial D_\alpha}[\lambda af_xn_x+O(1)]v_x^2\,dS+\oint_{\partial D_\alpha}bv_xv_tn_x\,dS\\ &-\iint_{D_\alpha}\{[\lambda^3(af_x^3)_x+O(\lambda^2)]v^2+[\lambda(af_x)_x+O(1)]v_x^2\}\,dx\,dt\\ &\leqq\iint_{D_\alpha}[L\phi-\mathscr{F}(\phi)]^2\,dx\,dt.\end{aligned}\tag{2.4}$$

If we now assume that

$$f_x\geqq c>0\quad\text{in } D_\alpha,$$

$$(a(x,t)f_x^3)_x\leqq -c<0\quad\text{in } D_\alpha,$$

and

$$(a(x,t)f_x)_x\leqq -c<0\quad\text{in } D_\alpha,$$

then for sufficiently large λ

$$-\iint_{D_\alpha}\{[\lambda^3(af_x^3)_x+O(\lambda^2)]v^2+[\lambda(af_x)_x+O(1)]v_x^2\}\,dx\,dt\geqq 0.$$

Thus for sufficiently large λ, (2.4) becomes

$$\begin{aligned}&\lambda^3\int_{S_\alpha}av^2|\nabla f|^{-1}\,dS+\lambda\int_{S_\alpha}v_x^2|\nabla f|^{-1}\,dS\\ &\qquad\leqq c\int_{S_\alpha}b(t)v_xv_t|\nabla f|^{-1}\,dS+c\int_{\Sigma_\alpha}(\lambda^3v^2+\lambda v_x^2+v_t^2)\,dS\\ &\qquad\quad+c\iint_{D_\alpha}[L\phi-F(x,t,\phi,\phi_x)]^2\,dx\,dt.\end{aligned}$$

Substituting $e^{-\lambda f}w$ for v and applying the arithmetic geometric mean inequality then yields

$$\begin{aligned}\lambda^2\int_{S_\alpha}w^2|\nabla f|^{-1}\,dS&\leqq c\int_{S_\alpha}b^2(t)w_t^2|\nabla f|^{-1}\,dS\\ &\quad+c\,e^{2\lambda\alpha}\int_{\Sigma_\alpha}(\lambda^3w^2+\lambda w_x^2+w_t^2)\,dS\\ &\quad+c\,e^{2\lambda\alpha}\iint_{D_\alpha}[L\phi-F(x,t,\phi,\phi_x)]^2\,dx\,dt.\end{aligned}\tag{2.5}$$

Integrating (2.5) with respect to α then yields the following theorem:

THEOREM 1. *Suppose u is a solution to problem* P1 *and ϕ is a function as described previously. Suppose further that there exists a function f such that*

1) $D_\alpha \subset D_\beta$ *and* $\Sigma_\alpha \subset \Sigma$ *for* $0<\alpha<\beta\leqq 1$,
2) $f_x \geqq c>0$ *in* D_1
3) $(a(x,t)f_x)_x \leqq -c<0$ *in* D_1, $\quad (a(x,t)f_x^3)_x \leqq -c<0$ *in* D_1.

Then, for all α, $0<\alpha\leqq 1$, we have the estimate

$$\lambda^3 \iint_{D_\alpha} (u-\phi)^2\,dx\,dt \leqq c\iint_{D_\alpha} b^2(t)(u_t-\phi_t)^2\,dx\,dt + c\,e^{2\lambda\alpha}\int_{\Sigma_\alpha}[\lambda^2(u-\phi)^2+(u_x-\phi_x)^2+\lambda^{-1}(u_t-\phi_t)^2]\,dS + c\lambda^{-1}e^{2\lambda\alpha}\iint_{D_\alpha}[L\phi-\mathcal{F}(\phi)]^2\,dx\,dt$$

for λ sufficiently large.

The inequality presented in the above theorem can now be used to prove uniqueness and continuous dependence on the data for problem P1. First, however, we must digress a moment to prove a technical lemma required to appropriately fix the stabilization class.

LEMMA. *Let $D \subset \Omega$ be a domain containing $\bar{D}_\alpha \cap \Omega$ and such that $\partial D \cap \partial\Omega \subset \Sigma$. Let ω be a C^∞ cutoff function such that $\omega \equiv 1$ on $\bar{D}_\alpha$ and $\omega \equiv 0$ on Ω/D. Then*

$$\iint_{D_\alpha} b^2(t)(u_t-\phi_t)^2\,dx\,dt \leqq c\int_\Sigma [(u-\phi)^2+(u_x-\phi_x)^2+(u_t-\phi_t)^2]\,dS + c\iint_D (u-\phi)^2\,dx\,dt + c\iint_D |L\phi-\mathcal{F}(\phi)|^2\,dx\,dt.$$

Proof. First we see that

$$\iint_{D_\alpha} b^2(u_t-\phi_t)^2\,dx\,dt \leqq \iint_D \omega^4 b^2(u_t-\phi_t)^2\,dx\,dt = \iint_D \omega^4 b(u_t-\phi_t)[a(\phi_{xx}-u_{xx})+\mathcal{F}(u)-\mathcal{F}(\phi)+\mathcal{F}(\phi)-L\phi]\,dx\,dt \leqq \frac{1}{2}\iint_D \omega^4 b^2(u_t-\phi_t)^2\,dx\,dt + c\iint_D (\omega^4(u-\phi)^2+\omega^2(u_x-\phi_x)^2)\,dx\,dt + c\int_\Sigma (u_t-\phi_t)^2+(u_x-\phi_x)^2\,dS + c\iint_D |\mathcal{F}(\phi)-L\phi|^2\,dx\,dt.$$

Next, observe that

$$\begin{aligned}
&\iint_D \omega^2(u_x-\phi_x)^2\,dx\,dt \\
&\quad=\oint_{\partial D}\omega^2(u-\phi)(u-\phi)_x n_x\,dS \\
&\qquad-\iint_D[(\omega^2)_x(u-\phi)_x(u-\phi)+\omega^2(u-\phi)_{xx}(u-\phi)]\,dx\,dt \\
&\quad=\oint_{\partial D}\omega^2(u-\phi)(u-\phi)_x n_x\,dS-\iint_D\omega^2 a^{-1}(u-\phi)[\mathcal{F}(u)-\mathcal{F}(\phi)]\,dx\,dt \\
&\qquad+\iint_D\omega^2\frac{b}{a}(u-\phi)(u-\phi)_t\,dx\,dt-\iint_D\omega_x^2(u-\phi)_x(u-\phi)\,dx\,dt \\
&\qquad+\iint_D\omega^2 a^{-1}(u-\phi)(L\phi-\mathcal{F}(\phi))\,dx\,dt \\
&\quad\leqq c\int_\Sigma(u-\phi)^2+(u_x-\phi_x)^2\,dS+\frac{1}{2}\iint_D\omega^2(u_x-\phi_x)^2\,dx\,dt \\
&\qquad+c\iint_D(u-\phi)^2\,dx\,dt+c\iint_D|L\phi-\mathcal{F}(\phi)|^2\,dx\,dt.
\end{aligned}$$

Combining these two results then yields the desired inequality. □

The inequality derived in the theorem can now be used to prove uniqueness and continuous dependence on the data within a restricted stabilization class.

Let

$$\mathcal{M}=\{u:\|u\|_{L^2(\Omega)}\leqq M\}.$$

The result follows from applying the inequality of the theorem, substituting for ϕ a function $\tilde{u}$ which solves a problem with data close to the data for problem P1, i.e., where

$$\|\tilde{u}-g\|^2_{H^1(\Sigma)}+\left\|\frac{\partial\tilde{u}}{\partial n}-h\right\|^2_{L^2(\Sigma)}+\|L\tilde{u}-\mathcal{F}(\tilde{u})\|^2_{L^2(\Omega)}=\varepsilon^2$$

is small.

We can then prove the desired uniqueness and continuous dependence results.

COROLLARY. *If a solution to* P1 *exists, then it is unique. Furthermore, for* $u,\tilde{u}\in\mathcal{M}$ *satisfying the properties described above we have the following estimate for* $0<\alpha\leqq 1$:

$$\iint_{D_\alpha}(u-\tilde{u})^2\,dx\,dt<c\left\{\varepsilon^{2(1-\alpha)}M^{2\alpha}+\frac{4M^2}{\log(M^2/\varepsilon^2)}\right\}\frac{1}{\log(M^2/\varepsilon^2)}.$$

Proof. In the situation described above, the inequality in the theorem becomes

$$\iint_{D_\alpha}(u-\tilde{u})^2\,dx\,dt<\frac{c}{\lambda^3}M^2+\frac{c\,e^{2\lambda\alpha}}{\lambda}\varepsilon^2.$$

For uniqueness we have $\varepsilon=0$. Letting $\lambda\to\infty$ then gives the desired result. For continuous dependence, letting $\lambda=\frac{1}{2}\log(M^2/\varepsilon^2)$ yields the desired result. □

3. Fourth order problems. In this section we treat problems analogous to problem P1 for operators (1.2) and (1.3). We will begin with operator (1.2) which, in the classical theory, describes the deformation of an isotropic beam loaded normal to the beam. Some generality may be added to the problem by allowing k to be a function of x and t; however, we will restrict our attention to the constant coefficient case for the sake of simplicity. We are thus led to the problem P2:

$$
\text{P2}\qquad
\begin{aligned}
&Lu \equiv u_{tt} + k u_{xxxx} = F(x, t, u, u_x, u_t, u_{xx}, u_{xt}, u_{xxx}) \equiv \mathcal{F}(u) \text{ in } \Omega,\\
&u = \zeta(t) \qquad \text{on } \Sigma\\
&\frac{\partial u}{\partial n} = \eta(t) \qquad \text{on } \Sigma,\\
&\frac{\partial^2 u}{\partial n^2} = \theta(t) \qquad \text{on } \Sigma,\\
&\frac{\partial^3 u}{\partial n^3} = \kappa(t) \qquad \text{on } \Sigma.
\end{aligned}
$$

where F is Lipschitz in its last 6 arguments. We will now prove an estimate analogous to Theorem 1 from which we can deduce uniqueness and continuous dependence on the data for problem P2.

THEOREM 2. *Suppose u satisfies problem* P2 *and ϕ is any function possessing bounded fourth order derivatives in Ω and third order derivatives in $\Omega \cup \Sigma$. Furthermore, assume there exists a function f which satisfies*

$$
\begin{aligned}
&D_\alpha \subset D_\beta, \qquad \Sigma_\alpha \subset \Sigma, \qquad 0 < \alpha < \beta \leqq 1,\\
&f_x \geqq c > 0 \quad \text{in } D_1, \qquad f_{xx} \leqq -c < 0 \quad \text{in } D_1,
\end{aligned}
$$

and

$$
f(x, t) = f_1(x) + f_2(t).
$$

Then, for sufficiently large λ, we have

$$
\begin{aligned}
\iint_{D_\alpha} (u-\phi)^2\, dx\, dt \leqq c \iint_{D_\alpha} &\lambda^{-4}(u_t-\phi_t)^2 + \lambda^{-7}(u_{tt}-\phi_{tt})^2\, dx\, dt\\
&+ c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} [\lambda^{-1}(u-\phi)^2 + \lambda^{-3}(u_x-\phi_x)^2 + \lambda^{-5}(u_t-\phi_t)^2\\
&\qquad + \lambda^{-5}(u_{xx}-\phi_{xx})^2 + \lambda^{-7}(u_{xxx}-\phi_{xxx})^2\\
&\qquad + \lambda^{-7}(u_{xt}-\phi_{xt})^2 + \lambda^{-8}(u_{tt}-\phi_{tt})^2]\, dS\\
&+ c\lambda^{-8} \iint_\Omega (L\phi - \mathcal{F}(\phi))^2\, dx\, dt.
\end{aligned}
\tag{3.1}
$$

Proof. Without loss of generality, we may take $k = 1$. As before we let $(u - \phi) = e^{\lambda f} v$. Thus v satisfies

$$\begin{aligned} L_v &= (\lambda^4 f_x^4 + O(\lambda^3))v + (4\lambda^3 f_x^3 + O(\lambda^2))v_x + (6\lambda^2 f_x^2 + O(\lambda))v_{xx} \\ &\quad + 4\lambda^3 v_{xxx} + v_{xxxx} + 2\lambda f_t v_t + v_{tt} \\ &= [\mathcal{F}(u) - \mathcal{F}(\phi) + (\mathcal{F}(\phi) - L\phi)]\, e^{-\lambda f}. \end{aligned}$$

Since we intend λ to be large, we may omit lower order terms in λ from the remainder of the proof. This leads to definitions of odd and even parts of the operator as was done previously; viz.,

$$L_e v \equiv \lambda^4 f_x^4 v + 6\lambda^2 f_x^2 v_{xx} + v_{xxxx} + v_{tt},$$

and

$$L_o v \equiv 4\lambda^3 f_x^3 v_x + 4\lambda f_x v_{xxx} + 2\lambda f_t v_t.$$

The key to success in the use of the weighted energy technique employed is the reduction of the initial inequality to a form in which the volume terms can be discarded. To accomplish that we must, in this case, begin with a more complex initial inequality. Hence, we form

$$\begin{aligned} &\iint_{D_\alpha} \{(L_e v) x (L_o v) + 10\lambda^3 [(L_e + L_o)v - e^{-\lambda f}(\mathcal{F}(u) - \mathcal{F}(\phi)] f_x^2 f_{xx}\}\, dx\, dt \\ &\leqq \frac{1}{2} \iint_{D_\alpha} \{e^{-2\lambda f}[(\mathcal{F}(u) - \mathcal{F}(\phi)) + (\mathcal{F}(\phi) - L\phi)]^2\}\, dx\, dt \\ &\quad + \iint_{D_\alpha} 10\lambda^3\, e^{-\lambda f}(\mathcal{F}(\phi) - L\phi) f_x^2 f_{xx} v\, dx\, dt. \end{aligned} \tag{3.2}$$

Expanding the first term we find

$$\begin{aligned} &\iint_{D_\alpha} (\lambda^4 f_x^4 v + 6\lambda^2 f_x^2 v_{xx} + v_{xxxx} + v_{tt}) \cdot (4\lambda^3 f_x^3 v_x + 4\lambda f_x v_{xxx} + 2\lambda f_t v_t)\, dx\, dt \\ &= \oint_{\partial D_\alpha} (2\lambda^7 f_x^7 n_x v^2 - O(\lambda^5) v v_x + 10\lambda^5 f_x^5 n_x v_x^2 \\ &\qquad + 4\lambda^5 f_x^5 n_x v x_{xx} - O(\lambda^3) v_{xx} v_x + 4\lambda^3 f_x^3 n_x v_{xxx} v_x \\ &\qquad + 10\lambda^3 f_x^3 n_x v_{xx}^2 + O(\lambda^3) v_x v_t - 2\lambda^3 f_x^3 n_x v_t^2 + 2\lambda f_x n_x v_{xxx}^2 \\ &\qquad + 2\lambda f_x n_x v_{tx}^2 + O(\lambda) v_{xx} v_{xt} + O(\lambda) v_{xx} v_t + O(\lambda) v_{xx} v_{tt} \\ &\qquad\qquad + O(\lambda) v_{xt} v_t + O(\lambda) v_{xxx} v_t)\, dS \\ &\quad + \iint_{D_\alpha} (-14\lambda^7 f_x^6 f_{xx} v^2 - 30\lambda^5 f_x^4 f_{xx} v_x^2 \\ &\qquad - 18\lambda^3 f_x^2 f_{xx} v_{xx}^2 - 2\lambda f_{xx} v_{xxx}^2 - 6\lambda f_{xx} v_{xt}^2 \\ &\qquad\qquad + 6\lambda^3 f_x^2 f_{xx} v_t^2 + O(\lambda^3) v_x v_t)\, dx\, dt. \end{aligned} \tag{3.3}$$

The expansion of the second term yields

$$
\begin{aligned}
&\iint_{D_\alpha} \lambda^3(L_e v + L_o v - e^{-\lambda f}(\mathcal{F}(u)-\mathcal{F}(\phi)))10 f_x^2 f_{xx} v \, dx\, dt \\
&\quad = \iint_{D_\alpha} \lambda^7 10 f_x^6 f_{xx} v^2 \, dx\, dt + O(\lambda^6)\oint_{\partial D_\alpha} v^2 \, dS \\
&\qquad - \iint_{D_\alpha} 60\lambda^5 (f_x^4 f_{xx}) v_x^2 \, dx\, dt + O(\lambda^5)\oint_{\partial D_\alpha} v v_x \, dS \\
&\qquad + O(\lambda^4)\oint_{\partial D_\alpha} v_{xx} v \, dS + O(\lambda^4)\oint_{\partial D_\alpha} v_x^2 \, dS + O(\lambda^3)\oint_{\partial D_\alpha} v_{xxx} v \, dS \\
&\qquad + O(\lambda^3)\oint_{\partial D_\alpha} v_{xx} v_x \, dS + 10\lambda^3 \iint_{D_\alpha} f_x^2 f_{xx} v_{xx}^2 \, dx\, dt \\
&\qquad - 10\lambda^3 \iint_{D_\alpha} f_x^2 f_{xx} v_t^2 \, dx\, dt + O(\lambda^3)\oint_{\partial D_\alpha} v_t v \, dS \\
&\qquad - \iint_{D_\alpha} 10\lambda^3 e^{-\lambda f}(\mathcal{F}(u)-\mathcal{F}(\phi)) f_x^2 f_{xx} v \, dx\, dt.
\end{aligned} \tag{3.4}
$$

Note that

$$
\begin{aligned}
&\lambda^3 \iint_{D_\alpha} 10\, e^{-\lambda f}(\mathcal{F}(u)-\mathcal{F}(\phi)) f_x^2 f_{xx} v \, dx\, dt \\
&\quad \leqq c\lambda^3 \iint_{D_\alpha} e^{-\lambda f}(|u-\phi| + |(u-\phi)_x| + |(u-\phi)_{xx}| \\
&\qquad\qquad + |(u-\phi)_{xxx}| + |(u-\phi)_{xt}| + |(u-\phi)_t|) v \, dx\, dt \\
&\quad \leqq c \iint_{D_\alpha} (\lambda^6 v^2 + \lambda^4 v_x^2 + \lambda^2 v_{xx}^2 + v_{xxx}^2 + v_{xt}^2 + \lambda^2 v_t^2) \, dx\, dt.
\end{aligned} \tag{3.5}
$$

Combining (3.3) and (3.4) and using the estimate (3.5), we find that (3.2) becomes

$$
\begin{aligned}
&\oint_{\partial D_\alpha} \{2\lambda^7 f_x^7 n_x v^2 + O(\lambda^5) v_x v + 10\lambda^5 f_x^5 n_x v_x^2 + 4\lambda^5 f_x^5 n_x v_{xx} v \\
&\quad + O(\lambda^3) v_{xx} v_x + O(\lambda^3) v_{xxx} v + 4\lambda^3 f_x^3 n_x v_{xxx} v_x + O(\lambda^3) v_x v_t \\
&\quad + 10\lambda^3 f_x^3 n_x v_{xx}^2 + 2\lambda f_x n_x v_{xt}^2 + 2\lambda f_x n_x v_{xxx}^2 - 2\lambda^3 f_x^3 n_x v_t^2 \\
&\quad + O(\lambda) v_{xx} v_{xt} + O(\lambda) v_{xx} v_t + O(\lambda) v_{xx} v_{tt} + O(\lambda) v_{xt} v_t + O(\lambda) v_{xxx} v_t\} \, dS \\
&\leqq \iint_{D_\alpha} (4\lambda^7 f_x^6 f_{xx} v^2 + 90\lambda^5 f_x^4 f_{xx} v_x^2 + 8\lambda^3 f_x^2 f_{xx} v_{xx}^2 \\
&\qquad + 2\lambda f_{xx} v_{xxx}^2 + 6\lambda f_{xx} v_{xt}^2 + 4\lambda^3 f_x^2 f_{xx} v_t^2 + O(\lambda^3) v_x v_t) \, dx\, dt \\
&\quad + \iint_{D_\alpha} (L\phi - \mathcal{F}(\phi))^2 \, dx\, dt,
\end{aligned} \tag{3.6}
$$

for λ sufficiently large.

Applying Schwarz's inequality and the arithmetic-geometric mean inequality to (3.6) yields

$$\begin{aligned}\int_{S_\alpha} &(2\lambda^7 f_x^8 v^2 + 10\lambda^5 f_x^6 v_x^2 + 4\lambda^5 f_x^6 v_{xx} v + 4\lambda^3 f_x^4 v_{xxx} v_x \\ &\qquad + 10\lambda^3 f_x^4 v_{xx}^2 + 2\lambda_x^2 v_{xxx}^2 + 2\lambda f_x^2 v_{tx})|\nabla f|^{-1}\, dS \\ \leqq c &\int_{S_\alpha} (\lambda^3 v_t^2 + v_{tt}^2)|\nabla f|^{-1}\, dS \\ &\qquad + c\int_{\Sigma_\alpha} (\lambda^7 v + \lambda^5 v_x^2 + \lambda^3 v_t^2 + \lambda^3 v_{xx}^2 + \lambda v_{xxx}^2 + \lambda v_{xt}^2 + v_{tt}^2)\, dS \\ &+ \iint_{D_\alpha} (L\phi - \mathscr{F}(\phi))^2\, dx\, dt.\end{aligned} \tag{3.7}$$

Next observe that

$$4\lambda^5 f_x^6 v_{xx} v + (8\lambda^3 f_x^4 v_{xx}^2 + \tfrac{1}{2}\lambda^7 f_x^8 v^2) \geqq 0$$

and

$$4\lambda^3 f_x^4 v_{xxx} v_x + (8\lambda^5 f_x^6 v_x^2 + \tfrac{1}{2}\lambda f_x^2 v_{xxx}^2) \geqq 0.$$

Discarding unneeded positive terms on the left of (3.7) and using the two preceding inequalities we see that for λ sufficiently large (3.7) becomes

$$\begin{aligned}\lambda^7 &\int_{S_\alpha} v^2 |\nabla f|^{-1}\, dS \\ &\leqq c\int_{S_\alpha} (\lambda^3 v_t^2 + v_{tt}^2)|\nabla f|^{-1}\, dS \\ &\qquad + \int_{\Sigma_\alpha} (\lambda^7 v^2 + \lambda^5 v_x^2 + \lambda^3 v_t^2 + \lambda^3 v_{xx}^2 + \lambda v_{xxx}^2 + \lambda v_{xt}^2 + v_{tt}^2)\, dS \\ &\qquad + \iint_{D_\alpha} (L\phi - \mathscr{F}(\phi))^2\, dx\, dt.\end{aligned}$$

Substituting for v we obtain

$$\begin{aligned}\int_{S_\alpha} &(u-\phi)^2 |\nabla f|^{-1}\, dS \\ &\leqq c\int_{S_\alpha} (\lambda^{-4}(u_t - \phi_t)^2 + \lambda^{-7}(u_{tt} - \phi_{tt})^2\, dS \\ &\quad + c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} \Big((u-\phi)^2 + \lambda^{-2}(u_x - \phi_x)^2 + \lambda^{-4}(u_t - \phi_t)^2 \\ &\qquad + \lambda^{-4}(u_{xx} - \phi_{xx})^2 + \lambda^{-6}(u_{xxx} - \phi_{xxx})^2 \\ &\qquad + \lambda^{-6}(u_{xt} - \phi_{xt})^2 + \lambda^{-7}(u_{tt} - \phi_{tt})^2 \Big)\, dS \\ &\quad + \frac{c\, e^{2\lambda\alpha}}{\lambda} \iint_{D_\alpha} (L\phi - \mathscr{F}(\phi))^2\, dx\, dt\end{aligned}$$

for λ large. Integrating with respect to α then yields the desired result. □

For this problem we require a more restrictive stabilization class. Let

$$\mathcal{M} = \{u: \|u_t\|^2_{L^2(\Omega)} + \|u_{tt}\|^2_{L^2(\Omega)} \leqq M^2\}.$$

We can now establish uniqueness and continuous dependence on the data for problem P2 by letting $\phi = \tilde{u}$ corresponding to the solution of a related problem, i.e., where

$$\|\zeta - \tilde{u}\|^2_{H^3(\Sigma)} + \left\|\eta - \frac{\partial \tilde{u}}{\partial n}\right\|^2_{H^2(\Sigma)} + \left\|\theta - \frac{\partial^2 \tilde{u}}{\partial n^2}\right\|^2_{H^1(\Sigma)} + \left\|\kappa - \frac{\partial^3 \tilde{u}}{\partial n^3}\right\|^2_{L^2(\Sigma)} + \|L\tilde{u} - \mathcal{F}(\tilde{u})\|^2_{L^2(\Omega)} = \varepsilon^2$$

is small.

COROLLARY. *If it exists, the solution to problem* P2 *is unique. Furthermore, for* $u \in \mathcal{M}$, *a solution of* P2 *and* $\tilde{u} \in \mathcal{M}$ *as described above we have the continuous dependence estimate*

$$\int\int_{D_\alpha} (u - \tilde{u})^2 \, dx \, dt \leqq c\left\{M^{2\alpha}\varepsilon^{2(1-\alpha)} + \frac{8M^2}{[\log (M^2/\varepsilon^2)]^3}\right\} \frac{1}{\log (M^2/\varepsilon^2)}$$

for $0 < \alpha \leqq 1$.

Proof. In the above described situation, inequality (3.2) reduces to

$$\int\int_{D_\alpha} (u - \tilde{u})^2 \, dx \, dt \leqq \lambda^{-4} M^2 + \lambda^{-1}\varepsilon^3 e^{2\lambda\alpha}. \tag{3.8}$$

For uniqueness we have $\varepsilon = 0$ and let $\lambda \to \infty$ in (3.8). Continuous dependence on the data follows immediately upon setting $\lambda = \frac{1}{2}\log (M^2/\varepsilon^2)$ □

Notice that we needed to impose a much more restrictive stabilization condition to prove continuous dependence on the data for this problem.

The cause of the problem is, in fact, the sign of k. We will now turn our attention to operator (1.3). This operator corresponds to the composition of a forward heat operator and its formal adjoint; namely, a backward heat operator.

We now pose the problem P3:

P3
$$\begin{aligned}
&Lu \equiv -u_{tt} + ku_{xxxx} = \mathcal{F}(u) && \text{in } \Omega,\\
&u = \zeta(t) && \text{on } \Sigma,\\
&\frac{\partial u}{\partial n} = \eta(t) && \text{on } \Sigma,\\
&\frac{\partial^2 u}{\partial n^2} = \theta(t) && \text{on } \Sigma,\\
&\frac{\partial^3 u}{\partial n^3} = \kappa(t) && \text{on } \Sigma.
\end{aligned}$$

Proceeding as before we prove the following theorem.

THEOREM 3. *Suppose* u *is a solution to* P3 *and* ϕ *is a function having bounded fourth derivatives in* Ω *and bounded third derivatives on* $\Omega \cup \Sigma$. *Furthermore suppose there exists a function* f *such that*

$$\begin{aligned}
&D_\alpha \subset D_\beta, \qquad \Sigma_\alpha \subset \Sigma, && 0 < \alpha < \beta \leqq 1,\\
&f_x \geqq c > 0 && \text{in } \Omega,\\
&f_{xx} \leqq -c < 0 && \text{in } \Omega,\\
&f(x, t) = f_1(x) + f_2(t) && \text{in } \Omega.
\end{aligned}$$

Then, the following inequality holds for λ sufficiently large:

$$
\begin{aligned}
&\lambda^2 \iint_{D_\alpha} (u-\phi)^2\,dx\,dt \\
&\qquad \leqq c \iint_{D_\alpha} (u_{xxt}-\phi_{xxt})^2+\lambda^2(u_{xt}-\phi_{xt})^2\,dx\,dt \\
&\qquad\quad + c\lambda^6 e^{2\lambda\alpha} \int_{\Sigma_\alpha} [(u-\phi)+(u_x-\phi_x)^2+(u_{xx}-\phi_{xx})^2 \\
&\qquad\qquad + (u_t-\phi_t)^2+(u_{xt}-\phi_{xt})^2+(u_{xxt}-\phi_{xxt})^2]\,dS \\
&\qquad\quad + c\lambda^{-1} e^{2\lambda\alpha} \iint_{D_\alpha} |L\phi-\mathscr{F}(\phi)|^2\,dx\,dt.
\end{aligned} \tag{3.9}
$$

Proof. We again assume $k=1$ and ignore lower terms in λ. Substituting $(u-\phi)=e^{\lambda f}v$, we see that v satisfies

$$
\begin{aligned}
Lv &\equiv \lambda^4 f_x^4 v+4\lambda^3 f_x^3 v_x+6\lambda^2 f_x^2 v_{xx}+4\lambda f_x v_{xxx}+v_{xxxx}-2\lambda f_t v_t-v_{tt}+\text{l.o.t.} \\
&= e^{-\lambda f}(\mathscr{F}(u)-\mathscr{F}(\phi))+e^{-\lambda f}(\mathscr{F}(\phi)-L\phi).
\end{aligned}
$$

Breaking this expression into odd and even parts we form the inequality

$$
\begin{aligned}
&\iint_{D_\alpha} [(L_e v)\times(L_o v)f_x^{-1}-\lambda^{1/2}(L_e v+L_o v)v_{xx}]\,dx\,dt \\
&\qquad \leqq \iint_{D_\alpha} f_x^{-1} e^{-\lambda f}|(\mathscr{F}(u)-\mathscr{F}(\phi))+(\mathscr{F}(\phi)-L\phi)|^2\,dx\,dt \\
&\qquad\quad + \lambda^{1/2} \iint_{D_\alpha} e^{-\lambda f}(|\mathscr{F}(u)-\mathscr{F}(\phi)|+|\mathscr{F}(\phi)-L(\phi)|)v_{xx}\,dx\,dt,
\end{aligned} \tag{3.10}
$$

where

$$
L_e v=\lambda^4 f_x^4 v+6\lambda^2 f_x^2 v_{xx}+v_{xxxx}-v_{tt}+\text{l.o.t.}
$$

and

$$
L_o v=4\lambda^3 f_x^3 v_x+4\lambda f_x v_{xxx}-2\lambda f_t v_t+\text{l.o.t.}
$$

Expanding, integrating by parts and applying the arithmetic geometric mean inequality as was done previously yields

$$
\begin{aligned}
&\oint_{\partial D_\alpha} (\lambda^7 f_x^7 v^2+2\lambda^5 f_x^5 v_x^2+2\lambda^3 f_x^3 v_{xx}^2+\lambda f_x v_{xxx}^2+2\lambda^3 f_x^3 v_t^2 \\
&\qquad\qquad + O(\lambda)v_{xt}^2+O(\lambda)v_{xxt}v_t+O(\lambda)v_{xx}v_{xt})|\nabla f|^{-1}\,dS \\
&\qquad \leqq \iint_{D_\alpha} (12\lambda^7 f_x^5 f_{xx}v^2+24\lambda^5 f_x^3 f_{xx}v_x^2+12\lambda^3 f_x f_{xx}v_{xx}^2 \\
&\qquad\qquad + 4\lambda^3 f_x f_{xx}v_t^2+\lambda^{1/2}v_{xxx}^2+\lambda^{1/2}v_{xt}^2)\,dx\,dt \\
&\qquad\quad + \iint_{D_\alpha} |L\phi-\mathscr{F}(\phi)|^2 e^{-2\lambda f}\,dx\,dt.
\end{aligned} \tag{3.11}
$$

The first volume term on the right-hand side of (3.11) is negative. Thus, when we substitute $e^{-\lambda f}(u-\phi)$ for v in (3.11), the inequality becomes

$$\begin{aligned}
\lambda^7 \int_{S_\alpha} (u-\phi)^2 |\nabla f|^{-1}\, dS
&\leqq c \int_{S_\alpha} [(u_{xxt}-\phi_{xxt})^2 + \lambda (u_{xt}-\phi_{xt}]|\nabla f|^{-1}\, dS \\
&\quad + c\lambda^7 e^{2\lambda\alpha} \int_{\Sigma_\alpha} [(u-\phi)^2 + (u_x-\phi_x)^2 + (u_{xx}-\phi_{xx})^2 + (u_{xxx}-\phi_{xxx})^2 + (u_t-\phi_t)^2 \\
&\qquad\qquad \times (u_{xt}-\phi_{xt})^2 + (u_{xxt}-\phi_{xxt})^2]\, dS \\
&\quad + c\, e^{2\lambda\alpha} \int\int_{D_\alpha} |L\phi - \mathcal{F}(\phi)|^2\, dx\, dt,
\end{aligned}$$

for λ sufficiently large. Integrating with respect to α then yields the desired result. □

As was the case for problem P1, we can prove a technical lemma which will allow us to use a less restrictive stabilization class than was used for problem P2.

LEMMA. *Let D be a subdomain of Ω such that*

$$\bar{D}_\alpha \cap \Omega \subseteq D \quad \text{and} \quad \partial D \cap \partial\Omega \subseteq \Sigma.$$

Then, for λ sufficiently large, if u satisfies (1.3) *and ϕ is as described in Theorem* 3 *we have*

$$\begin{aligned}
&\int\int_{D_\alpha} (u_{xxt}-\phi_{xxt})^2 + \lambda^2 (u_{xt}-\phi_{xt})^2\, dx\, dt \\
&\quad \leqq c\lambda^4 \int\int_D (u-\phi)^2\, dx\, dt \\
&\qquad + c\lambda^2 \int_\Sigma [(u-\phi)^2 + (u_t-\phi_t)^2 + (u_x-\phi_x)^2 + (u_{xx}-\phi_{xx})^2 \\
&\qquad\qquad + (u_{xt}-\phi_{xt})^2 + (u_{xxx}-\phi_{xxx})^2 + (u_{xxt}-\phi_{xxt})^2 + (u_{tt}-\phi_{tt})^2]\, dS \\
&\qquad + c\lambda^2 \int\int_D |L\phi - \mathcal{F}(\phi)|^2\, dx\, dt.
\end{aligned}$$

Proof. Let ω be a C^∞ cutoff function such that $\omega \equiv 1$ in D_α and $\omega \equiv 0$ in $\Omega \cap D^c$. If we let w denote $u-\phi$, we then have

$$\int\int_{D_\alpha} (w_{xxt}^2 + \lambda^2 w_{xt}^2)\, dx\, dt \leqq \int\int_D (\omega^n w_{xxt}^2 + \omega^{n-2}\lambda^2 w_{xt}^2)\, dx\, dt. \tag{3.12}$$

Now,

$$
\begin{aligned}
\iint_D \omega^n w_{xxt}^2\,dx\,dt &= \int_{\partial D} \omega^n w_{xxt} w_{xt} n_x\,dS - \iint_D \omega_x^n w_{xxt} w_{xt}\,dx\,dt \\
&\quad - \iint_D \omega^n w_{xt} w_{xxxt}\,dx\,dt \\
&\leqq \oint_{\partial D} \omega^n w_{xxt} w_{xt} n_x\,dS - \oint_{\partial D} \omega^n w_{xt} w_{xxx} n_t\,dS \\
&\quad + \frac{1}{2}\iint_D \omega^n w_{xxt}^2\,dx\,dt + c\iint_D \omega^{n-2} w_{xt}^2\,dx\,dt + \iint_D \omega_t^n w_{xt} w_{xxx}\,dx\,dt \\
&\quad + \iint_D \omega^n w_{xtt} w_{xxx}\,dx\,dt \\
&\leqq \oint_{\partial D} (\omega^n w_{xxt} w_{xt} n_x - \omega^n w_{xt} w_{xxx} n_t + \omega^n w_{tt} w_{xxx} n_x)\,dS \\
&\quad + \iint_D \omega^n w_{xxt}^2\,dx\,dt + c\iint_D \omega^{n-2} w_{xt}^2\,dx\,dt \\
&\quad + c\iint_D \omega^n w_{xxx}^2\,dx\,dt - \iint_D \omega_x^n w_{tt} w_{xxx}\,dx\,dt - \iint_D \omega^n w_{tt} w_{xxxx}\,dx\,dt.
\end{aligned} \tag{3.13}
$$

If we now substitute for w_{tt} using the equation, (3.13) becomes

$$
\begin{aligned}
&\frac{1}{2}\iint_D \omega^n w_{xxt}^2\,dx\,dt \\
&\quad \leqq \oint_{\partial D} (\omega^n w_{xxt} w_{xt} n_x - \omega^n w_{xt} w_{xxx} n_t + \omega^n w_{tt} w_{xxx} n_x)\,dS \\
&\quad + c\iint_D \omega^{n-2} w_{xt}^2 + \omega^n w_{xxx}^2\,dx\,dt \\
&\quad - \iint_D \omega_x^n (w_{xxxx} - [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(u) - L\phi)]) w_{xxx}\,dx\,dt \\
&\quad - \iint_D \omega^n (w_{xxxx} - [(\mathscr{F}(u) - \mathscr{F}(\phi)) + [\mathscr{F}(\phi) - L\phi]) w_{xxxx}\,dx\,dt \\
&\quad \leqq c\int_\Sigma (w_{xxt}^2 + w_{xt}^2 + w_{xxx}^2 + w_{tt}^2)\,dS \\
&\quad + c\iint_D \omega^{n-2}(w_{xt}^2 + w_{xxx}^2 + w^2 + w_x^2 + w_{xx}^2 + w_t^2)\,dx\,dt \\
&\quad - \frac{1}{2}\iint_D \omega^n w_{xxxx}^2\,dx\,dt + c\iint_D |\mathscr{F}(\phi) - L\phi|^2\,dx\,dt.
\end{aligned} \tag{3.14}
$$

Let us now examine our other term:

$$\iint_D \omega^{n-2} w_{xt}^2 \, dx\, dt$$

$$= \oint_{\partial D} \omega^{n-2} w_{xt} w_t n_x \, dS - \iint \omega_x^{n-2} w_{xt} w_t \, dx\, dt - \iint \omega^{n-2} w_{xxt} w_t \, dx\, dt$$

$$\leqq \oint_{\partial D} \omega^{n-2} (w_{xt} w_t n_x - w_{xx} w_t n_t) \, dS + \iint_D (\omega^{n-2} w_{xt}^2 + c\omega^{n-4} w_t^2) \, dx\, dt$$

$$+ \iint_D \omega_t^{n-2} w_t \, dx\, dt + \iint_D \omega^{n-2} w_{xx} w_{tt} \, dx\, dt \tag{3.15}$$

$$\leqq \oint_{\partial D} \omega^{n-2} (w_{xt} w_t n_x - w_{xx} w_t n_t) \, dS$$

$$+ \frac{1}{2} \iint_D \omega^{n-2} w_{xt}^2 \, dx\, dt + c \iint_D (\omega^{n-4} w_t^2 + \omega^{n-2} w_{xx}^2) \, dx\, dt$$

$$+ \iint_D \omega^{n-2} w_{xx} w_{tt} \, dx\, dt.$$

Substituting the equation into the last term of (3.15) we find

$$\iint_D \omega^{n-2} w_{xx} w_{tt} \, dx\, dt$$

$$= \iint_D \omega^{n-2} w_{xx} (w_{xxxx} - [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(\phi) - L\phi)]) \, dx\, dt$$

$$= \oint_{\partial D} \omega^{n-2} w_{xx} w_{xxx} n_x \, dS - \iint_D \omega^{n-2} w_{xxx}^2 \, dx\, dt - \iint_D \omega_x^{n-2} w_{xx} w_{xxx} \, dx\, dt$$

$$- \iint_D \omega^{n-2} w_{xx} [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(\phi) - L\phi)] \, dx\, dt \tag{3.16}$$

$$\leqq c \int_\Sigma (w_{xx}^2 + w_{xxx}^2) \, dS - \frac{1}{2} \iint \omega^{n-2} w_{xxx}^2 \, dx\, dt$$

$$+ c \iint_D [\omega^{n-4} w_{xx}^2 + \omega^{n-2} (w^2 + w_t^2 + w_x^2)] \, dx\, dt$$

$$+ \frac{1}{4} \iint_D \omega^{n-2} w_{xt}^2 \, dx\, dt + c \iint_D |\mathscr{F}(\phi) - L\phi|^2 \, dx\, dt.$$

Combining (3.14), (3.15) and (3.16) we see that for λ sufficiently large

$$\iint_D (\omega^n w_{xxt}^2 + \lambda^2 \omega^{n-2} w_{xt}^2) \, dx\, dt$$

$$\leqq c \int_\Sigma [w_{xxt}^2 + w_{xt}^2 + w_{xxx}^2 + w_{tt}^2 + \lambda^2 (w_{xt}^2 + w_t^2 + w_{xx}^2 + w_{xxx}^2)] \, dS$$

$$- \iint_D \omega^n w_{xxxx}^2 \, dx\, dt - \lambda^2 \iint_D \omega^{n-2} w_{xxx}^2 \, dx\, dt \tag{3.17}$$

$$+ c\lambda^2 \iint_D [\omega^{n-4} w_{xx}^2 + \omega^{n-2} (w^2 + w_t^2 + w_x^2)] \, dx\, dt + \iint_D |\mathscr{F}(\phi) - L\phi|^2 \, dx\, dt.$$

Finally, we need to examine

$$\iint_D [\omega^{n-4} w_{xx}^2 + \omega^{n-2}(w_t^2 + w_x^2)]\, dx\, dt.$$

Integrating by parts we find that

$$\begin{aligned}
&\iint_D [\omega^{n-4} w_{xx}^2 + \omega^{n-2}(w_t^2 + w_x^2)]\, dx\, dt \\
&\quad = \oint_{\partial D} [\omega^{n-4} w_{xx} w_x n_x + \omega^{n-2}(w w_t n_t + w w_x n_x)]\, dS \\
&\quad\quad - \iint_D (\omega_x^{n-4} w_{xx} w_x + \omega^{n-4} w_{xxx} w_x + \omega_t^{n-2} w w_t \\
&\quad\quad\quad + \omega^{n-2} w w_{tt} + \omega_x^{n-2} w w_x + \omega^{n-2} w w_{xx})\, dx\, dt \qquad (3.18) \\
&\quad \leqq c \int_\Sigma (w_{xx}^2 + w_x^2 + w_t^2 + w^2)\, dS \\
&\quad\quad + \iint_D \left(\frac{1}{4}\omega^{n-4} w_{xx}^2 + \frac{1}{2c}\omega^{n-2} w_{xxx}^2 + \frac{1}{2c}\omega^{n-2} w_t^2\right) dx\, dt \\
&\quad\quad + c \iint (\omega^{n-6} w_x^2 + \omega^{n-4} w^2)\, dx\, dt - \iint_D \omega^{n-2} w w_{tt}\, dx\, dt.
\end{aligned}$$

However,

$$\begin{aligned}
&\iint \omega^{n-6} w_x^2\, dx\, dt \\
&\quad = \oint_{\partial D} \omega^{n-6} w w_x n_x\, ds - \iint_D \omega_x^{n-6} w w_x\, dx\, dt - \iint \omega^{n-6} w w_{xx}\, dx\, dt \\
&\quad \leqq \int_\Sigma w^2 + w_x^2\, dS + \frac{1}{2}\iint_D \omega^{n-6} w_x^2\, dx\, dt \\
&\quad\quad + \frac{1}{4c}\iint_D \omega^{n-4} w_{xx}^2\, dx\, dt + c \iint_D \omega^{n-8} w^2\, dx\, dt.
\end{aligned}$$

Thus, (3.18) reduces to

$$\begin{aligned}
&\iint_D [\omega^{n-4} w_{xx}^2 + \omega^{n-2}(w_t^2 + w_x^2)]\, dx\, dt \\
&\quad \leqq c \int_\Sigma (w_{xx}^2 + w_x^2 + w_t^2 + w^2)\, dS + \frac{1}{2c}\iint_D \omega^{n-2} w_{xxx}^2\, dx\, dt \qquad (3.19) \\
&\quad\quad + c \iint_D \omega^{n-8} w^2\, dx\, dt - \iint_D \omega^{n-2} w w_{tt}\, dx\, dt.
\end{aligned}$$

Finally, observe that

$$
\begin{aligned}
&-\iint_D \omega^{n-2} w w_{tt}\, dx\, dt \\
&\quad = -\iint_D \omega^{n-2} w (w_{xxxx} - [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(\phi) - L\phi)])\, dx\, dt \\
&\quad \leqq \frac{1}{2}\frac{1}{c\lambda^2}\iint_D \omega^n w_{xxxx}^2\, dx\, dt + \frac{1}{c\lambda^2}\iint_D \omega^n (w_x^2 + w_t^2 + w_{xt}^2 + w_{xx}^2 + w_{xxx}^2)\, dx\, dt \\
&\qquad + c\lambda^2 \iint_D \omega^{n-4} w^2\, dx\, dt + c\iint_D |\mathscr{F}(\phi) - L\phi|^2\, dx\, dt.
\end{aligned}
$$

Using this estimate in (3.19) and applying (3.19) to (3.17) then yields the desired result when we take $n \geqq 8$. □

If we now let $\mathscr{M} = \{u : \|u\|_{L^2(\Omega)} \leqq M\}$, we can prove uniqueness and stability for problem P3.

As before we let $\phi = \tilde{u}$, the solution to a related problem in the sense that

$$
\begin{aligned}
\varepsilon^2 = \|\tilde{u} - \zeta\|_{H^3(\Sigma)}^2 &+ \left\|\frac{\partial \tilde{u}}{\partial n} - \eta\right\|_{H^2(\Sigma)}^2 + \left\|\frac{\partial^2 \tilde{u}}{\partial n^2} - \theta(t)\right\|_{H^1(\Sigma)}^2 \\
&+ \left\|\frac{\partial^2 u}{\partial n^3} - \kappa(t)\right\|_{L^2(\Sigma)}^2 + \|L\tilde{u} - \mathscr{F}\tilde{u}\|_{L^2(\Omega)}^2
\end{aligned}
$$

is small. Using the lemma, the following corollary follows from Theorem 3 by the same argument as used earlier.

COROLLARY. *If a solution u exists for* P3, *then it is unique and depends continuously on the data. In particular for $u \in \mathscr{M}$ a solution of* P3 *and $\tilde{u} \in \mathscr{M}$ defined above we have*

$$
\iint_{D_\alpha} (u - \tilde{u})^2\, dx\, dt \leqq c\left\{M^{2\alpha}\varepsilon^{2(1-\alpha)} + \frac{4}{[\log (M^2/\varepsilon^2)]^2}\right\}\frac{1}{\log (M^2/\varepsilon^2)}
$$

for $0 < \alpha \leqq 1$.

4. Extensions. In this section, we show how to extend the results of the previous two sections to weakly coupled systems of equations. The treatment of systems will be given explicitly for second order systems. The techniques and results for the two fourth order problems are completely analogous.

We wish to consider systems of the form

$$
\begin{aligned}
L_k u \equiv a_k(x,t)u_{k,xx} + b_k(t)u_{k,t} &= F_k(x, t, u_1, \cdots, u_K, u_{1,x}, \cdots, u_{K,x}) \\
&\equiv \mathscr{F}_k(u_1, \cdots, u_K) \quad \text{in } \Omega, \qquad k = 1, \cdots, K,
\end{aligned} \tag{4.1}
$$

$$
u_k = g_k \quad \text{on } \Sigma, \qquad k = 1, \cdots, K, \tag{4.2}
$$

$$
\frac{\partial u_k}{\partial n} = h_k \quad \text{on } \Sigma, \qquad k = 1, \cdots, K, \tag{4.3}
$$

where each a_k is positive and bounded away from zero and each F_k is Lipschitz in its last $2K$ arguments. We now extend our basic inequality to systems.

THEOREM 4. *Suppose* $(u_1, \cdots, u_K)$ *satisfies* (4.1), (4.2) *and* (4.3) *and suppose* $\phi_1, \cdots, \phi_k$ *are functions with bounded second derivatives in* Ω *and bounded first*

derivatives in $\Omega \cup \Sigma$. *Suppose further that there exists a function* f *satisfying*

$$D_\alpha \subset D_\beta, \qquad \Sigma_\alpha \subset \Sigma, \qquad 0 < \alpha < \beta \leqq 1,$$
$$f_x \geqq c > 0 \quad \text{in } \Omega,$$
$$(a_k f_x)_x \leqq -c < 0, \qquad (a_k f_x^3)_x \leqq -c < 0, \qquad k = 1, \cdots, K \text{ in } \Omega.$$

Then, for all α, $0 < \alpha \leqq 1$ *we have the estimate*

$$\begin{aligned}
&\lambda^3 \iint_{D_\alpha} \sum_{k=1}^{K} (u_k - \phi_k)^2 \, dx\, dt \\
&\qquad \leqq c \iint_{D_\alpha} \sum_{k=1}^{K} b_k^2(t)(u_{k,t} - \phi_{k,t})^2 \, dx\, dt \\
&\qquad\quad + c \iint_{D_\alpha} \sum_{k=1}^{K} [L_k \phi_k - \mathcal{F}_k(u_1, \cdots, u_k)] \, dx\, dt \\
&\qquad\quad + c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} \sum_{k=1}^{K} (\lambda^2 (u_k - \phi_k)^2 + (u_{k,x} - \phi_{k,x})^2 + \lambda^{-1}(u_{k,t} - \phi_{k,t})^2) \, dS
\end{aligned} \tag{4.4}$$

provided λ *is sufficiently large.*

Proof. The idea of the extension to systems is to begin the treatment of each equation separately, then combine everything to essentially uncouple the equations.

More precisely, we set $u_k - \phi_k = e^{\lambda f} v_k$ for $k = 1, \cdots, k$. Then form odd and even groupings for each equations as in the proof of Theorem 1. We then form

$$\begin{aligned}
&\iint_{D_\alpha} \sum_{k=1}^{k} a_k^{-1} (L_{ke} v) \cdot (L_{ko} v) \, dx\, dt \\
&\qquad \leqq c \iint_{D_\alpha} e^{-2\lambda f} \Big(\sum_{k=1}^{K} [(\mathcal{F}_k(u_1, \cdots, u_K) - \mathcal{F}_k(\phi_1, \cdots, \phi_K)) \\
&\qquad\qquad + (\mathcal{F}_k(u_1, \cdots, \phi_K) - L_k \phi_k)]^2 \Big) \, dx\, dt \\
&\qquad \leqq c \iint_{D_\alpha} e^{-2\lambda f} \sum_{k=1}^{K} ((u_k - \phi_k)^2 + (u_{k,x} - \phi_{k,x})^2) \, dx\, dt \\
&\qquad\quad + c \sum_{k=1}^{k} \iint_{D_\alpha} |\mathcal{F}_k(\phi_1, \cdots, \phi_K) - L_k \phi_k|^2 \, dx\, dt.
\end{aligned}$$

Note that we have used the Lipschitz behavior of the $\mathcal{F}_k$ in deriving this inequality.

Observe that we now have the same inequality we would obtain by treating k uncoupled equations (with the exception of the last term which is a data term). The remainder of the proof proceeds as though we were treating k separate equations. □

Uniqueness and continuous dependence on the data are also derived by essentially summing the results for one equation.

COROLLARY. *Let* $\mathcal{M} = \{(u_1, \cdots, u_K) : \sum_{k-1}^{K} \|u_k\|^2_{L^2(\Omega)} \leqq M^2\}$. *Suppose there exists* $(u_1, \cdots, u_K) \in \mathcal{M}$ *which satisfies* (4.1)–(4.3). *Then* $(u_1, \cdots, u_K)$ *is unique. Furthermore, if we substitute for* $(\phi_1, \cdots, \phi_K)$ *in* (4.4), *the functions* $(\tilde{u}_1, \cdots, \tilde{u}_K)$ *which solve a related problem, i.e., that*

$$\varepsilon^2 = \sum_{k=1}^{k} \left(\|g_k - \tilde{u}_k\|^2_{H^1(\Sigma)} + \left\| h_k - \frac{\partial \tilde{u}_k}{\partial n} \right\|^2_{L^2(\Sigma)} + \|L\tilde{u}_k - \mathcal{F}(\tilde{u}_1, \cdots, \tilde{u}_k)\|^2_{L^2(\Omega)} \right)$$

is small, then we have

$$\int\int_{D_\alpha} \sum_{k=1}^{K} |u_k - \tilde{u}_k|^2 \, dx\, dt \leqq c\left\{M^{2\alpha}\varepsilon^{2(1-\alpha)} + \frac{4M^2}{[\log (M^2/\varepsilon^2)]^2}\right\} \frac{1}{\log (M^2/\varepsilon^2)}$$

for $0 < \alpha \leqq 1$.

5. Examples and comments. Now that we have derived conditions on a function f which allow us to prove uniqueness and continuous dependence on the data estimates, it is beneficial to exhibit an f satisfying these conditions.

If we restrict our attention in (1.1) to constant coefficient case, then the conditions in Theorem 2 reduce to $f(x) = f_1(x) + f_2(t)$, where $f_1' \geqq c > 0$ and $f_1'' \leqq -c < 0$. For any point $(x_0, t_0) \in \mathcal{R}^2/\bar{\Omega}$ near Σ and any $\zeta_1, \zeta_2, \zeta_3 > 0$, if we let

$$f(x, t) = \zeta_1(t - t_0)^2 + \zeta_2 \ln [\zeta_3(x - x_0)]$$

then f clearly satisfies the above conditions. The level surfaces of this function are illustrated in Fig. 1.

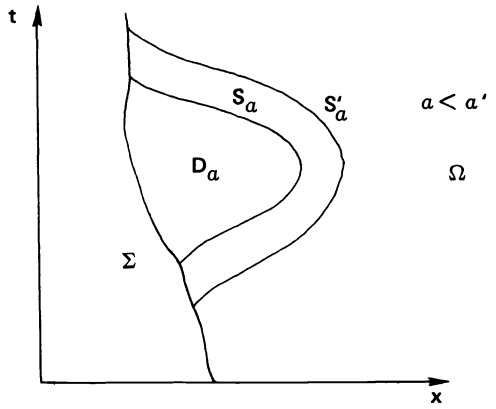


FIG. 1

This work raises two questions. The first concerns whether the results contained here can be extended to multiple space dimensions. For the second order case the answer is affirmative.

These results will be expostulated in Part II [8]. Our technique appears insufficient to answer this question for the generalization of problems P2 and P3 to multiple dimensions.

The second question concerns the continuous dependence estimate. It is well known (cf. [4]) that for the 1-d heat equation with a somewhat more restrictive stabilizing class, a stronger Hölder continuous dependence holds. The use of a weighted energy technique usually forces one to accept logarithmic continuous dependence on the data. Thus, one may well ask for what class of equations does a stronger continuous dependence result hold. As yet, no progress has been achieved on this front either.

Acknowledgment. The author would like to thank Professor L. E. Payne for many valuable discussions during the course of this research.

REFERENCES

[1] J. R. CANNON, *A Cauchy problem for the heat equation*, Ann. Mat. Pura Appl., 66 (1964), pp. 155–165.

[2] ———, *A priori estimates for continuation of the solution of the heat equation in the space variable*, Ann. Mat. Pura Appl., 65 (1964), pp. 377–388.

[3] J. R. CANNON AND J. DOUGLAS, JR., *The Cauchy problem for the heat equation*, SIAM J. Numer. Anal., 3 (1966), pp. 451–466.

[4] J. R. CANNON AND R. E. EWING, *A direct numerical procedure for the Cauchy problem for the heat equation*, J. Math. Anal. Appl., 56 (1976), pp. 7–17.

[5] R. E. EWING AND R. S. FALK, *Numerical Approximation of a Cauchy Problem for a Parabolic Partial Differential Equation*, MCR Technical Summary Report #1898, University of Wisconsin, Madison, November, 1978.

[6] F. GINSBERG, F., *On the Cauchy problem for the one-dimensional heat equation*, Math. Comp., 17 (1963), pp. 257–269.

[7] L. E. PAYNE, *On a priori bounds in the Cauchy problem for elliptic equations*, this Journal, 1 (1970), pp. 82–89.

[8] J. B. BELL, *The noncharacteristic Cauchy problem for a class of equations with time dependence, II. Multidimensional problems*, this Journal, this issue, pp. 778–797.

SIAM J. MATH. ANAL.
Vol. 12, No. 5, September 1981

0036-1410/81/1205-0009 $01.00/0

THE NONCHARACTERISTIC CAUCHY PROBLEM FOR A CLASS OF EQUATIONS WITH TIME DEPENDENCE II. MULTIDIMENSIONAL PROBLEMS*

JOHN B. BELL†

Abstract. The noncharacteristic Cauchy problem is treated for a class of equations which are second order in space and first order in time. The spatial part of the operator considered is multidimensional. A weighted energy technique is used to prove uniqueness and continuous dependence on the data within a restricted class of functions. The results are then generalized to treat systems. The technique is also applied to the noncharacteristic Cauchy problem for the time-dependent Navier–Stokes equations. Uniqueness and continuous dependence on the data within a restricted class of functions are again shown to hold.

1. Introduction. In [1] we treated the noncharacteristic Cauchy problem for some operators in one space dimension. In this work we present the generalization of the results for operators which are spatially second order elliptic in several space dimensions.

The situation in multiple dimensions is much more complex. The straightforward generalization of the weighted energy technique used in [1] leads to essentially contradictory conditions on the function f defining the surfaces used in the analysis. To surmount this difficulty it is necessary to distinguish one of the spatial variables. More precisely, we will only consider functions f which depend on one space variable and time.

Needless to say, in this framework Cartesian coordinates yield satisfying results only in cases with unusual problem geometries. We must, therefore, introduce other coordinate systems to obtain meaningful results.

The type of system we need is a generalized polar coordinate system where $\omega_1, \cdots, \omega_{n-1}$ denote the coordinates on a surface and ξ denotes the coordinate which is orthogonal to the surface. In our new coordinate system, the Laplace operator takes the form[1]

$$\Delta u = \bar{\rho}\frac{\partial^2 u}{\partial \xi^2} + \bar{\gamma}^{ij}u_{,ij} + \mu\frac{\partial u}{\partial \xi} + \kappa_i u_{,i},$$

where $\bar{\rho}$, $\bar{\gamma}^{ij}$, μ and κ_i depend, in general, on ξ, $\omega_1, \cdots, \omega_{n-1}$. Furthermore, the ellipticity of the Laplacian at each point is equivalent to the conditions

$$\bar{\rho} \geqq c > 0 \quad \text{and} \quad \bar{\gamma}^{ij}\nu_i\nu_j \geqq c\nu_i\nu_i$$

for arbitrary $(n-1)$-vectors $(\nu_1, \cdots, \nu_{n-1})$.

Motivated by these considerations, we will study the noncharacteristic Cauchy problem for the more general equation

$$(1.1)\quad Lu \equiv \rho u_{,\xi\xi} + \gamma^{ij}u_{,ij} + bu_{,t} = F(t, \xi, \omega_1, \cdots, \omega_{n-1}u, u_{,\xi}, u_{,1} \cdots, u_{,n-1}) \equiv \mathcal{F}(u),$$

* Received by the editors February 1, 1980, and in revised form December 16, 1980. This work was supported by a National Science Foundation Graduate Research Fellowship and the NSWC IR Fund.

† Naval Surface Weapons Center-R44, White Oak, Maryland 20910.

[1] Summation convention is used throughout, and a comma is used to denote differentiation. In a $(\xi, \omega, \cdots, \omega_{n-1})$-coordinate system summation is from 1 to $n-1$. In Cartesian coordinates, summation is from 1 to n. In other cases the limits of the summation are clear from the context. When an explicit summation symbol is present, the summation convention does not apply. Note that for notational convenience $v_{,\xi}^2 \equiv (\partial v/\partial \xi)^2$.

where ρ and the γ^{ij} depend on $\xi, \omega_1, \cdots, \omega_{n-1}$, and t and b depend on t. In addition, we require that ρ, γ^{ij} and b be bounded and have bounded derivatives, and that the following technical conditions be satisfied:

$$\rho \geqq c > 0,$$

$$\gamma^{ij} = \gamma^{ji} \quad \text{and} \quad \gamma^{ij}\nu_i\nu_j \geqq c\nu_i\nu_i, \qquad (\nu_1, \cdots, \nu_{n-1}) \in \mathscr{R}^{n-1}$$

and

$$|F(t, \xi, \omega_1, \cdots, \omega_{n-1}, u^1, \nu_1^1, \cdots, \nu_n^1) - F(t, \xi, \omega_1, \cdots, \omega_{n-1}, u^2, \nu_1^2, \cdots, \nu_n^2)|$$
$$\leqq c(|u^1 - u^2| + |\nu_1^1 - \nu_1^2| + \cdots + |\nu_{n-1}^1 - \nu_{n-1}^2| + |\nu_n^1 - \nu_n^2|).$$

It should be emphasized that although ρ and the γ^{ij} were motivated by the expression for the Laplacian in alternate coordinate systems, we do not intend that they be interpreted solely as arising from the metrical coefficients of some coordinate system. Rather, the theory is intended to fit any equation which can be transformed to the form of (1.1).

We are now ready to pose the noncharacteristic Cauchy problem for (1.1). We let Ω be a domain in $\mathscr{R}^{n+1}$ and $\Sigma \subset \partial\Omega$. We require that Σ be closed and nowhere characteristic. The second condition is equivalent to saying that the spatial components of the normal to Σ never vanish simultaneously at any point of Σ. The first condition guarantees that Σ does not degenerate at its boundary. We may now pose the problem[2]

$$\text{P1} \quad \begin{cases} \rho u_{,\xi\xi} + \gamma^{ij}u_{,ij} + bu_{,t} = \mathscr{F}(u) & \text{in } \Omega, \\ u = g & \text{on } \Sigma, \\ \dfrac{\partial u}{\partial n} = h & \text{on } \Sigma, \end{cases}$$

where ρ, γ^{ij}, b and F satisfy the properties specified earlier.

As in [1], a weighted energy technique is used to prove uniqueness and logarithmic continuous dependence on the data within a restricted class of functions. As before, the method involves the use of the level sets of a function f. Hence, we introduce, for a function $f(t, \xi)$,

$$D_\alpha \equiv \{(\xi, \omega_1, \cdots, \omega_{n-1}, t): f(t, \xi) \leqq \alpha\} \cap \Omega,$$

$$S_\alpha \equiv \partial D_\alpha \cap \Omega,$$

$$\Sigma_\alpha \equiv \partial D_\alpha \cap \Sigma.$$

It is also assumed that $\partial D_\alpha = S_\alpha \cup \Sigma_\alpha$.

In § 3 we indicate how to extend the uniqueness and continuous dependence results to systems of equations which are coupled in their lower order derivative terms. The above ideas are modified to treat the noncharacteristic Cauchy problem for the time-dependent Navier–Stokes equations in § 4. In the final section we exhibit a suitable coordinate system and function f satisfying the hypotheses of the theorem.

2. Problem 1. The theorem which follows shows that under certain conditions on the coordinate system, the function f and an auxiliary weighting function, it is possible to bound certain integrals over D_α in terms of available data. We will then use this bound to deduce uniqueness and continuous dependence on the data for problem P1. Thus we prove the following estimate.

[2] The arguments of all functions are suppressed throughout the remainder of this section.

THEOREM 1. *Suppose u satisfies (1.1) in Ω. Let ϕ be any function with bounded second derivatives in Ω and bounded first derivatives in $\Omega \cup \Sigma$. Suppose also that there exists a function $f(t, \xi)$ and a function $\sigma(t, \xi, \omega_1, \cdots \omega_{n-1})$ which satisfy:*

1) $D_\alpha \subset D_\beta$, $\Sigma_\alpha \subset \Sigma$, $0 < \alpha \leqq 1$,
2) $f_{,\xi} \geqq c > 0$ *in* D_1,
3) $(\sigma^2 \rho^2 f^3{}_{,\xi})_{,\xi} \leqq -c < 0$, $(\sigma^2 \rho^2 f_{,\xi})_{,\xi} \leqq -c < 0$ *in* D_1,
4) $(\sigma^2 \rho f_{,\xi} \gamma^{ij})_{,\xi}$ *is positive semidefinite in* D_1,
5) $(\sigma^2 f_{,\xi} \rho \gamma^{ij})_{,j} = 0$ *in* D_1,
6) $(\sigma^2 b \gamma^{ij})_{,j} = 0$ *in* D_1,
7) $(\sigma^2 \rho)_{,\xi} \geqq c > 0$ *in* D_1.

Then, for all $\alpha \in (0, 1]$, we have the estimate

$$
\begin{aligned}
&\lambda^3 \iint_{D_\alpha} (u-\phi)^2 \, dm \, dt \\
&\quad \leqq c \iint_{D_\alpha} \lambda (u_{,i} - \phi_{,i})(u_{,i} - \phi_{,i}) + (u_{,t} - \phi_{,t})^2 \, dm \, dt \\
&\quad + c e^{2\lambda\alpha} \int_{\Sigma_\alpha} (\lambda^2 (u-\phi)^2 + (u_{,\xi} - \phi_{,\xi})^2 + (u_{,i} - \phi_{,i})(u_{,i} - \phi_{,i}) + \lambda^{-1} (u_{,t} - \phi_{,t})^2) \, dS \\
&\quad + c \lambda^{-1/2} e^{2\lambda\alpha} \iint_{D_\alpha} |\mathscr{F}(\phi) - L\phi|^2 \, dm \, dt
\end{aligned} \tag{2.1}
$$

for λ sufficiently large.

Proof. We let $u - \phi = e^{\lambda f} v$. Substituting for u in (1.1) and multiplying through by σ yields

$$
\begin{aligned}
&[\lambda^2 \tilde{\rho} f^2{}_{,\xi} + O(\lambda)] v + 2\lambda \tilde{\rho} f_{,\xi} v_{,\xi} + \tilde{\rho} v_{,\xi\xi} + \tilde{\gamma}^{ij} v_{,ij} + \tilde{b} v_{,t} \\
&\quad = \sigma e^{-\lambda f} [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(\phi) - L\phi)]
\end{aligned}
$$

where $\tilde{\rho} = \sigma\rho$, $\tilde{\gamma}^{ij} = \sigma\gamma^{ij}$ and $\tilde{b} = \sigma b$. We now define odd and even parts of the operator as

$$
L_e v = [\lambda^2 \tilde{\rho} f^2{}_{,\xi} + O(\lambda)] v + \tilde{\rho} v_{,\xi\xi} + \tilde{\gamma}^{ij} v_{,ij}, \qquad L_o v = 2\lambda \tilde{\rho} f_{,\xi} v_{,\xi} + \tilde{b}_{,t}.
$$

We now form the inequality

$$
\begin{aligned}
&\iint_{D_\alpha} \{L_e v \times L_o v + \tfrac{1}{2}(L_o v)^2 - \lambda^{1/2} (L_o v + L_e v) v\} \, dm \, dt \\
&\quad \leqq \iint_{D_\alpha} \{\tfrac{1}{2}\sigma^2 e^{-\lambda f} (|\mathscr{F}(u) - \mathscr{F}(\phi)|^2 + |\mathscr{F}(\phi) - L\phi|^2) \\
&\qquad - \lambda^{1/2} \sigma e^{-\lambda f} [(\mathscr{F}(u) - \mathscr{F}(\phi)) + (\mathscr{F}(\phi) - L\phi)] v\} \, dm \, dt.
\end{aligned} \tag{2.2}
$$

We now expand the first term on the left of (2.2) and integrate the result by parts. Thus,

$$
\begin{aligned}
&\iint_{D_\alpha} \{L_e v \times L_o v\} \, dm \, dt \\
&\quad = \iint_{D_\alpha} [2\lambda^3 \tilde{\rho}^2 f^3{}_{,\xi} + O(\lambda^2)] v v_{,\xi} \, dm \, dt + \iint_{D_\alpha} [\lambda^2 \tilde{\rho} \tilde{b} f^2{}_{,\xi} + O(\lambda)] v v_{,t} \, dm \, dt \\
&\qquad + \iint_{D_\alpha} 2\lambda \tilde{\rho}^2 f_{,\xi} v_{,\xi\xi} v_{,\xi} \, dm \, dt + \iint_{D_\alpha} 2\lambda \tilde{\rho} f_{,\xi} \tilde{\gamma}^{ij} v_{,ij} v_{,\xi} \, dm \, dt
\end{aligned}
$$

$$
\begin{aligned}
&+\int\int_{D_\alpha}\tilde{\rho}\tilde{b}v_{,\xi\xi}v_{,t}\,dm\,dt+\int\int_{D_\alpha}\tilde{\gamma}^{ij}\tilde{b}v_{,ij}v_{,t}\,dm\,dt\\
&=\oint_{\partial D_\alpha}[\lambda^3\tilde{\rho}^2f^3{}_{,\xi}+O(\lambda^2)]n_\xi v^2\,dS-\int\int[\lambda^3(\tilde{\rho}^2f^3{}_{,\xi})_{,\xi}+O(\lambda^2)]v^2\,dm\,dt\\
&+\oint_{\partial D_\alpha}\frac{\lambda^2}{2}\tilde{\rho}\tilde{b}f^2{}_{,\xi}v^2n_t\,dS-\int\int_{D_\alpha}O(\lambda^2)v^2\,dm\,dt+\oint_{\partial D_\alpha}\lambda\tilde{\rho}^2f_{,\xi}v^2{}_{,\xi}n_\xi\,dS\\
&-\int\int_{D_\alpha}\lambda(\tilde{\rho}^2f_{,\xi})_{,\xi}v^2{}_{,\xi}\,dm\,dt+2\oint_{\partial D_\alpha}\lambda\tilde{\rho}f_{,\xi}\tilde{\gamma}^{ij}v_{,i}v_{,\xi}n_j\,dS\\
&-\oint_{\partial D_\alpha}\lambda\tilde{\rho}f_{,\xi}\tilde{\gamma}^{ij}v_{,i}v_{,j}n_\xi\,dS-2\int\int_{D_\alpha}\lambda(\tilde{\rho}f_{,\xi}\tilde{\gamma}^{ij})_{,j}v_{,i}v_{,\xi}\,dm\,dt\\
&+\int\int_{D_\alpha}\lambda(\tilde{\rho}f_{,\xi}\gamma^{ij})_{,\xi}v_{,i}v_{,j}\,dm\,dt+\oint_{\partial D_\alpha}\tilde{\rho}\tilde{b}v_{,\xi}v_{,t}n_\xi\,dS-\oint_{\partial D_\alpha}\tfrac{1}{2}\tilde{\rho}\tilde{b}v^2{}_{,\xi}n_t\,dS\\
&-\int\int_{D_\alpha}(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}\,dm\,dt+\frac{1}{2}\int\int_{D_\alpha}(\tilde{\rho}\tilde{b})_{,t}v^2{}_{,\xi}\,dm\,dt+\oint_{\partial D_\alpha}\tilde{\gamma}^{ij}\tilde{b}v_{,i}v_{,t}n_j\,dS\\
&-\frac{1}{2}\oint_{\partial D}\tilde{\gamma}^{ij}\tilde{b}v_{,i}v_{,j}n_t\,dS-\int\int_{D_\alpha}(\tilde{\gamma}^{ij}\tilde{b})_{,j}v_{,i}v_{,t}\,dm\,dt+\frac{1}{2}\int\int_{D_\alpha}(\tilde{\gamma}^{ij}\tilde{b})_{,t}v_{,i}v_{,j}\,dm\,dt.
\end{aligned}
\tag{2.3}
$$

The third term on the left of (2.2) can also be expanded and integrated by parts yielding

$$
\begin{aligned}
&\lambda^{1/2}\int\int_{D_\alpha}(L_e v+L_o v)v\,dm\,dt\\
&\quad=\lambda^{1/2}\Big\{\int\int_{D_\alpha}[\lambda^2\tilde{\rho}f^2{}_{,\xi}+O(\lambda)]v^2\,dm\,dt-\int\int_{D_\alpha}\tilde{\rho}v^2{}_{,\xi}\,dm\,dt\\
&\qquad-\int\int_{D_\alpha}\tilde{\gamma}^{ij}v_{,i}v_{,j}\,dm\,dt+\oint_{\partial D_\alpha}[\lambda\tilde{\rho}f_{,\xi}n_\xi+O(1)]v^2\,dS\\
&\qquad+\oint_{\partial D_\alpha}\tilde{\rho}v_{,\xi}vn_\xi\,dS+\int_{\partial D_\alpha}\tilde{\gamma}^{ij}v_{,i}vn_j\,dS\Big\}.
\end{aligned}
\tag{2.4}
$$

Substituting (2.3) and (2.4) into (2.2) and using the Lipschitz properties of $\mathscr{F}$, we find that

$$
\begin{aligned}
&\lambda^3\oint_{\partial D_\alpha}[\tilde{\rho}^2f^3{}_{,\xi}n_\xi+O(\lambda^{-1})]v^2\,dS+\frac{\lambda}{2}\oint_{\partial D_\alpha}[\tilde{\rho}^2f_{,\xi}n_\xi+O(\lambda^{-1/2})]v^2{}_{,\xi}\,dS\\
&\quad-\lambda\oint_{\partial D_\alpha}[c+O(\lambda^{-1/2})]v_{,i}v_{,i}|\nabla f|^{-1}\,dS-c\oint_{\partial D_\alpha}\tilde{b}^2(t)v^2{}_{,t}|\nabla f|^{-1}\,dS\\
&\quad-\lambda^3\int\int_{D_\alpha}[(\tilde{\rho}^2f^3{}_{,\xi})_{,\xi}+O(\lambda^{-1/2})]v^2\,dm\,dt\\
&\quad-\lambda\int\int_{D_\alpha}[(\tilde{\rho}^2f_{,\xi})_{,\xi}+O(\lambda^{-1/2})]v^2{}_{,\xi}\,dm\,dt\\
&\quad-2\lambda\int\int_{D_\alpha}(\tilde{\rho}f_{,\xi}\tilde{\gamma}^{ij})_{,j}v_{,i}v_{,\xi}\,dm\,dt+\lambda\int\int_{D_\alpha}(\tilde{\rho}f_{,\xi}\tilde{\gamma}^{ij})_{,\xi}v_{,i}v_{,j}\,dm\,dt
\end{aligned}
\tag{2.5}
$$

$$+\lambda^{1/2}\iint_{D_\alpha}[c+O(\lambda^{-1/2})]v_{,i}v_{,i}\,dm\,dt-\iint_{D_\alpha}(\tilde{b}\tilde{\gamma}^{ij})_{,j}v_{,i}v_{,t}\,dm\,dt$$

$$\leqq\iint_{D_\alpha}(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}\,dm\,dt-\tfrac{1}{2}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2\,dm\,dt$$

$$+c\lambda^{1/2}\iint_{D_\alpha}e^{-2\lambda f}|\mathscr{F}(\phi)-L\phi|^2\,dm\,dt.$$

Applying assumptions 2)–6) to the volume terms on the left side of the inequality, we find that they are all greater than or equal to zero for λ sufficiently large. Thus, (2.5) becomes

$$\begin{aligned}&\lambda^3\oint_{\partial D_\alpha}[\tilde{\rho}^2f^3{}_{,\xi}n_\xi+O(\lambda^{-1})]v^2\,dS+\frac{\lambda}{2}\oint_{\partial D_\alpha}[\tilde{\rho}f_{,\xi}n_\xi+O(\lambda^{-1/2})]v^2{}_{,\xi}\,dS\\&\leqq c\lambda\oint_{\partial D_\alpha}v_{,i}v_{,i}|\nabla f|^{-1}\,dS+\int_{\partial D_\alpha}\tilde{b}^2(t)v^2{}_{,t}|\nabla f|^{-1}\,dS\\&+\iint_{D_\alpha}\lambda^{1/2}\,e^{-2\lambda f}(\mathscr{F}(\phi)-L\phi)^2\,dm\,dt\\&+\iint_{D_\alpha}[(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}-\tfrac{1}{2}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2]\,dm\,dt.\end{aligned}\tag{2.6}$$

We now claim that assumption 7) implies that

$$\iint_{D_\alpha}[(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}-\tfrac{1}{2}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2]\,dm\,dt\leqq 0,$$

provided λ is sufficiently large. To see this, observe that

$$(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}\leqq\frac{(\tilde{\rho}\sigma)_{,\xi}}{4\lambda\tilde{\rho}f_{,\xi}\sigma}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2,$$

since by 7) $(\tilde{\rho}\sigma)_{,\xi}\geqq c>0$ and the denominator is, by definition, positive. Hence, we may conclude that

$$(\tilde{\rho}\tilde{b})_{,\xi}v_{,\xi}v_{,t}-\tfrac{1}{2}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2\leqq\left(\frac{(\tilde{\rho}\sigma)_{,\xi}}{2\lambda\tilde{\rho}f_{,\xi}\sigma}-1\right)\tfrac{1}{2}(2\lambda\tilde{\rho}f_{,\xi}v_{,\xi}+\tilde{b}v_{,t})^2\leqq 0$$

for λ large.

Therefore, we may conclude from (2.6) that for λ sufficiently large

$$\begin{aligned}&\lambda^3\int_{S_\alpha}v^2|\nabla f|^{-1}\,dS\\&\leqq c\int_{S_\alpha}(\lambda v_{,i}v_{,i}+b^2(t)v_t^2)|\nabla f|^{-1}\,dS+c\int_{S_\alpha}(\lambda^3v^2+\lambda v^2{}_{,\xi}+\lambda v_{,i}v_{,i}+v^2{}_{,t})\,dS\\&+\lambda^{1/2}\iint_{D_\alpha}|L\phi-\mathscr{F}(\phi)|^2\,e^{-2\lambda f}\,dm\,dt.\end{aligned}$$

Substituting $e^{-\lambda f}(u-\phi)=v$ then yields

$$\lambda^3\int_{S_\alpha}(u-\phi)^2|\nabla f|^{-1}\,dS$$

$$\leqq c\int_{S_\alpha}(\lambda(u-\phi)_{,i}(u-\phi)_{,i}+b^2(u_{,t}-\phi_{,t})^2)|\nabla f|^{-1}\,dS$$

$$+c\,e^{2\lambda\alpha}\int_{\Sigma_\alpha}(\lambda^3(u-\phi)^2+\lambda(u_{,\xi}-\phi_{,\xi})^2+\lambda(u-\phi)_{,i}(u-\phi)_{,i}+(u_{,t}-\phi_{,t})^2)\,dS$$

$$+\int\int_{D_\alpha}\lambda^{1/2}\,e^{-2\lambda f}|L^\phi-\mathcal{F}(\phi)|^2\,dm\,dt.$$

Integrating with respect to α gives

$$\lambda^3\int\int_{D_\alpha}(u-\phi)^2\,dm\,dt$$

$$\leqq c\int\int_{D_\alpha}(\lambda(u-\phi)_{,i}(u-\phi)_{,i}+b^2(u_{,t}-\phi_{,t})^2)\,dm\,dt$$

$$+c\,e^{2\lambda\alpha}\int\int_{\Sigma_\alpha}\left(\lambda^2(u-\phi)^2+(u_{,\xi}-\phi_{,\xi})^2+(u-\phi)_{,i}(u-\phi)_{,i}+\frac{(u_{,t}-\phi_{,t})^2}{\lambda}\right)dS$$

$$+\lambda^{-1/2}\,e^{2\lambda\alpha}\int\int_{D_\alpha}|L\phi-\mathcal{F}(\phi)|^2\,dm\,dt$$

for λ sufficiently large. □

We are now almost ready to use the estimate of the theorem to prove uniqueness and continuous dependence on the data. First, however, we prove a lemma bounding

$$\int\int_{D_\alpha}[\lambda(u-\phi)_{,i}(u-\phi)_{,i}+b^2(u_{,t}-\phi_{,t})^2]\,dm\,dt$$

in terms of $\|u-\phi\|^2_{L^2(\Omega)}$ and data terms.

LEMMA. (The lemma is stated and proven in Cartesian coordinates.) *Suppose u satisfies*

$$Lu\equiv a_{ij}(t,x_1,\cdots,x_n)u_{,ij}+b(t)u_{,t}=F\left(t,x_1,\cdots,x_n,u,\frac{\partial u}{\partial x_1},\cdots,\frac{\partial u}{\partial x_n}\right)\equiv\mathcal{F}(u)$$

in Ω where $a_{ij}\nu_i\nu_i\geqq c\nu_i\nu_i$ for all $\nu_1,\cdots,\nu_n$, at each point of Ω and $a_{ij}=a_{ji}$ and ϕ is a function as defined earlier. In addition we assume the a_{ij} and b are bounded and possess bounded derivatives in Ω. F is assumed to be Lipschitz in its last $n+1$ arguments. Then,

$$\int\int_{D_\alpha}(b^2(t)(u_{,t}-\phi_{,t})^2+\lambda(u-\phi)_{,i}(u-\phi)_{,i})\,dm\,dt$$

$$\leqq\lambda\Big\{c\int_\Sigma((u_{,t}-\phi_{,t})^2+(u-\phi)_{,i}(u-\phi)_{,i}+u^2)\,dS$$

$$+c\int\int_D(u-\phi)^2\,dm\,dt+\int\int_D|L\phi-\mathcal{F}(\phi)|^2\,dm\,dt\Big\},$$

where $D\subset\Omega$ such that $\bar{D}_\alpha\cap\Omega\subset D$ and $\Sigma\subset\partial D\cap\partial\Omega$.

Proof. Let ω be a C^∞ cutoff function such that $\omega\equiv 1$ on $\bar{D}_\alpha$ and $\omega\equiv 0$ on $\Omega\backslash D$.

Using the equation and integrating by parts we have, substituting $w = u - \phi$,

$$
\begin{aligned}
&\iint_{D_\alpha} (b^2(t)w^2{}_{,t} + \lambda w_{,i}w_{,i})\, dm\, dt \\
&\quad \leqq \iint_D (\omega^4 b^2(t)w^2{}_{,t} + \lambda\omega^2 w_{,i}w_{,i})\, dm\, dt \\
&\quad = -\iint_D \omega^4 b(t)w_{,t}a_{ij}w_{,ij}\, dm\, dt + \iint_D \lambda\omega^2 w_{,i}w_{,i}\, dm\, dt \\
&\qquad + \iint_D \omega^4 b(t)w_{,t}[(\mathcal{F}(u) - \mathcal{F}(\phi)) + (\mathcal{F}(\phi) - L\phi)]\, dm\, dt \\
&\quad = -\oint_{\partial D} \omega^4 b(t)w_{,t}a_{ij}w_{,i}n_j\, dS + \frac{1}{2}\oint_{\partial D} \omega^4 b(t)a_{ij}w_{,i}w_{,j}n_t\, dS \\
&\qquad + \iint_D (\omega^4 b(t)a_{ij})_{,j}w_{,t}w_{,i}\, dm\, dt - \frac{1}{2}\iint_D (\omega^4 b(t)a_{ij})_{,t}w_{,i}w_{,j}\, dm\, dt \\
&\qquad + \lambda \iint_D \omega^2 w_{,i}w_{,i}\, dm\, dt + \iint_D \omega^4 b(t)w_{,t}[(\mathcal{F}(u) - \mathcal{F}(\phi)) + (\mathcal{F}(\phi) - L\phi)]\, dm\, dt \\
&\quad \leqq \frac{1}{2}\iint_D \omega^2 b^2(t)w^2{}_{,t}\, dm\, dt \\
&\qquad + c\lambda \iint_D \omega^2 w_{,i}w_{,i}\, dm\, dt + c\iint_D w^2\, dm\, dt + c\int_\Sigma (w^2{}_{,t} + \lambda w_{,i}w_{,i})\, dS \\
&\qquad + c\iint_D (|\mathcal{F}(u) - \mathcal{F}(\phi)|^2 + |\mathcal{F}(\phi) - L\phi|^2)\, dm\, dt.
\end{aligned} \tag{2.7}
$$

To complete the proof we must show how to estimate the gradient of w in terms of w plus data terms. We again make use of the equation and the ellipticity of $a_{ij}\,\partial^2/\partial x_i\,\partial x_j$ to obtain

$$
\begin{aligned}
&\iint_D \omega^2 w_{,i}w_{,i}\, dm\, dt \\
&\quad \leqq c\iint_D \omega^2 a_{ij}w_{,i}w_{,j}\, dm\, dt \\
&\quad = c\oint_{\partial D} \omega^2 a_{ij}ww_{,j}n_i\, dS - c\iint_D \omega\omega_{,i}a_{ij}ww_{,j}\, dm\, dt - c\iint_D \omega^2 wa_{ij}w_{,ij}\, dm\, dt \\
&\quad \leqq c\oint_{\partial D} \omega^2 a_{ij}ww_{,j}n_i\, dS + c\iint_D w^2\, dm\, dt + \frac{1}{4}\iint_D \omega^2 w_{,i}w_{,i}\, dm\, dt \\
&\qquad + c\iint_D \omega^2 b(t)ww_{,t}\, dm\, dt \\
&\qquad + c\iint_D w[(\mathcal{F}(u) - \mathcal{F}(\phi)) + (\mathcal{F}(\phi) - L\phi)]\, dm\, dt
\end{aligned} \tag{2.8}
$$

$$\leqq c\oint_{\partial D}(\omega^2 a_{ij}ww_{,j}n_i+\omega^2 b(t)w^2 n_t)\,dS+\frac{1}{2}\int\int_D \omega^2 w_{,i}w_{,i}\,dm\,dt$$

$$+c\int\int_D w^2\,dm\,dt+c\int\int_D(|\mathscr{F}(u)-\mathscr{F}(\phi)|^2+|\mathscr{F}(\phi)-L\phi|^2)\,dm\,dt.$$

Combining (2.8) with (2.7) and using the Lipschitz behavior of $\mathscr{F}$ completes the proof of the lemma. □

Let $\mathscr{M}=\{u:\|u\|_{L^2(\Omega)}\leqq M\}$. Using the lemma, we now obtain uniqueness and continuous dependence on the data within $\mathscr{M}$ as a corollary of the theorem by letting $\phi=\tilde{u}$ which satisfies a problem close to P1 in the sense that

$$\varepsilon^2=\|g-\tilde{u}\|^2_{H^1(\Sigma)}+\left\|h-\frac{\partial\tilde{u}}{\partial n}\right\|^2_{L^2(\Sigma)}+\|\mathscr{F}(\tilde{u})-L\tilde{u}\|^2_{L^2(\Omega)}$$

is small. In this situation we have the corollary:

COROLLARY. *If there exists a coordinate system, a function f, and a function σ satisfying* 1)–7) *of the previous theorem for $\alpha\in(0,1]$ then, if* P1 *has a solution, it is unique and depends continuously on the data in D_1 within the class $\mathscr{M}$. In particular, for u, $\tilde{u}\in\mathscr{M}$ as described above we have*

$$\int\int_{D_\alpha}(u-\tilde{u})^2\,dm\,dt\leqq c\left\{\varepsilon^{2(1-\alpha)}M^{2\alpha}+\frac{2M^2}{\log(M^2/\varepsilon^2)}\right\}\frac{1}{\log(M^2/\varepsilon^2)},\qquad 0<\alpha\leqq 1.$$

Proof. By use of the lemma, in the above situation the estimate in the theorem becomes

$$\lambda^3\int\int_{D_\alpha}(u-\tilde{u})^2\,dm\,dt\leqq c\lambda M^2+c\lambda^2 e^{2\lambda\alpha}\varepsilon^2.$$

For uniqueness, $\varepsilon=0$, so let $\lambda\to\infty$. For continuous dependence, setting $\lambda=\frac{1}{2}\log(M^2/\varepsilon^2)$ yields the desired result. □

3. Systems. In this section, we will extend the results of the previous section to systems of equations which are coupled in their lower order derivative terms. More precisely, we wish to consider systems of the form

$$\text{P2}\qquad\begin{cases}\rho_k u_{k,\xi\xi}+\gamma_k^{ij}u_{k,ij}+b_k u_{k,t}=\mathscr{F}_k(u_1,\cdots,u_K) & \text{in }\Omega\text{ for }k=1,\cdots,K,\\ u_k=g_k & \text{on }\Sigma\text{ for }k=1,\cdots,K,\\ \dfrac{\partial u_k}{\partial n}=h_k & \text{on }\Sigma\text{ for }k=1,\cdots,K,\end{cases}$$

where

$$\mathscr{F}_k(u_1,\cdots,u_k)\equiv F_k(t,\xi,\omega_1,\cdots,\omega_{n-1},u_1,\cdots,u_K,u_{1,\xi},\cdots,u_{K,\xi},$$
$$u_{1,1},\cdots,u_{K,n-1})\quad\text{for }k=1,\cdots,K$$

for functions F_k which are each Lipschitz in their last $K(n+1)$ arguments. The following analogue of Theorem 1 then holds:

THEOREM 2. *Let $u_1,\cdots,u_K$ solve* P2 *and let $\phi_1,\cdots,\phi_K$ be K functions possessing bounded second order derivatives in Ω and bounded first order derivatives in $\Omega\cup\Sigma$. Suppose further that there exists a function f and functions $\sigma_1,\cdots,\sigma_K$ satisfying* 1)–7)

of Theorem 1 *for each k. Then, for all* $\alpha \in (0, 1]$ *we have the estimate*

$$\lambda^3 \int\int_{D_\alpha} (u_k - \phi_k)(u_k - \phi_k)\, dx\, dt$$

$$\leqq c \int\int_{D_\alpha} [\lambda (u_{k,i} - \phi_{k,i})(u_{k,i} - \phi_{k,i}) + \sum_{k=1}^{K} b_k^2(t)(u_{k,t} - \phi_{k,t})^2]\, dm\, dt$$

$$+ c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} [\lambda^2 (u_k - \phi_k)(u_k - \phi_k) + (u_{k,\xi} - \phi_{k,\xi})(u_{k,\xi} - \phi_{k,\xi})$$

$$+ (u_{k,i} - \phi_{k,i})(u_{k,i} - \phi_{k,i}) + \lambda^{-1}(u_{k,t} - \phi_{k,t})(u_{k,t} - \phi_{k,t})]\, dS,$$

provided λ *is sufficiently large.*

Proof. The proof is essentially a summation of the proof of Theorem 1. For each k we set $u_k = \phi_k + e^{\lambda f} v_k$ and expand each equation as done before. We then define (without summation convention)

$$L_{ke} v_k = [\lambda^2 \tilde{\rho}_k f^2{}_{,\xi} + O(\lambda)] v_k + \tilde{\rho}_k v_{k,\xi\xi} + \tilde{\gamma}_k^{ij} v_{k,ij}$$

and

$$L_{ko} v_k = 2\lambda \tilde{\rho}_k f_{,\xi} v_{k,\xi} + \tilde{b}_k(t) v_{k,t},$$

where

$$\tilde{\rho}_k = \rho_k \sigma_k, \tilde{\gamma}_k^{ij} = \gamma_k^{ij} \sigma_k \quad \text{and} \quad \tilde{b}_k(t) = \tilde{b}_k(t) \sigma_k.$$

We now form

$$\int\int_{D_\alpha} \left\{ \sum_{k=1}^{K} (L_{ke} v_k \times L_{ko} v_k) + \tfrac{1}{2} \sum_{k=1}^{K} (L_{ko} v_k)^2 - \tfrac{1}{2} \sum_{k=1}^{K} (L_{ke} v_k + L_{ko} v_k) v_k \right\} dm\, dt$$

$$\leqq \sum_{k=1}^{K} \int\int_{D_\alpha} \tfrac{1}{2} \sigma_k^2 e^{-\lambda f} (|\mathscr{F}_k(u_1, \cdots, u_k) - \mathscr{F}_k(\phi_1, \cdots, \phi_K)|^2$$

$$+ |\mathscr{F}_k(\phi_1, \cdots, \phi_K) - L\phi_K|^2)\, dm\, dt.$$

After applying the Lipschitz property of $\mathscr{F}_1, \cdots, \mathscr{F}_K$ we have essentially decoupled the system. If we now expand term by term the result follows from the same argument as was used in the proof of Theorem 1. □

Proceeding analogously to the one-equation case, we can now prove uniqueness and continuous dependence on the data for problem P2. In particular, let

$$\mathscr{M} = \left\{ (u_1, \cdots, u_K) : \sum_{k=1}^{K} \|u_k\|^2_{L^2(\Omega)} \leqq M^2 \right\}.$$

We now wish to replace $(\phi_1, \cdots, \phi_K)$ by the solution $(\tilde{u}_1, \cdots, \tilde{u}_K)$ to a problem close to P2 in the sense that

$$\varepsilon^2 = \sum_{k=1}^{K} \left\{ \|g_k - \tilde{u}_k\|^2_{H^1(\Sigma)} + \left\| h_k - \frac{\partial \tilde{u}_k}{\partial n} \right\|^2_{L^2(\Sigma)} + \|L_k \tilde{u}_k - \mathscr{F}_k(\tilde{u}_1, \cdots, \tilde{u}_K)\|^2_{L^2(\Omega)} \right\}$$

is small. We then have the following result:

COROLLARY. *The solution to problem* P2, *if it exists, is unique and depends continuously on the data within the class* $\mathscr{M}$. *In particular, if* $(u_1, \cdots, u_K), (\tilde{u}_1, \cdots, \tilde{u}_K) \in \mathscr{M}$ *are as described above and there exists a function f and functions* $\sigma_1, \cdots, \sigma_K$ *satisfying*

the hypotheses of Theorem 2, then

$$\iint_{D_\alpha} (u_k - \tilde{u}_k)(u_k - \tilde{u}_k)\, dm\, dt \leqq c\left\{M^{2\alpha}\varepsilon^{2(1-\alpha)} + \frac{2M^2}{\log (M^2/\varepsilon^2)}\right\} \frac{1}{\log (M^2/\varepsilon^2)}$$

for $0 < \alpha \leqq 1$.

4. Navier–Stokes equations. In this section we wish to consider the noncharacteristic Cauchy problem for the time-dependent Navier–Stokes equations

$$u_{i,t} + u_j u_{i,j} = -\frac{p_{,i}}{\rho} + F_i + \frac{\nu}{\rho}\Delta u_i, \qquad i = 1, 2, 3, \tag{4.1}$$

$$u_{i,i} = 0. \tag{4.2}$$

More precisely, we wish to study the stability properties of (4.1) and (4.2) with data given only on Σ. However, the question remains as to what data is required to pose the problem. The answer comes from "Cauchy Kovalevsky" type arguments; i.e., we wish to pose data which would allow us to formally resolve the power series of the u_i if everything were analytic.

It should be recalled that for the initial boundary value problem, because of special orthogonality conditions one does not specify any boundary condition for the pressure (cf. [3]). However, since this orthogonality condition cannot be applied to the noncharacteristic Cauchy problem, one would not expect specifying u_i and $\partial u_i/\partial n$ for $i = 1$, 2, 3 to be sufficient to resolve the problem. This is indeed the case, for we find that we cannot resolve

$$\nu u_{s,nn} - p_{,s}, \tag{4.3}$$

where s represents a tangential direction on Σ and n the normal to Σ. (Note also that the divergence condition imposes a compatibility relation on our boundary data.) We will circumvent (4.3) by assuming that the pressure gradient is available on Σ. Thus, we pose the problem P3,

$$\text{P3} \quad \begin{cases} \dfrac{\partial u_i}{\partial t} + u_j u_{i,j} = \mu \Delta u_i - \dfrac{p_{,i}}{\rho} + F_i & \text{in } \Omega, \quad i = 1, 2, 3, \\ u_{i,i} = 0 & \text{in } \Omega, \\ u_i = g_i & \text{on } \Sigma, \quad i = 1, 2, 3, \\ \dfrac{\partial u_i}{\partial n} = h_i & \text{on } \Sigma, \quad i = 1, 2, 3, \\ p_{,i} = \psi_i & \text{on } \Sigma, \quad i = 1, 2, 3, \end{cases}$$

where $\mu = \nu/\rho$ and ρ are constants as defined before. In posing the data for P3 we have simply used boundary data which are convenient for the analysis; other choices are possible. (In fact, the normal derivative of both the pressure and the normal velocity component can be expressed in terms of the other available data; but this does not affect our analysis.) Two other possibilities merit consideration. We can resolve all of the data required from the velocity field and the first and second normal derivatives of the tangential velocity components. From a more physical perspective, the required data can also be determined from the surface tractions

$$(u_{i,j} + u_{j,i})n_j + pn_i \quad \text{on } \Sigma.$$

In addition to the fully nonlinear equations, we also wish to examine the equations in their linearized form. Aside from interest for its own sake, we will treat the nonlinear equations simply by restricting our consideration to a class of functions in which the nonlinearity can be treated as a Lipschitz function. We will then discuss the effect of the nonlinearity on our stability results. Rather than linearize about a constant velocity field, we will use any divergence free, C^2 velocity field. Thus, we are led to the linearized version of P3; viz.,

$$\text{P4}\quad \begin{cases} \dfrac{\partial \bar{u}_i}{\partial t} + U_j \bar{u}_{i,j} + \bar{u}_j U_{i,j} = \mu \Delta \bar{u}_i - \dfrac{p_{,i}}{\rho} + \bar{F}_i & \text{in } \Omega, \\ \bar{u}_{i,i} = 0 & \text{in } \Omega, \\ \bar{u}_i = \bar{g}_i & \text{on } \Sigma, \\ \dfrac{\partial \bar{u}_i}{\partial n} = \bar{h}_i & \text{on } \Sigma, \\ p_{,i} = \psi_i & \text{on } \Sigma, \end{cases}$$

where $\bar{u}_i$ represents u_i linearized about U_i and the boundary data and forcing terms have been suitably modified.

As was the case in §§ 2–3, to obtain meaningful results we need to use a coordinate system other than the usual Cartesian system. However, in this case we will restrict our consideration to a triply orthogonal coordinate system defined by the variables ξ, ω_1 and ω_2 with metrical coefficients γ_ξ, γ_1 and γ_2. We will assume that γ_ξ, γ_1 and γ_2 are bounded, bounded away from zero and possess bounded derivatives.

If we let w_ξ, w_1 and w_2 represent the velocity components in our new coordinate direction, the equations of motion become

$$-w_{k,t} + \mu\left(\frac{1}{\gamma_\xi^2} w_{k,\xi\xi} + \frac{1}{\gamma_1^2} w_{k,11} + \frac{1}{\gamma_2^2} w_{k,22}\right) \equiv G_k - F_k - \frac{1}{\rho\gamma} p_{,k}, \qquad k = \xi, \omega_1, \omega_2, \tag{4.4}$$

where F_ξ, F_{ω_1} and F_{ω_2} are the components of the external force. G_ξ, G_1 and G_2 contain the remaining terms of the equation. Nash and Patel [5] contains the exact form of these functions. For our purposes the exact form of these equations is not important; their behavior is sufficiently characterized by

$$\begin{aligned} &|\mathcal{G}_k(w_\xi^1, w_1^1, w_2^1)| - |\mathcal{G}_k(w_\xi^2, w_1^2, w_2^2)| \\ &\quad \equiv |G_k(t, \xi, \omega_1, \omega_2, w_\xi^1, w_1^1, w_2^1, w_{\xi,\xi}^1, w_{\xi,1}^1, \cdots, w_{2,2}^1) \\ &\qquad\qquad - G_k(t, \xi, \omega_1\omega_2, w_\xi^2, w_1^2, w_2^2, \cdots, w_{2,2}^2)| \\ &\quad \leqq c(\max(|w_\xi^1|, |w_1^1|, |w_2^1|) + \max(|w_\xi^2|, |w_1^2|, |w_2^2|)) \\ &\qquad \times (|w_\xi^1 - w_\xi^2| + |w_1^1 - w_1^2| + \cdots + |w_{1,2}^1 - w_{1,2}^2| + |w_{2,2}^1 - w_{2,2}^2|) \quad \text{for } k = \xi, \omega_1, \omega_2. \end{aligned} \tag{4.5}$$

The linearized equations for $\bar{w}_\xi$, $\bar{w}_1$ and $\bar{w}_2$ also takes the form of (4.4) with right-hand side functions $\bar{F}_\xi, \bar{F}_{\omega_1}, \bar{F}_{\omega_2}, \bar{G}_\xi, \bar{G}_1$ and $\bar{G}_2$. The difference between the two cases is that $\bar{G}_\xi$, $\bar{G}_1$ and $\bar{G}_2$ are actually Lipschitz functions (the Lipschitz constants depending, of course, on the velocity field U).

Having established the framework of the problem, we now prove an a priori type inequality which will be used to treat the pressure term in the equation.

Using the previously defined triply orthogonal coordinate system in $\mathcal{R}^3$, we define for f, a function only of ξ, and D a domain in $\mathcal{R}^3$,

$$d_\alpha = \{(\xi, \omega_1, \omega_2): f(\xi) \leqq \alpha\} \cap D, \quad s_\alpha = \partial d_\alpha \cap D, \quad \Gamma_\alpha = \partial d_\alpha \cap \partial D.$$

We then have the lemma:

LEMMA. *Suppose there exists a function $f(\xi)$ and a $\sigma(\xi, \omega_1, \omega_2) > 0$ satisfying*

1) $d_\alpha \subset d_\beta, \quad 0 \leqq \alpha < \beta \leqq A \quad$ *in* d_A,

2) $f_{,\xi} \geqq c > 0, \quad f_{,\xi\xi} \leqq -c < 0 \quad$ *in* d_A,

3) $\left(\left(\frac{f^2{}_{,\xi}\sigma}{\gamma_\xi^2}\right)\right)_{,\xi} \leqq -c < 0 \quad$ *in* d_A,

4) $\left(\frac{\sigma}{\gamma_i^2}\right)_{,\xi} \geqq 0 \quad$ *for* $i = 1, 2 \quad$ *in* d_A.

Then, for any function p,

$$\begin{aligned}
&\iint_{d_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dm \\
&\quad \leqq c \iint_{d_\alpha} \frac{e^{-2\lambda f}}{\lambda} (\Delta p)^2\, dm + c\, e^{-2\lambda\alpha} \int_{s_\alpha} (p^2{}_{,1} + p^2{}_{,2})\, ds \\
&\qquad + c \int_{\Gamma_\alpha} (\lambda^2 p^2 + p^2{}_{,\xi} + p^2{}_{,1} + p^2{}_{,2})\, ds,
\end{aligned}$$

for all λ sufficiently large where $q = e^{\lambda f} p$.

Proof. The lemma is, essentially, an abortive proof of stability for the Cauchy problem for the Laplace equation using the technique of the previous sections. In the $(\xi, \omega_1, \omega_2)$-coordinate system

$$\Delta \equiv \frac{1}{\gamma_\xi^2}\frac{\partial^2}{\partial\xi^2} + \frac{1}{\gamma_1^2}\frac{\partial^2}{\partial\omega_1^2} + \frac{1}{\gamma_2^2}\frac{\partial^2}{\partial\omega_2^2} + n_\xi \frac{\partial}{\partial\xi} + n_1 \frac{\partial}{\partial\omega_1} + n_2 \frac{\partial}{\partial\omega_2};$$

so, letting $q = e^{\lambda f} p$, we find that q satisfies an equation of the form

$$\begin{aligned}
&\tilde{\gamma}_\xi[(\lambda^2 f^2{}_{,\xi} + O(\lambda))q + 2\lambda f_{,\xi} q_{,\xi} + q_{,\xi\xi}] + \tilde{\gamma}_1 q_{,11} + \tilde{\gamma}_2 q_{,22} \\
&\quad = \sigma\, e^{-\lambda f}(\Delta p - \lambda f_{,\xi} n_\xi q - n_\xi q_{,\xi} - n_1 q_{,1} - n_2 q_{,2})
\end{aligned} \tag{4.6}$$

where

$$\tilde{\gamma}_\xi = \frac{\sigma}{\gamma_\xi^2}, \quad \tilde{\gamma}_1 = \frac{\sigma}{\gamma_1^2}, \quad \tilde{\gamma}_2 = \frac{\sigma}{\gamma_2^2}.$$

We now multiply (4.6) by $(q_{,\xi} + \frac{1}{2} f^{-1}{}_{,\xi} f_{,\xi\xi} q)$ and integrate over d_α to obtain

$$\begin{aligned}
\iint_{d_\alpha} \Big\{ &\lambda^2 (\tilde{\gamma}_\xi f^2{}_{,\xi} + O(\lambda^{-1})) q q_{,\xi} + 2\lambda \tilde{\gamma}_\xi f_{,\xi} q^2{}_{,\xi} + \tilde{\gamma}_\xi q_{,\xi\xi} q_{,\xi} + \tilde{\gamma}_1 q_{,11} q_{,\xi} \\
&+ \tilde{\gamma}_2 q_{,22} q_{,\xi} + \frac{\lambda^2}{2} \tilde{\gamma}_\xi (f_{,\xi} f_{,\xi\xi} + O(\lambda^{-1})) q^2 + \lambda \tilde{\gamma}_\xi f_{,\xi\xi} q q_{,\xi} \\
&+ \frac{\tilde{\gamma}_\xi}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,\xi\xi} q + \frac{\tilde{\gamma}_1}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,11} q + \frac{\tilde{\gamma}_2}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,22} q \Big\}\, dm
\end{aligned}$$

$$
\begin{aligned}
(4.7)\qquad &= \iint_{d_\alpha} \left\{ \sigma e^{-\lambda f} \Delta p \left(q_{,\xi} + \frac{1}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q \right) - \sigma \left(q_{,\xi} + \frac{1}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q \right) \right. \\
&\qquad \left. \times (\lambda f_{,\xi} n_\xi q + n_\xi q_{,\xi} + n_{,} q_{,1} + n_2 q_{,2}) \right\} dm \\
&\leqq c \iint_{d_\alpha} \sigma^2 e^{-2\lambda f} \frac{(\Delta p)^2}{\lambda} dm + \iint_{d_\alpha} (\lambda \tilde{\gamma}_\xi f_{,\xi} q^2{}_{,\xi} + \lambda q^2)\, dm \\
&\quad + c \iint_{d_\alpha} \frac{1}{\lambda} (q^2{}_{,1} + q^2{}_{,2})\, dm.
\end{aligned}
$$

Integrating by parts and regrouping terms we find

$$
\begin{aligned}
&- \iint_{d_\alpha} \lambda^2 \frac{\gamma_\xi}{2} [f_{,\xi} f_{,\xi\xi} + O(\lambda^{-1})] q^2\, dm + \iint_{d_\alpha} \lambda \tilde{\gamma}_\xi [f_{,\xi} + O(\lambda^{-1})] q^2{}_{,\xi}\, dm \\
&- \iint_{d_\alpha} \frac{\tilde{\gamma}_1}{2} \left[\frac{f_{,\xi\xi}}{f_{,\xi}} + O(\lambda^{-1}) \right] q^2{}_{,1}\, dm - \iint_{d_\alpha} \frac{\tilde{\gamma}_2}{2} \left[\frac{f_{,\xi\xi}}{f_{,\xi}} + O(\lambda^{-1}) \right] q^2{}_{,2}\, dm \\
&+ \iint_{d_\alpha} (\tilde{\gamma}_{1,\xi} q^2{}_{,1} + \tilde{\gamma}_{2,\xi} q^2{}_{,2})\, dm + \lambda^2 \oint_{\partial d_\alpha} \left[\frac{f^2{}_{,\xi}}{2} + O(\lambda^{-1}) \right] q^2 n_\xi\, dS + \oint_{\partial d_\alpha} \tilde{\gamma}_\xi q^2{}_{,\xi} n_\xi\, dS \\
(4.8)\qquad &+ \oint_{\partial d_\alpha} \frac{\tilde{\gamma}_1}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,1} q n_1\, dS + \oint_{\partial d_\alpha} \frac{\tilde{\gamma}}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,2} q n_2\, dS + \int_{\partial d_\alpha} \frac{\tilde{\gamma}_\xi}{2} \frac{f_{,\xi\xi}}{f_{,\xi}} q_{,\xi} q n_\xi\, dS \\
&+ \oint_{\partial d_\alpha} (\tilde{\gamma}_1 q_{,1} n_1 + \tilde{\gamma}_2 q_{,2} n_2) q_{,\xi}\, dS - \oint_{\partial d_\alpha} (\tilde{\gamma}_1 q^2{}_{,1} + \tilde{\gamma}_2 q^2{}_{,2}) n_\xi\, dS \\
&\qquad \leqq c \iint_{d_\alpha} \sigma^2 \frac{e^{-2\lambda f}}{\lambda} (\Delta p)^2\, dm.
\end{aligned}
$$

Applying the assumptions for f and σ we see that the volume terms in (4.8) are greater than

$$
c \left(\iint_{d_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dm \right)
$$

for λ sufficiently large. Similarly, the sum of the boundary terms can be bounded above by

$$
-c \int_{s_\alpha} (q^2{}_{,1} + q^2{}_{,2})\, dS - c \left[\int_{\Gamma_\alpha} \lambda^2 q^2\, dS + \int_{\Gamma_\alpha} (q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dS \right]
$$

for large λ. Combining these results we find that for λ large

$$
\begin{aligned}
&\iint_{d_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dm \\
&\quad \leqq c \iint_{d_\alpha} e^{-2\lambda f} \frac{(\Delta p)^2}{\lambda} dS + c \int_{s_\alpha} (q^2{}_{,1} + q^2{}_{,2})\, dS + c \int_{\Gamma_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dS.
\end{aligned}
$$

Substituting back for p yields

$$\iint_{d_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dm$$
$$\leqq c \iint_{d_\alpha} \frac{e^{-2\lambda f}}{\lambda} (\Delta p)^2\, dS + c\, e^{-2\lambda\alpha} \int_{s_\alpha} (p^2{}_{,1} + p^2{}_{,2})\, dS$$
$$+ c \int_{\Gamma_\alpha} (\lambda^2 p^2 + p^2{}_{,\xi} + p^2{}_{,1} + p^2{}_{,2})\, dS$$

for sufficiently large λ. □

We are now ready to prove the analogue of Theorem 2 for the fluid flow equations. We will treat the linearized version of the equation first, then indicate the modification required to treat the full nonlinear equations.

THEOREM 3. *Suppose* $(\bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{p})$ *satisfies* P4. Let $(\phi_1, \phi_2, \phi_3, p_\phi)$ *be functions such that each* ϕ_i *and* p_ϕ *possess bounded second order derivatives in* Ω *and bounded first order derivatives in* $\Sigma \cup \Omega$ *and such that*

$$\left(\frac{\partial \phi_i}{\partial t} + U_j \phi_{i,j} + \phi_j U_{i,j} - \mu \Delta \phi_i - \frac{p_{\phi,i}}{\rho}\right)_{,i} \equiv F_{\phi i,i}$$

is bounded in Ω. *Suppose further that there exists a function* $f(t, \xi)$ *and a* $\sigma(t, \xi, \omega_1, \omega_2)$ *satisfying*

1) $D_\alpha \subset D_\beta, \qquad \Sigma_\alpha \subset \Sigma, \qquad 0 < \alpha < \beta \leqq 1,$

2) $f_{,\xi} \geqq c > 0, \qquad f_{,\xi\xi} \leqq -c < 0 \quad in\ \Omega,$

3) $\left(\frac{\sigma^2}{\gamma_\xi^2} f^3{}_{,\xi}\right)_{,\xi} \leqq -c < 0, \qquad \left(\frac{\sigma}{\gamma_\xi^2} f_{,\xi}\right)_{,\xi} \leqq -c < 0 \quad in\ \Omega,$

4) $\left(\left(\frac{\sigma^2}{\gamma_\xi^2 \gamma_i^2}\right)\right)_{,\xi} \geqq 0, \qquad i = i, 2 \quad in\ \Omega,$

5) $\left(\left(\frac{\sigma^2}{\gamma_i^2}\right)\right)_{,i} = 0, \qquad i = 1, 2 \quad in\ \Omega,$

6) $\left(\left(\frac{\sigma^2}{\gamma_\xi^2}\right)\right)_{,\xi} \geqq 0 \quad in\ \Omega,$

for the triply orthogonal coordinate system defined earlier in this section. Then, for all α, $0 < \alpha \leqq 1$,

$$\lambda^3 \iint_{D_\alpha} (\bar{u}_i - \phi_i)(\bar{u}_i - \phi_i)\, dm\, dt$$
$$\leqq c \iint_{D_\alpha} [\lambda (\bar{u}_{j,i} - \phi_{j,i})(\bar{u}_{j,i} - \phi_{j,i}) + (\bar{u}_{i,t} - \phi_{i,t})(\bar{u}_{i,t} - \phi_{i,t})$$
$$+ (\bar{p}_{,i} - p_{\phi,i})(\bar{p}_{,i} - p_{\phi,i})]\, dm\, dt$$
$$+ c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} [\lambda^2 (\bar{u}_i - \phi_i)(\bar{u}_i - \phi_i) + (\bar{u}_{j,i} - \phi_{j,i})(\bar{u}_{j,i} - \phi_{j,i})$$
$$+ (\bar{u}_{i,t} - \phi_{i,t})(\bar{u}_{i,t} - \phi_{i,t})]\, dS$$
$$+ c\, e^{2\lambda\alpha} \int_{\Sigma} (\bar{p}_{,i} - p_{\phi,i})(\bar{p}_{,i} - p_{\phi,i})\, dS$$

for all λ *sufficiently large.*

Proof. We will prove the estimate using the transformed form of the equation given by (4.4). To denote the transform of ϕ_1, ϕ_2 and ϕ_3 we will use ϕ_ξ, ϕ_{ω_1} and ϕ_{ω_2}. Furthermore, F_{ϕ_ξ}, $F_{\phi_{\omega_1}}$ and $F_{\phi_{\omega_2}}$ will denote the transformed equivalents of the F_{ϕ_i}.

We now let $v_k = \bar{e}^{\lambda f}(\bar{w}_k - \bar{\phi}_k)$ for $k = \xi, \omega_1, \omega_2$ and $q = \bar{e}^{\lambda f}(\bar{p} - p_\phi)$. Substituting into (4.4) we obtain

$$
\begin{aligned}
&-\sigma v_{k,t} + \mu[\lambda^2 \tilde{\gamma}_\xi (f^2{}_{,\xi} + O(\lambda^{-1}))v_k + 2\lambda \tilde{\gamma}_\xi f_{,\xi} v_{k,\xi} + \tilde{\gamma}_\xi v_{k,\xi\xi} + \tilde{\gamma}_1 v_{k,11} + \tilde{\gamma}_2 v_{k,22}] \\
&\quad = e^{-\lambda f}\sigma[\bar{\mathcal{G}}_k(w_\xi, w_{\omega_1}, w_{\omega_2}) - \bar{\mathcal{G}}_k(\phi_\xi, \phi_{\omega_1}, \phi_{\omega_2}) - (\bar{F}_k - F_{\phi k_i})] \\
&\quad\quad + \sigma \frac{\delta_{k\xi}\lambda f_{,\xi} q - q_{,k}}{\rho \gamma_k}, \qquad k = \xi, \omega_1 \omega_2
\end{aligned} \tag{4.9}
$$

where $\tilde{\gamma}_\xi = \sigma/\gamma_\xi^2$, $\tilde{\gamma}_1 = \sigma/\sigma_1^2$ and $\tilde{\gamma}_2 = \sigma/\gamma_2^2$. Also, $\delta_{k\xi}$ denotes the Kronecker δ which is 1 if $k = \xi$ and zero otherwise. We now group terms as odd or even and define the operators

$$
L_{ke} v_k = \mu[\lambda^2 \tilde{\gamma}_\xi (f^2{}_{,\xi} + O(\lambda^{-1}))v_k + \tilde{\gamma}_\xi v_{k,\xi\xi} + \tilde{\gamma}_1 v_{k,11} + \tilde{\gamma}_2 v_{k,22}, \qquad k = \xi, \omega_1, \omega_2
$$

and

$$
L_{ko} v_k = -\sigma v_{k,t} + 2\lambda \tilde{\gamma}_\xi f_{,\xi} v_{k,\xi}, \qquad k = \xi, \omega_1, \omega_2.
$$

Next, we form

$$
\begin{aligned}
&\sum_{k=\xi,\omega_1,\omega_2} (L_{ke} v_k \times L_{ke} v_k) + \sum_{k=\xi,\omega_1,\omega_2} \tfrac{1}{2}(L_{ko} v_k)^2 - \lambda^{1/2} \sum_{k=\xi,\omega_1,\omega_2} (L_{ke} v_k + L_{ko} v_k) v_k \\
&\quad \leqq c \sum_{k=\xi,\omega_1,\omega_2} \Big\{ c\, e^{-2\lambda f}\sigma^2 (\bar{\mathcal{G}}_k(w_\xi, w_{\omega_1}, w_{\omega_2}) - \bar{\mathcal{G}}_k(\phi_\xi, \phi_{\omega_1}, \phi_{\omega_2})) + \sigma^2 \Big(\frac{\delta_{k\xi}\lambda f_{,\xi} q - q_{,k}}{\rho \gamma_k} \Big)^2 \\
&\quad\quad + c\, e^{-2\lambda f}\sigma^2 |\bar{F}_k - F_{\phi_k}|^2 \\
&\quad\quad - \lambda^{1/2} \Big[e^{-\lambda f}\sigma |\bar{\mathcal{G}}_k(w_\xi, w_{\omega_1}, w_{\omega_2}) - \bar{\mathcal{G}}_k(\phi_\xi, \phi_{\omega_1}, \phi_{\omega_2})| \\
&\quad\quad\quad + \sigma\, e^{-\lambda f} |\bar{F}_k - F_{\phi_k}| - \sigma \delta_{k\xi} \frac{\lambda f_{,\xi}}{\rho \gamma_\xi} q - \sigma \frac{q_{,k}}{\gamma_k} \Big] v_k \Big\}.
\end{aligned} \tag{4.10}
$$

Note that the left-hand side of (4.10) is exactly the same as the left-hand side of the inequality used to prove Theorem 2. In fact, after integrating over D_α we find that the only term on the right-hand side which causes problems is

$$
c \iint_{D_\alpha} (\lambda^2 q^2 + q^2{}_{,\xi} + q^2{}_{,1} + q^2{}_{,2})\, dm\, dt. \tag{4.11}
$$

To treat this term observe that, for fixed t, $f(\xi, t)$ by the above assumptions satisfies the hypotheses of the lemma. Computing $\Delta(p - p_\phi)$ by taking the divergence of the system treated as a single vector equation, we note that for λ sufficiently large we may bound (4.11) by

$$
\begin{aligned}
&c \iint_{D_\alpha} \lambda (v_\xi^2 + v_{\omega_1}^2 + v_{\omega_2}^2) + \lambda^{-1}(v_{\xi,\xi}^2 + v_{\omega_1,\xi}^2 + \cdots + v_{\omega_2,\omega_2}^2)\, dm\, dt \\
&\quad + c\, e^{-2\lambda\alpha} \int_{S_\alpha} (p^2{}_{,\omega_1} + p^2{}_{,\omega_2} - p_{\phi,\omega_1}^2 - p_{\phi,\omega_1}^2) |\nabla f|^{-1}\, dS \\
&\quad + c \int_{\Sigma_\alpha} \lambda^2 (p - p_\phi)^2 + (p_{,\xi} - p_{\phi,\xi})^2 + (p_{,\omega_1} - p_{\phi,\omega_1})^2 + (\bar{p}_{,\omega_2} - p_{\phi,\omega_2})^2\, dS \\
&\quad + c\lambda^{-1} \iint_{D_\alpha} \sum_{k=\xi,\omega_1,\omega_2} |\bar{F}_{k,k} - F_{\phi_k,k}|^2\, dm\, dt.
\end{aligned}
$$

Proceeding as in the proof of Theorem 2, and returning to our original variables, we obtain

$$\lambda^3 \iint_{D_\alpha} (\bar{u}_i - \phi_i)(u_i - \phi_i)\, dm\, dt$$

$$\leqq c \iint_{D_\alpha} \lambda(\bar{u}_{j,i} - \phi_{j,i})(\bar{u}_{j,i} - \phi_{j,i}) + (\bar{u}_{i,t} - \phi_{i,t})(\bar{u}_{i,t} - \phi_{i,t}) + (p_{,i} - p_{\phi,i})(p_{,i} - p_{\phi,i})\, dm\, dt$$

$$+ c\lambda^{-1} e^{2\lambda\alpha} \int_{\Sigma_\alpha} \lambda^3(\bar{u}_i - \phi_i)(\bar{u}_i - \phi_i) + \lambda(\bar{u}_{j,i} - \phi_{j,i})(\bar{u}_{j,i} - \phi_{j,i})$$

$$+ (\bar{u}_{i,t} - \phi_{i,t})(\bar{u}_{i,t} - \phi_{i,t})\, dS$$

$$+ c\lambda^{-1} e^{2\lambda\alpha} \int_{\Sigma} [\lambda^2(\bar{p} - p_\phi)^2 + (\bar{p} - p_\phi)_{,i}(\bar{p} - p_\phi)_{,i}]\, dS$$

$$+ c\lambda^{-2} e^{2\lambda\alpha} \iint_{D_\alpha} \sum_{k=1}^{3} (\bar{F}_{k,k} - F_{\phi_k,k})^2\, dm\, dt.$$

Finally, since p is only determined up to a constant at each time level, by appropriately selecting the constants we can bound $\int_\Sigma \lambda^2(\bar{p} - p_\phi)^2\, dS$ by $c \int_\Sigma (\bar{p} - p_\phi)_{,i}(\bar{p} - p_\phi)_{,i}\, dS$ which yields the desired result. □

We can now apply the estimate to prove uniqueness and continuous dependence on the data for problem P4. Let[3]

$$\bar{\mathcal{M}} = \left\{(\bar{u}_1, \bar{u}_2, \bar{u}_3): \iint_\Omega (\bar{u}_{i,t}\bar{u}_{i,t} + \bar{u}_{j,i}\bar{u}_{j,i})\, dm\, dt \leqq M^2\right\}.$$

We establish the following corollary to Theorem 3:

COROLLARY. *Given a triply orthogonal coordinate system ξ, ω_1 and ω_2, suppose there exist functions f and σ satisfying conditions* 1)–6) *of Theorem* 3 *for* $0 < \alpha \leqq 1$. *Then, if there is a flow solution to* P4 *it is unique and depends continuously on the data in D_1 within the class $\bar{\mathcal{M}}$. In particular, if $(\tilde{\bar{u}}_1, \tilde{\bar{u}}_2, \tilde{\bar{u}}_3) \in \bar{\mathcal{M}}$ with associated pressure $\tilde{\bar{p}}$ solves a problem close to* P4 *in the sense that*

$$\varepsilon^2 = \sum_{i=1}^{3} \left(\|\bar{g}_i - \tilde{\bar{u}}_i\|^2_{H^1(\Sigma)} + \left\| \bar{h}_i - \frac{\partial \tilde{\bar{u}}_i}{\partial n} \right\|^2_{L^2(\Sigma)} + \|\bar{p}_{,i} - \tilde{\bar{p}}_{,i}\|^2_{L^2(\Sigma)} + \|\bar{F}_i - \bar{F}_{\tilde{u}_i}\|^2_{L^2(\Omega)} \right)$$

is small, then

$$\iint_{D_\alpha} (\bar{u}_i - \tilde{\bar{u}}_i)(\bar{u}_i - \tilde{\bar{u}}_i)\, dm\, dt \leqq c\left(M^{2\alpha} \varepsilon^{2(1-\alpha)} + \frac{2M^2}{\log(M^2/\varepsilon^2)} \right) \frac{1}{\log(M^2/\varepsilon^2)}.$$

Proof. Substituting $\tilde{\bar{u}}_1$, $\tilde{\bar{u}}_2$, $\tilde{\bar{u}}_3$ and $\tilde{\bar{p}}$ for ϕ_1, ϕ_2, ϕ_3 and p_ϕ in Theorem 3, we see that we need only bound $\iint_{D_\alpha} (\bar{p}_{,i} - \tilde{\bar{p}}_{,i})(\bar{p} - \tilde{\bar{p}}_{,i})\, dm\, dt$ in terms of M and data terms. Let ω be a C^∞ cutoff function which is one on D_α and identically zero outside a subdomain D of Ω, satisfying $(\bar{D}_\alpha \cap \Omega) \subseteq D$ and $\partial D \cap \partial\Omega \subseteq \Sigma$. Then, denoting $\hat{p} = \bar{p} - \tilde{\bar{p}}$, $\hat{u} = \bar{u}_i - \tilde{\bar{u}}_i$,

[3] Note that we have chosen to eliminate information about the pressure from our stabilization class. If pressure information is available, less restrictive conditions on the velocity field can be used.

and $\hat{F}_i = \bar{F}_i - \bar{F}_{\bar{u}_i}$ yields

$$\iint_{D_\alpha} \hat{p}_{,i}\hat{p}_{,i}\,dm\,dt \leqq \iint_D \omega^2 \hat{p}_{,i}\hat{p}_{,i}\,dm\,dt$$

$$= \iint_D \omega^2 \hat{p}_{,i}(\rho\mu\Delta\hat{u}_i - \rho U_j \hat{u}_{i,j} - \rho\hat{u}_j\bar{U}_{i,j} - \rho\hat{u}_{i,t} - \hat{F}_i)\,dm\,dt$$

$$\leqq \iint_D \frac{\omega^2}{2}\hat{p}_{,i}\hat{p}_{,i}\,dm\,dt + c\iint_D \hat{u}_{j,i}\hat{u}_{j,i} + \hat{u}_i\hat{u}_i + \hat{u}_{i,t}\hat{u}_{i,t}\,dm\,dt$$

$$+ \iint_D \omega^2\hat{p}_{,i}\rho\mu\Delta\hat{u}_i\,dm\,dt + c\iint_D \hat{F}_i\hat{F}_i\,dm\,dt.$$

Now observe that

$$\hat{p}_{,i}\Delta\hat{u}_i = \hat{p}_{,i}(\hat{u}_{i,j} - \hat{u}_{j,i})_{,j}.$$

Thus,

$$\iint_D \omega^2\hat{p}_{,i}\rho\mu\Delta\hat{u}_i\,dm\,dt$$

$$= \iint_D \omega^2\hat{p}_{,i}\rho\mu(\hat{u}_{i,j} - \hat{u}_{j,i})_{,j}\,dm\,dt$$

$$= \oint_{\partial D} \omega^2\hat{p}_{,i}\rho\mu(\hat{u}_{i,j} - \hat{u}_{j,i})n_j\,dS - \iint_D \omega\omega_{,j}\hat{p}_{,i}\rho\mu(\hat{u}_{i,j} - \hat{u}_{j,i})\,dm\,dt$$

$$\leqq \int_\Sigma \hat{p}_{,i}\hat{p}_{,i}\,dS + \int_\Sigma \hat{u}_{i,j}\hat{u}_{j,i}\,dS + \iint_D \frac{\omega^2}{4}\hat{p}_{,i}\hat{p}_{,i}\,dm\,dt + c\iint_D \hat{u}_{j,i}\hat{u}_{j,i}\,dm\,dt.$$

Using the resultant estimate for $\int_{D_\alpha} \hat{p}_{,i}\hat{p}_{,i}\,dm\,dt$ we find that

$$\lambda^3 \iint_{D_\alpha} \hat{u}_i\hat{u}_i\,dm\,dt$$

$$\leqq c\iint_D (\lambda\hat{u}_{i,j}\hat{u}_{i,j} + \hat{u}_{i,t}\hat{u}_{i,t})\,dm\,dt + c\,e^{2\lambda\alpha}\int_{\Sigma_\alpha} (\lambda^2\hat{u}_i\hat{u}_i + \hat{u}_{i,j}\hat{u}_{i,j} + \hat{u}_{i,t}\hat{u}_{i,t})\,dS \tag{4.12}$$

$$+ c\,e^{2\lambda\alpha}\int_\Sigma \hat{p}_{,i}\hat{p}_{,i}\,dS + c\iint_D \hat{F}_i\hat{F}_i\,dm\,dt.$$

The estimate of the theorem now becomes

$$\iint_{D_\alpha} \hat{u}_i\hat{u}_i\,dm\,dt \leqq c\lambda^{-2}M^2 + \lambda^{-1}\varepsilon^2\,e^{2\lambda\alpha}. \tag{4.13}$$

Uniqueness follows by letting $\lambda \to \infty$ since $\varepsilon = 0$. The continuous dependence result follows by letting $\lambda = \frac{1}{2}\log(M^2/\varepsilon^2)$ in (4.13)

We are now ready to extend the results for problem P4 to the full nonlinear case. We will handle the nonlinearity by initially restricting consideration to a class of functions in which we can deduce a linear bound on the nonlinear functions. Thus, we let $\mathcal{N} = \{(u_1, u_2, u_3)\colon |u_{i,j}|^2 \leqq m$ at each point of $\Omega\}$. Within this class all of our functions are Lipschitz. The analogue of Theorem 3 can now be shown with no alteration in the proof.

THEOREM 4. *Suppose* (u_1, u_2, u_3, p) *where* $(u_1, u_2, u_3) \in \mathcal{N}$ *satisfies* P3. *Let* $(\phi_1, \phi_2, \phi_3, p_\phi)$ *be functions such that each* ϕ_i *and* p_ϕ *possess bounded second order*

derivatives in Ω *and bounded first order derivatives in* $\Sigma \cup \Omega$ *and such that*

$$\left(\phi_{i,t} + \phi_j \phi_{i,j} - \mu \Delta \phi_i + \frac{p_{\phi,i}}{\rho}\right)_{,i} \equiv (F_{\phi i})_{,i}$$

is bounded in Ω. *Suppose further that there exist a function* $f(t, \xi)$ *and a* $\sigma(t, \xi, \omega_1, \omega_2)$ *satisfying* 1)–6) *of Theorem* 3 *for the triply orthogonal coordinate system described earlier. Then, for* $0 < \alpha \leqq 1$,

$$\lambda^3 \int\int_{D_\alpha} (u_i - \phi_i)(u_i - \phi_i)\, dm\, dt$$

$$\leqq c \int\int_{D_\alpha} [\lambda (u_{i,j} - \phi_{i,j})(u_{i,j} - \phi_{i,j}) + (u_{i,t} - \phi_{i,t})(u_{i,t} - \phi_{i,t}) + (p_{,i} - p_{\phi,i})(p_{,i} - p_{\phi,i})]\, dm\, dt$$

$$+ c\, e^{2\lambda\alpha} \int_{\Sigma_\alpha} [\lambda^2 (u_i - \phi_i)(u_i - \phi_i) + (u_{i,j} - \phi_{i,j})(u_{i,j} - \phi_{i,j}) + \lambda^{-1}(u_{i,t} - \phi_{i,t})(u_{i,t} - \phi_{i,t})]\, dS$$

$$+ c\, e^{2\lambda\alpha} \int_{\Sigma} \lambda (p_{,i} - p_{\phi,i})(p_{,i} - p_{\phi,i})\, dS + c\, e^{2\lambda\alpha} \int\int_{D_\alpha} (F_i - F_{\phi i})(F_i - F_{\phi i})\, dm\, dt,$$

provided $\lambda \geqq \lambda_0(1 + m)$ *where* λ_0 *is a computable constant which depends on the problem coefficients, the geometry and coordinate system with its associated* f *and* σ.

The following corollary examines the effect of the condition $\lambda \geqq \lambda_0(1 + m)$ on the uniqueness and stability results. Let $\mathcal{M} = \bar{\mathcal{M}} \cap \mathcal{N}$. We then have:

COROLLARY. *Suppose there exist a function* f *and* σ *satisfying the assumption of Theorem* 4 *for all* $\alpha \in (0, 1]$. *Then if there is a flow field solution to problem* P3, *it is unique and depends continuously on the data in* D_1 *within the class* $\mathcal{M}$. *In particular, suppose* $(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3, \tilde{p}) \in \mathcal{M}$ *satisfies a problem which is close to* P3 *in the sense that*

$$\varepsilon^2 = \sum_{i=1}^{3} \left(\|g_i - \tilde{u}_i\|^2_{H^1(\Sigma)} + \left\| h_i - \frac{\partial \tilde{u}_i}{\partial n} \right\|^2_{L^2(\Sigma)} + \|\psi_i - \tilde{p}_{,i}\|^2_{L^2(\Sigma)} + \|F_i - F_{\tilde{u}_i}\|^2_{L^2(\Omega)} \right)$$

satisfies $M^2 \geqq \varepsilon^2\, e^{2\tilde{\lambda}_0(1+m)}$, *where* $\tilde{\lambda}_0$ *is a computable constant. Then*

$$\int\int_{D_\alpha} (u_i - \tilde{u}_i)(u_i - \tilde{u}_i)\, dm\, dt \leqq c \left\{ M^{2\alpha} \varepsilon^{2(1-\alpha)} + \frac{2M^2}{\log (M^2/\varepsilon^2)} \right\} \frac{1}{\log (M^2/\varepsilon^2)}.$$

Proof. The proof is essentially identical to the proof of the corollary to Theorem 3. The condition

$$M^2 \geqq \varepsilon^2\, e^{2\tilde{\lambda}_0(1+m)}$$

results from requiring $\lambda = \frac{1}{2} \log (M^2/\varepsilon^2)$ for continuous dependence while simultaneously requiring λ to be sufficiently large to be used to dominate m. □

The effect of the nonlinearity is thus seen to create a relationship between the data and the stabilization class. In particular, the estimate sets an explicit requirement on the size of the boundary data for two related problems in terms of the stabilization hypothesis. We must in addition restrict our consideration from H^1 functions to $W^{1,\infty}$ functions.

Finally, it should be remarked that conditions 1)–6) of Theorems 3 and 4 are simply the restriction of the conditions of Theorem 1 to the representation of the Laplacian in a triply orthogonal coordinate system. Thus, the example of the next section will satisfy the conditions of the theorems in this section.

5. Examples of surfaces. In this section we will exhibit a coordinate system and functions f and σ satisfying the hypotheses of Theorem 1. For our discussion we will initially restrict ourselves to equations of the form

$$a(t)\Delta u + b(t)u_t = F\left(t, x_1, \cdots, x_n, u, \frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_n}\right). \tag{5.1}$$

The first step in treating multidimensional problems is to introduce a generalized polar coordinate system $\xi, \omega_1, \cdots, \omega_{n-1}$. To introduce our new coordinate system we must first locate our axes in a convenient place. This will be done by moving the origin in the $(t=0)$-plane until the t-axis is near Σ but never intersects $\bar{\Omega}$. Fig. 1 is an illustration of what is intended here.

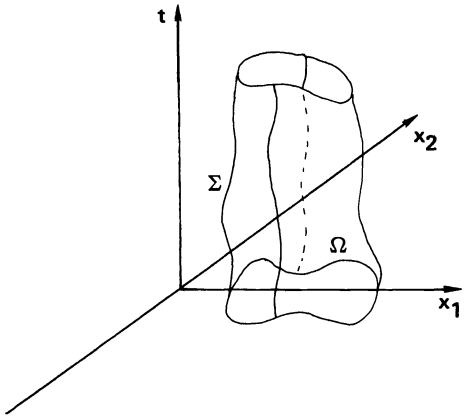


FIG. 1

We next transform to polar coordinates, where (5.1) becomes

$$a(t)\left[u_{,rr} + \frac{1}{r^2}\Delta_\omega u\right] + b(t)u_{,t} = G\left(t, r, \omega_1, \cdots, \omega_{n-1} u, u_{,r} \frac{\partial u}{\partial \omega_1}, \cdots, \frac{\partial u}{\partial \omega_{n-1}}\right)$$

where Δ_ω is the Laplacian on an $(n-1)$-dimensional sphere. We now transform again, letting $\xi = \log r$. We then get

$$a(t)[e^{-2\xi}u_{,\xi\xi} + e^{-2\xi}\Delta_\omega u] + b(t)u_{,t} = H\left(t, \xi, \omega_1, \cdots, \omega_{n-1}, u, u_{,\xi}, \frac{\partial u}{\partial \omega_1}, \cdots, \frac{\partial u}{\partial \omega_{n-1}}\right).$$

If we now require $f_{,\xi} \geqq c > 0$, $\sigma = e^{2\xi} f_{,\xi}^{-1/2}$, the conditions of Theorem 1 reduce to

$$f_{,\xi\xi} \leqq -c < 0,$$

since $\gamma^{ij} = 0$ if $i \neq j$, $\gamma_{ii,i} = 0$ and, $\gamma^{ij} = e^{-2\xi}\bar{\gamma}^{ij}$ where $\bar{\gamma}^{ij}_{,\xi} = 0$ so that $a(t)\sigma^2 e^{-2\xi} f_{,\xi} e^{-2\xi}\bar{\gamma}^{ij}$ is independent of ξ. The conditions on f are now just the one-dimensional condition used in [1]. Therefore, for all $\zeta_1, \zeta_2, \zeta_3 > 0$,

$$f(\xi, t) = \zeta_1(t - t_0)^2 + \zeta_2 \log(\zeta_3 + \xi)$$

satisfies the conditions of Theorem 1.

Finally we note that in (5.1) we may replace $a(t)\Delta u$ by a smoothly varying, time dependent elliptic operator. The technique is simply to change the independent variable diagonalizing the spatial operator as was done in Protter and Weinberger [6] for a constant coefficient operator. Since the problem here is assumed smoothly varying

in time, we can perform their change of variable at each constant time level smoothly in time. Thus, we can explicitly exhibit surfaces for the equation

$$a_{ij}(t)u_{,ij} + b(t)u_{,t} = F(x, t, u, u_{,1}, \cdots, u_{,n}).$$

In conclusion, it should be remarked that we have raised some difficult questions here because of the use of the weighted energy technique: Namely, is there a class of equation for which we can prove a sharper Hölder continuous dependence on the data. The answer to this question remains unknown except in cases with special symmetries which render them essentially one-dimensional.

Acknowledgment. The author would like to thank Professor Larry Payne for many valuable conversations during the course of this research.

REFERENCES

[1] J. BELL, *The noncharacteristic Cauchy problem for a class of equations with time dependence, I. Problems in one space dimension*, this Journal, this issue, pp. 759–777.

[2] F. JOHN, *Partial Differential Equations*, Springer-Verlag, New York, 1971.

[3] O. LADYZHENSKAYA, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1968.

[4] L. M. MILNE-THOMSON, *Theoretical Hydrodynamics*, Macmillan, New York, 1949.

[5] J. F. NASH AND V. C. PATEL, *Three-Dimensional Turbulent Boundary Layers*, SBC Technical Books, Atlanta, 1972.

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

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0001 $01.00/0

A CONTINUUM LIMIT OF MATRIX INVERSE PROBLEMS*

P. DEIFT† AND E. TRUBOWITZ‡

Abstract. In this paper the authors consider the Lanczos algorithm for periodic Jacobi-matrix inverse problems. In the usual form this algorithm does not have a continuum limit, but by mixing periodic and antiperiodic spectra and introducing a set of *weights*, the authors show that the modified Lanczos scheme does indeed have a continuum limit. Moreover, this limit is precisely the inverse scheme for the Schrödinger equation which was introduced by the authors in [Comm. Pure Appl. Math., 32 (1979), pp. 121–151].

1. Introduction. In the past few years much work has been done on the inverse spectral theory of ordinary differential operators, and many unexpected connections have been discovered with integrable Hamiltonian systems, Riemann surfaces and the representation theory of Lie groups. At the same time the inverse spectral problem for difference operators (equivalently, band matrices) has been studied and similar connections have been found. However, the exact relationship between the discrete and continuous inverse spectral theories has not been fully understood, and clearly it is of theoretical and computational interest to make the relationship precise.

Case and Kac [1973] and Case [1973] discretized the Gel'fand–Levitan inverse method for second order ordinary differential operators, but they found that the corresponding discrete operator is not the usual centered second difference approximation. In the other direction, Boley and Golub [1977], [1978] refined and generalized the Lanczos computational inverse scheme for periodic Jacobi matrices (i.e., centered second difference operators), but their algorithm has no immediate continuous analogue. The purpose of this paper is to show that by supplying suitable *weights*, the Lanczos algorithm for periodic Jacobi matrices does indeed have a continuum limit.

Let $q(x)$ be a real smooth function of period 1 and let $L_{n,p}$ $(n \geqq 1)$ be the periodic difference operator on $\mathbb{R}^n$ defined by[1]

$$(L_P y)(l) = \begin{cases} [-n^2 y(n) + 2n^2 y(1) - n^2 y(2)] + q\left(\frac{1}{n}\right) y(1), & l = 1, \\ [-n^2 y(l-1) + 2n^2 y(l) - n^2 y(l+1)] + q\left(\frac{l}{n}\right) y(l), & 1 < l < n, \\ [-n^2 y(n-1) + 2n^2 y(n) - n^2 y(1)] + q(1) y(n), & l = n. \end{cases}$$

Let $\lambda_{P,1} < \lambda_{P,2} \leqq \lambda_{P,3} \leqq \cdots \leqq \lambda_{P,n}$ be the spectrum of L_P and $g_1, \cdots, g_n$ an associated orthogonal frame of eigenfunctions normalized by $1/n \sum_{l=1}^{n} g_k^2(l) = 1$. Now, by the spectral theorem for matrices, we have

$$L_P = \left(\frac{1}{\sqrt{n}} g_j(i)\right) \begin{pmatrix} \lambda_{P,1} & & 0 \\ & \ddots & \\ 0 & & \lambda_{P,n} \end{pmatrix} \left(\frac{1}{\sqrt{n}} g_i(j)\right)$$

* Received by the editors July 2, 1980 and in revised form February 25, 1981.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was supported by the National Science Foundation under grant NSF-MCS-76-07521.

‡ Courant Institute and Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139. The research of this author was supported by the National Science Foundation, under grant NSF-MCS-78-18222.

[1] We will frequently suppress the dependence on n.

and, in particular

$$q\left(\frac{l}{n}\right) = -2n^2 + \frac{1}{n}\sum_{k=1}^{n} \lambda_{P,k} g_k^2(l), \qquad 1 \leqq l \leqq n.$$

The following is a common inverse problem. Recover $q(l/n)$ $(1 \leqq l \leqq n)$ from $\lambda_{P,k}$, $(1 \leqq k \leqq n)$ and the auxiliary data $g_k(1)$, $g_k(2)$ $(1 \leqq k \leqq n)$[2]. There is a simple solution. First, reconstruct $q(1/n)$ from $\lambda_{P,k}$, $g_k(1)$, $(1 \leqq k \leqq n)$ and the identity

$$q\left(\frac{1}{n}\right) = -2n^2 + \frac{1}{n}\sum_{k=1}^{n} \lambda_{P,k} g_k^2(1).$$

Next, construct $g_k(3)$, $(1 \leqq k \leqq n)$ from $q(1/n)$ and $\lambda_{P,k}$, $g_k(1)$, $g_k(2)$, $(1 \leqq k \leqq n)$ by writing the difference equation as

$$g_k(3) = n^{-2}\left[q\left(\frac{1}{n}\right) + 2n^2 - \lambda_{P,k}\right] g_k(2) - g_k(1), \qquad 1 \leqq k \leqq n.$$

Now, reconstruct $q(2/n)$ from $\lambda_{P,k}$, $g_k(2)$ $(1 \leqq k \leqq n)$ and the identity $q(2/n) = -2n^2 + (1/n)\sum_{k=1}^{n} \lambda_{P,k} g_k^2(2)$. Once again, use $q(2/n)$, $\lambda_{P,k}$, $g_k(2)$, $g_k(3)$ and the difference equation to construct $g_k(4)$, $(1 \leqq k \leqq n)$. Iterating this procedure we obtain $q(l/n)$, $(1 \leqq k \leqq n)$.

This scheme cannot converge as $n \to \infty$, because

$$q\left(\frac{l}{n}\right) = -2n^2 + \frac{1}{n}\sum_{k=1}^{n} \lambda_{P,k} g_k^2(l),$$

as it stands, has no continuum analogue (see also § 4). Our point is that the way around this difficulty is to introduce a set of weights, the antiperiodic spectrum and modified auxiliary data.

Let L_A be the antiperiodic difference operator on $\mathbb{R}^n$ defined by

$$(L_A y)(l) = \begin{cases} [n^2 y(n) + 2n^2 y(1) - n^2 y(2)] + q\left(\frac{1}{n}\right) y(1), & l = 1, \\ (L_P y)(l), & 1 < l < n, \\ [-n^2 y(n-1) + 2n^2 y(n) + n^2 y(1)] + q(1) y(n), & l = n. \end{cases}$$

Let $\lambda_{A,1} \leqq \lambda_{A,2} < \cdots \leqq \lambda_{A,n}$ be its spectrum and $h_1, \cdots, h_n$ a corresponding orthogonal frame of eigenfunctions normalized by $(1/n)\sum_{l=1}^{n} h_k^2(l) = 1$. The spectra $\lambda_{P,k}$, $\lambda_{A,k}$ $(1 \leqq k \leqq n)$ have a band structure

$$\lambda_{n,0} < \lambda_{n,1} \leqq \lambda_{n,2} < \lambda_{n,3} \leqq \lambda_{n,4} < \cdots < \lambda_{n,2n-3} \leqq \lambda_{n,2n-2} < \lambda_{n,2n-1},$$

where $\lambda_{n,0} = \lambda_{P,1}$, $\lambda_{n,1} = \lambda_{A,1}$, $\lambda_{n,2} = \lambda_{A,2}$, $\lambda_{n,3} = \lambda_{P,2}$, $\lambda_{n,4} = \lambda_{P,3}$ and so on. Let $f_{n,0} = g_1$, $f_{n,1} = h_1$, $f_{n,2} = h_2$, $f_{n,3} = g_2$, $f_{n,4} = g_3 \cdots$ and introduce the weights[3]

$$\varepsilon_{n,k} = \frac{-n^{2n-1}\dot{\Delta}_n(\lambda_{n,2k})}{\left[\prod_{j=0, j\neq k}^{n-1} (\lambda_{n,2j} - \lambda_{n,2k})\right]} \geqq 0, \qquad 0 \leqq k \leqq n-1,$$

[2] There is some redundancy in this data; a reduction to minimal data can be found in Boley and Golub [1978].

[3] $\dot{} = \frac{\partial}{\partial \lambda}$.

where

$$\Delta(\lambda)=2+(-1)^{n+1}(\lambda_{n,0}-\lambda)\Big[\prod_{i=2,4,6,\cdots}(\lambda_{n,2i-1}-\lambda)(\lambda_{n,2i}-\lambda)\Big].$$

Then, as we will show in § 2,

$$1=\sum_{k=0}^{n-1}\varepsilon_{n,k},$$

$$1=\sum_{k=0}^{n-1}\varepsilon_{n,k}f_{n,2k}^2(l),\qquad 1\leqq l\leqq n,$$

and

$$q\Big(\frac{l}{n}\Big)=\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l)-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}(f_{n,2k}(l+1)-f_{n,2k}(l))^2.$$

Now, the appropriately modified inverse problem is: Recover $q(l/n)$ $(1\leqq l\leqq n)$ from $\lambda_{P,k}$, $(1\leqq k\leqq n)$ and the auxiliary data $f_{n,2k}(1)$, $f_{n,2k}(2)$ $(0\leqq k\leqq n-1)$. To solve this problem, first recover the λ_A's from the λ_P's (this is done in § 2) and then construct the $\varepsilon_{n,k}$'s. Next, reconstruct $q(1/n)$ from $\lambda_{n,2k}$, $f_{n,2k}(1)$, $f_{n,2k}(2)$, $(0\leqq k\leqq n-1)$ and the identity

$$q\Big(\frac{1}{n}\Big)=\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(1)-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}(f_{n,2k}(2)-f_{n,2k}(1))^2,$$

and then continue iteratively, as before, to recover $q(l/n)$ $(1\leqq l\leqq n)$.

This scheme has a continuum limit as follows.

Consider Hill's operator $-(d^2/dx^2)+q(x)$, with band spectrum

$$\lambda_0<\lambda_1\leqq\lambda_2<\lambda_3\leqq\lambda_4<\cdots$$

and normalized eigenfunctions $f_0, f_1, f_2, \cdots$. We will prove in § 3 that $\lim_{n\to\infty}\varepsilon_{n,k}=\varepsilon_k$ exists and

$$1=\sum_{k\geqq 0}\varepsilon_k f_{2k}^2(x),\qquad 0\leqq x\leqq 1.$$

This identity among squares of eigenfunctions appears in McKean–Trubowitz [1976] and motivated the use of the weights $\varepsilon_{n,k}$. Now differentiate the identity twice to obtain

$$q(x)=\sum_{k\geqq 0}\lambda_{2k}\varepsilon_k f_{2k}^2(x)-\sum_{k\geqq 0}\varepsilon_k(f_{2k}'(x))^2,$$

which is the continuum limit of

$$q\Big(\frac{l}{n}\Big)=\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l)-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}(f_{n,2k}(l+1)-f_{n,2k}(l))^2.$$

The iterative inversion procedure becomes the infinite system of ordinary differential equations

$$f_{2n}''(x)+\lambda_{2n}f_{2n}(x)=\Big[\sum_{k\geqq 0}\lambda_{2k}\varepsilon_k f_{2k}^2(x)-\varepsilon_k(f_{2k}'(x))^2\Big]f_{2n}(x),\qquad n\geqq 0,$$

and the auxiliary data become, in the continuum, the initial data. Deift and Trubowitz [1979] used an analogous system to solve the problem of inverse scattering on the line.

We will not carry out all the technical details necessary to justify this picture, but the heart of the matter is in § 3, where we show that for each $0 \leqq x \leqq 1$ and $k \geqq 0$,

$$\lim_{n\to\infty} \varepsilon_{n,k} f_{n,2k}^2([nx]) = \varepsilon_k f_{2k}^2(x),$$

where $[nx]$ is the smallest integer $\geqq nx$.

In § 4 we point out some of the peculiarities of this continuum limit.

2. Some identities for Jacobi matrices. Let

$$J_P = \begin{bmatrix} b_1 & a_1 & 0 & \cdots & & & a_n \\ a_1 & b_2 & a_2 & & & & \\ 0 & a_2 & b_2 & a_3 & & & \\ & & & & a_{n-2} & b_{n-1} & a_{n-1} \\ a_n & \cdots & & & & a_{n-1} & b_n \end{bmatrix}$$

and

$$J_A = \begin{bmatrix} b_1 & a_1 & 0 & \cdots & & & -a_n \\ a_1 & b_2 & a_2 & & & & \\ 0 & a_2 & b_3 & a_3 & & & \\ & & & & a_{n-2} & b_{n-1} & a_{n-1} \\ -a_n & \cdots & & & & a_{n-1} & b_n \end{bmatrix}$$

denote $n \times n$ periodic and antiperiodic Jacobi matrices, respectively, with b_l real and $a_l < 0$. *We will only consider the case n odd*, which is sufficient for § 3. The spectra[4] $\{\lambda_{P,i}\}_{i=1,\cdots,n}$, $\{\lambda_{A,i}\}_{i=1,\cdots,n}$ of J_P and J_A have a *band structure*

$$\lambda_0 < \lambda_1 \leqq \lambda_2 < \lambda_3 \leqq \lambda_4 < \cdots < \lambda_{2n-3} \leqq \lambda_{2n-2} < \lambda_{2n-1},$$

where

$$\begin{aligned} \lambda_0 &= \lambda_{P,1}, & \lambda_1 &= \lambda_{A,1}, \\ \lambda_3 &= \lambda_{P,2} & \lambda_2 &= \lambda_{A,2}, \\ \lambda_4 &= \lambda_{P,3}, & &\vdots \\ &\vdots & & \\ \lambda_{2n-3} &= \lambda_{P,n-1}, & & \\ \lambda_{2n-2} &= \lambda_{P,n}, & \lambda_{2n-1} &= \lambda_{A,n}. \end{aligned}$$

The gap (λ_1, λ_2) closes if and only if $\lambda_1 = \lambda_2$ is a double eigenvalue of L_A, and so on.

Set $A = (\prod_{l=1}^{n} a_l)$ and extend a_l and b_l by periodicity; $a_{l+n} = a_l$, $b_{l+n} = b_l$ to all of $\mathbb{Z}$. Fix $h > 0$, and let $y_1(l, \lambda, a, b)$, $y_2(l, \lambda, a, b)$[5] denote the solutions of

$$a_{l-1} y_j(l-1, \lambda) + b_l y_j(l, \lambda) + a_l y_j(l+1, \lambda) = \lambda y_j(l, \lambda), \qquad \lambda \in \mathbb{C}, \quad j = 1, 2, \quad l \in Z$$

with initial conditions

$$y_1(0, \lambda) = 1, \quad y_1(1, \lambda) = 1, \qquad y_2(0, \lambda) = 0, \quad y_2(1, \lambda) = h,$$

respectively. $y_1(l, \lambda)$ and $[y_2(l, \lambda)/h]$ are polynomials in λ, with identical leading terms

[4] See Hochstadt [1974] and van Moerbeke [1979] for the basic spectral theory of Jacobi matrices.

[5] $a = (a_1 \cdots a_n)$, $b = (b_1, \cdots, b_n)$. The dependence on a or b will most often be suppressed.

$\lambda^{l-1}/[\prod_{j=1}^{l-1} a_j]$. Now define the *discriminant*

$$\Delta(\lambda, a, b) \equiv y_1(n, \lambda) + \frac{y_2(n+1, \lambda) - y_2(n, \lambda)}{h},$$

which is a polynomial with leading term λ^n/A. The zeros of $\Delta(\lambda) - 2$ are the eigenvalues of L_P while the zeros of $\Delta(\lambda) + 2$ are the eigenvalues of L_A. Thus,

$$\Delta(\lambda) - 2 = \frac{\lambda_0 - \lambda}{-A}\left[\prod_{i=2,4,\cdots,n-1} (\lambda_{2i-1} - \lambda)(\lambda_{2i} - \lambda)\right]$$

and

$$\Delta(\lambda) + 2 = \frac{\lambda_{2n-1} - \lambda}{-A}\left[\prod_{i=1,3,\cdots,n-2} (\lambda_{2i-1} - \lambda)(\lambda_{2i} - \lambda)\right].$$

It follows that the λ_A's can be recovered from the λ_P's. Finally,[6] $\dot{\Delta}(\lambda_{2k}) = 0$ if and only if $\lambda_{2k} = \lambda_{2k-1}$. For $i = 0, 1, 2, \cdots, 2n-1$, let $f_i(l)$ denote the normalized periodic or antiperiodic eigenvectors $h \sum_{l=1}^{n} f_i^2(l) = 1$, corresponding to λ_i.

LEMMA 2.1. *If λ_{2k} is a simple eigenvalue, then*

$$\frac{\partial \Delta}{\partial a_l}(\lambda_{2k}) = -2h\dot{\Delta}(\lambda_{2k}) f_{2k}(l) f_{2k}(l+1)$$

and

$$\frac{\partial \Delta}{\partial b_l}(\lambda_{2k}) = -h\dot{\Delta}(\lambda_{2k}) f_{2k}^2(l)$$

for all l. Otherwise

$$\frac{\partial \Delta}{\partial a_l}(\lambda_{2k}) = \frac{\partial \Delta}{\partial b_l}(\lambda_{2k}) = 0.$$

Proof. We calculate $(\partial\Delta/\partial b_l)(\lambda_{2k})$; the a_l-derivative is calculated similarly. Assume first that λ_{2k} is a simple eigenvalue. Then $\lambda_{2k}(b)$ is a smooth function of b and, by first order perturbation theory, $\partial\lambda_{2k}/\partial b_l = h f_{2k}^2(l)$; indeed,

$$\begin{aligned}
\frac{\partial}{\partial b_l}\lambda_{2k} &= \frac{\partial}{\partial b_l}[h(f_{2k}, L_P f_{2k})] \\
&= h\left(\frac{\partial f_{2k}}{\partial b_l}, L_P f_{2k}\right) + h\left(f_{2k}, L_P \frac{\partial f_{2k}}{\partial b_l}\right) + h(f_{2k}, \delta_{\cdot l} f_{2k}) \\
&= 2h\lambda_{2k}\left(\frac{\partial f_{2k}}{\partial b_l}, f_{2k}\right) + h f_{2k}^2(l).
\end{aligned}$$

Now differentiate the equation $\Delta(\lambda_{2k}(b), b) = 2(-1)^k$ to obtain

$$\begin{aligned}
0 &= \frac{d}{db_l}\Delta(\lambda_{2k}(b), b) \\
&= \frac{\partial}{\partial b_l}\Delta(\lambda_{2k}) + \dot{\Delta}(\lambda_{2k})\frac{\partial \lambda_{2k}}{\partial b_l} \\
&= \frac{\partial}{\partial b_l}\Delta(\lambda_{2k}) + \dot{\Delta}(\lambda_{2k}) h(f_{2k}^2(l)).
\end{aligned}$$

This proves the lemma when λ_{2k} is simple.

---

[6] Again $\dot{\cdot} = \frac{\partial}{\partial \lambda}$.

It is not hard to prove the second assertion by taking a limit through matrices with simple eigenvalues. □

Remark. We may write $(\partial\Delta/\partial b_l)(\lambda_{2k}) = -h\dot{\Delta}(\lambda_{2k})f_{2k}^2(l)$, even when $\lambda_{2k-1} = \lambda_{2k}$ is degenerate. In this case, obviously, any choice of f_{2k} will do.

THEOREM 2.1. *Let*

$$\varepsilon_k = \frac{hA\dot{\Delta}(\lambda_{2k})}{[\prod_{j=0, j\neq k}^{n-1}(\lambda_{2j}-\lambda_{2k})]}, \qquad k = 0, \cdots, n-1.$$

Then

$$1 = \sum_{k=0}^{n-1} \varepsilon_k f_{2k}^2(l),$$

$$(\lambda_0 - \lambda_{2n-1}) + \sum_{j=1}^{n-1}(\lambda_{2j}-\lambda_{2j-1}) = 4a_l \sum_{k=0}^{n-1} \varepsilon_k f_{2k}(l) f_{2k}(l+1)$$

and

$$b_l = \frac{\lambda_{2n-1}-\lambda_0}{2} + \frac{1}{2}\sum_{j=1}^{n-1}(\lambda_{2j-1}-\lambda_{2j}) + \sum_{k=0}^{n-1}\lambda_{2k}\varepsilon_k f_{2k}^2(l)$$

for all l. Moreover, $\varepsilon_0 > 0$ *and* $\varepsilon_k \geqq 0$, $(k = 1, \cdots, n-1)$ *with equality if and only if* $\lambda_{2k} = \lambda_{2k-1}$.

Proof.

$$\begin{aligned} -A(\Delta - 2) &= (\lambda_0 - \lambda)\prod_{i=2,4,\cdots,n-1}(\lambda_{2i-1}-\lambda)(\lambda_{2i}-\lambda) \\ &= \det(L_P - \lambda) \\ &= -\lambda^n + \Big[\sum_{l=1}^{n} b_l\Big]\lambda^{n-1} - \Big[\sum_{1\leqq i<j\leqq n} b_i b_j - \sum_{i=1}^{n} a_i^2\Big]\lambda^{n-2} + O(\lambda^{n-3}). \end{aligned}$$

Here $(\partial\Delta/\partial b_l)(\lambda)$ is a polynomial of degree $n-1$ which can be interpolated off the n points $\lambda_0 < \lambda_2 < \cdots < \lambda_{2n-2}$ to get

$$\begin{aligned} \frac{\partial\Delta}{\partial b_l}(\lambda) &= \sum_{k=0}^{n-1}\Big[\prod_{\substack{j=0\\ j\neq k}}^{n-1}\Big(\frac{\lambda-\lambda_{2j}}{\lambda_{2k}-\lambda_{2j}}\Big)\Big]\frac{\partial\Delta}{\partial b_l}(\lambda_{2k}) \\ &= \sum_{k=0}^{n-1}\Big[\prod_{\substack{j=0\\ j\neq k}}^{n-1}\Big(\frac{\lambda-\lambda_{2j}}{\lambda_{2k}-\lambda_{2j}}\Big)\Big][-h\dot{\Delta}(\lambda_{2k})f_{2k}^2(l)], \end{aligned}$$

so that

$$\begin{aligned} \lambda^{n-1} - \Big[\sum_{\substack{m=1\\ m\neq l}}^{n} b_m\Big]\lambda^{n-2} + O(\lambda^{n-3}) &= \sum_{k=0}^{n-1}\varepsilon_k f_{2k}^2(l)\Big[\prod_{\substack{j=0\\ j\neq k}}^{n-1}(\lambda_{2j}-\lambda)\Big] \\ &= \lambda^{n-1}\Big[\sum_{k=0}^{n-1}\varepsilon_k f_{2k(l)}^2\Big] - \lambda^{n-2}\Big[\sum_{k=0}^{n-1}\varepsilon_k\Big(\sum_{\substack{j=0\\ j\neq k}}^{n-1}\lambda_{2j}\Big)f_{2k}^2(l)\Big] \\ &\quad + O(\lambda^{n-3}). \end{aligned}$$

Equating coefficients we obtain $1=\sum_{k=0}^{n-1}\varepsilon_k f_{2k}^2(l)$ and

$$\sum_{\substack{m=1\\ m\neq l}}^{n} b_m=\sum_{k=0}^{n-1}\varepsilon_k\Big[\sum_{\substack{j=0\\ j\neq k}}^{n-1}\lambda_{2j}\Big]f_{2k}^2(l)$$

$$=\Big[\sum_{j=0}^{n-1}\lambda_{2j}\Big]\Big[\sum_{k=0}^{n-1}\varepsilon_k f_{2k}^2(l)\Big]-\sum_{k=0}^{n-1}\lambda_{2k}\varepsilon_k f_{2k}^2(l).$$

But

$$\sum_{\substack{m=1\\ m\neq l}}^{n} b_m=\Big[\sum_{m=1}^{n} b_m\Big]-b_l=\Big[\sum_{i=1}^{n}\lambda_{P,i}\Big]-b_l.$$

Also, $-A(\Delta-2)=-A(\Delta+2)+4A$, which implies $\sum_{i=1}^{n}\lambda_{P,i}=\sum_{i=1}^{n}\lambda_{A,i}$, so that $\sum_{i=1}^{n}\lambda_{P,i}=\frac{1}{2}\sum_{i=0}^{2n-1}\lambda_i$. Thus,

$$b_l=\sum_{i=1}^{n}\lambda_{P,i}-\sum_{i=0}^{n-1}\lambda_{2i}+\sum_{k=0}^{n-1}\lambda_{2k}\varepsilon_k f_{2k}^2(l)$$

$$=\tfrac{1}{2}\sum_{i=0}^{2n-1}\lambda_i-\sum_{i=0}^{n-1}\lambda_{2i}+\sum_{k=0}^{n-1}\lambda_{2k}\varepsilon_k f_{2k}^2(l)$$

$$=\tfrac{1}{2}(\lambda_{2n-1}-\lambda_0)+\tfrac{1}{2}\sum_{i=1}^{n-1}(\lambda_{2i-1}-\lambda_{2i})+\sum_{k=0}^{n-1}\lambda_{2k}\varepsilon_k f_{2k}^2(l).$$

It is easily checked that $\varepsilon_k\geqq 0$, and that $\varepsilon_k=0$ if and only if $\lambda_{2k}=\lambda_{2k-1}$.

The proof of the second identity is similar. □

Let $q(x)$ be a real smooth function of period 1. Setting $a_l=-n^2$, $b_l=q(l/n)+2n^2$ and $h=1/n$, we obtain

$$[-n^2 f(l-1)+2n^2 f(l)-n^2 f(l+1)]+q\Big(\frac{l}{n}\Big)f(l),$$

a discrete approximation to Hill's operator, $-f''(x)+q(x)f(x)$.

In an obvious notation the formulae of this section now become

$$\Delta_n(\lambda)=y_{n,1}(n,\lambda)+n(y_{n,2}(n+1,\lambda)-y_{n,2}(n,\lambda)),$$

$$\Delta_n(\lambda)-2=\frac{\lambda_{n,0}-\lambda}{n^{2n}}\Big[\prod_{i=1,4,\cdots,n-1}(\lambda_{n,2i-1}-\lambda)(\lambda_{n,2i}-\lambda)\Big],$$

$$\Delta_n(\lambda)+2=\frac{\lambda_{n,2n-1}-\lambda}{n^{2n}}\Big[\prod_{i=1,3,\cdots,n-2}(\lambda_{n,2i-1}-\lambda)(\lambda_{n,2i}-\lambda)\Big],$$

$$1=n^{-1}\sum_{k=1}^{n} f_{n,k}^2(l),\qquad 1=\sum_{k=0}^{n-1}\varepsilon_{n,k}f_{n,2k}^2(l),$$

$$q\Big(\frac{l}{n}\Big)=-2n^2+\frac{\lambda_{n,2n-1}-\lambda_{n,0}}{2}+\frac{1}{2}\sum_{k=1}^{n-1}(\lambda_{n,2k-1}-\lambda_{n,2k})+\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l),$$

$$\frac{(\lambda_{n,0}-\lambda_{n,2n-1})}{4}+\frac{1}{4}\sum_{k=1}^{n-1}(\lambda_{n,2k}-\lambda_{n,2k-1})=-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}f_{n,2k}(l)f_{n,2k}(l+1),$$

$$\varepsilon_{n,k}=\frac{-n^{2n-1}\dot{\Delta}_n(\lambda_{n,2k})}{[\prod_{j=0,j\neq k}^{n-1}(\lambda_{n,2j}-\lambda_{n,2k})]}.$$

Finally, we have $1=\sum_{k=0}^{n}\varepsilon_{n,k}$ by summing the identity $1=\sum_{k=0}^{n}\varepsilon_{n,k}f_{n,2k}^2(l)$ over l and

$$\begin{aligned}q\left(\frac{l}{n}\right)&=-2n^2+\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l)\\&\quad+n^2\sum_{k=0}^{n-1}2\varepsilon_{n,k}f_{n,2k}(l)f_{n,2k}(l+1)\\&=\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l)\\&\quad+n^2\sum_{k=0}^{n-1}2\varepsilon_{n,k}f_{n,2k}(l)f_{n,2k}(l+1)\\&\quad-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}(f_{n,2k}^2(l)+f_{n,2k}^2(l+1))\\&=\sum_{k=0}^{n-1}\lambda_{n,2k}\varepsilon_{n,k}f_{n,2k}^2(l)\\&\quad-n^2\sum_{k=0}^{n-1}\varepsilon_{n,k}(f_{n,2k}(l+1)-f_{n,2k}(l))^2.\end{aligned}$$

3. The continuum limit. Let $-(d^2/dx^2)+q(x)$ be the Hill's operator introduced in the last section. The associated band spectrum[7] is

$$\lambda_0<\lambda_1\leqq\lambda_2<\lambda_3\leqq\lambda_4<\cdots,$$

with corresponding normalized eigenfunctions $f_0, f_1, f_2, \cdots$. Let $y_1(x,\lambda)$, $y_2(x,\lambda)$ be the solutions of

$$-y''+q(x)y=\lambda y,$$

with $y_1(0,\lambda)=y_2'(0,\lambda)=1$, $y_1'(0,\lambda)=y_2(0,\lambda)=0$. The discriminant $\Delta(\lambda)=y_1(1,\lambda)+y_2'(1,\lambda)$ is an entire function of order $\frac{1}{2}$ with $\Delta(\lambda_0)=2$, $\Delta(\lambda_{2i-1})=\Delta(\lambda_{2i})=2(-1)^i$, $(i\geqq 1)$ and $\dot{\Delta}(\lambda_{2k})=0$ if and only if $\lambda_{2k}=\lambda_{2k-1}$. It follows from Hadamard's factorization theorem and the asymptotics $\Delta(\lambda)\sim 2\cos\sqrt{\lambda}$ $(\lambda\to-\infty)$ that

$$\Delta(\lambda)-2=(\lambda_0-\lambda)\prod_{i=2,4,6,\cdots}\frac{(\lambda_{2i-1}-\lambda)(\lambda_{2i}-\lambda)}{i^4\pi^4},$$

$$\Delta(\lambda)+2=4\prod_{i=1,3,5,\cdots}\frac{(\lambda_{2i-1}-\lambda)(\lambda_{2i}-\lambda)}{i^4\pi^4}.$$

Set

$$\varepsilon_o=\frac{-\dot{\Delta}(\lambda_0)}{\left[\prod_{i\geqq 1}\frac{\lambda_{2i}-\lambda_0}{i^2\pi^2}\right]}$$

and

$$\varepsilon_k=\frac{-k^2\pi^2\dot{\Delta}(\lambda_{2k})}{\left[(\lambda_0-\lambda_{2k})\prod_{i\neq k}\frac{\lambda_{2i}-\lambda_{2k}}{i^2\pi^2}\right]},\qquad k\geqq 1;$$

$\varepsilon_0>0$, and $\varepsilon_k\geqq 0$ with equality if and only if $\lambda_{2k}=\lambda_{2k-1}$.

[7] See McKean–Trubowitz [1976] for more information.

THEOREM 3.1 (pointwise convergence). *For each* $x \in [0, 1]$ *and* $k \geqq 0$,[8]

$$\lim_{n\to\infty} \varepsilon_{n,k} f_{n,2k}^2([nx]) = \varepsilon_k f_{2k}^2(x),$$

where $[nx]$ *is the smallest integer* $\geqq nx$.

Proof. Step 1. By standard numerical techniques (see, e.g., Isaacson and Keller [1966]),

$$\lim_{n\to\infty} y_{n,j}([nx], \lambda) = y_j(x, \lambda)$$

and

$$\lim_{n\to\infty} n[y_{n,j}([nx]+1, \lambda) - y_{n,j}([nx], \lambda)] = y_j'(x, \lambda), \qquad j = 1, 2,$$

locally uniformly in x and λ. Hence

$$\begin{aligned}\Delta_n(\lambda) &= y_{n,1}(n, \lambda) + n(y_{n,2}(n+1, \lambda) - y_{n,2}(n, \lambda)) \\ &\to y_k(1, \lambda) + y_2'(1, \lambda) = \Delta(\lambda),\end{aligned}$$

uniformly on bounded λ-sets, so that

$$(*) \qquad \lambda_{n,k} \to \lambda_k, \qquad k = 0, 1, 2, \cdots.$$

Step 2. If $q = 0$, we have

$$\lambda_{n,0} = 0,$$

$$\lambda_{n,2j-1} = \lambda_{n,2j} = 4n^2 \sin^2\left(\frac{j\pi}{2n}\right), \qquad j = 1, \cdots, n-1,$$

$$\lambda_{n,2n-1} = 4n^2$$

and, in general, it follows from min-max that[9]

$$-K \leqq \lambda_{n,0} \leqq K,$$

$$(**) \qquad 4j^2 - K \leqq 4n^2 \sin^2 \frac{j\pi}{2n} - K \leqq \lambda_{n,2j-1}, \lambda_{n,2j} \leqq 4n^2 \sin^2\left(\frac{j\pi}{2n}\right) + K < j^2\pi^2 + K, \qquad j = 1, \cdots, n-1,$$

$$4n^2 - K \leqq \lambda_{n,2n-1} \leqq 4n^2 + K,$$

where we have used the inequality $(\sin y/y) \geqq 2/\pi$ for $0 \leqq y \leqq \pi/2$.

Step 3. We show that

$$\lim_{n\to\infty} \varepsilon_{n,k} = \varepsilon_k.$$

We do this for even $k > 0$; if $k = 0$ or k is odd, the argument is similar. Evaluate

$$\frac{\Delta_n(\lambda) - 2}{\lambda_{n,2k} - \lambda} = \frac{(\lambda_{n,0} - \lambda)(\lambda_{n,2k-1} - \lambda)}{n^{2n}} \prod_{\substack{i=2,4,\cdots,n-1 \\ i \neq k}} (\lambda_{n,2i-1} - \lambda)(\lambda_{n,2i} - \lambda)$$

[8] Here and below $n \to \infty$ through *odd* values.

[9] We use K to denote a generic constant which depends only on q and its derivatives.

at $\lambda = \lambda_{n,2k}$ to obtain

$$\dot{\Delta}_n(\lambda_{n,2k}) = \frac{(\lambda_{n,0}-\lambda_{n,2k})(\lambda_{n,2k}-\lambda_{n,2k-1})}{n^{2n}} \prod_{\substack{i=2,4,\cdots,n-1 \\ i\neq k}} (\lambda_{n,2i-1}-\lambda_{n,2k})(\lambda_{n,2i}-\lambda_{n,2k}),$$

$$= \frac{(\lambda_{n,0}-\lambda_{n,2k})(\lambda_{n,2k}-\lambda_{n,2k-1})}{n^{2n}} \cdot \left[\prod_{\substack{i=2,4,\cdots,n-1 \\ i\neq k}} \left(\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right)\right]\left[\prod_{\substack{i=2,4,\cdots,n-1 \\ i\neq k}} (\lambda_{n,2i}-\lambda_{n,2k})\right]^2.$$

It follows from (*) that

$$\left[\frac{n^n \dot{\Delta}_n(\lambda_{n,2k})}{\prod_{i=2,4,\cdots,n-1, i\neq k}(\lambda_{n,2i}-\lambda_{n,2k})}\right]^2$$

$$= (\lambda_{n,0}-\lambda_{n,2k})(\lambda_{n,2k}-\lambda_{n,2k-1})\dot{\Delta}_n(\lambda_{n,2k}) \prod_{\substack{i=2,4,\cdots,n-1 \\ i\neq k}} \left(\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right)$$

$$= (\lambda_{n,0}-\lambda_{n,2k})(\lambda_{n,2k}-\lambda_{n,2k-1})\dot{\Delta}_n(\lambda_{n,2k}) \prod_{\substack{i=2,4,\cdots,n-1 \\ i\neq k}} \left(1+\frac{\lambda_{n,2i-1}-\lambda_{n,2i}}{\lambda_{n,2i}-\lambda_{n,2k}}\right),$$

which converges to

$$(\lambda_0-\lambda_{2k})(\lambda_{2k}-\lambda_{2k-1})\dot{\Delta}(\lambda_{2k}) \prod_{\substack{i=2,4,\cdots \\ i\neq k}} \left[1+\frac{\lambda_{2i-1}-\lambda_{2i}}{\lambda_{2i}-\lambda_{2k}}\right]$$

$$= (\lambda_0-\lambda_{2k})(\lambda_{2k}-\lambda_{2k-1})\dot{\Delta}(\lambda_{2k}) \prod_{\substack{i=2,4,\cdots \\ i\neq k}} \left[\frac{\lambda_{2i-1}-\lambda_{2k}}{\lambda_{2i}-\lambda_{2k}}\right],$$

because, by (**),

$$|\lambda_{n,2i-1}-\lambda_{n,2i}| \leqq K$$

and

$$\lambda_{n,2i}-\lambda_{n,2k} \geqq (4i^2-K)-(k^2\pi^2+K) = 4i^2-(k^2\pi^2+2K)$$

for all i and n. The identity

$$\Delta_n(\lambda)+2 = \frac{\lambda_{n,2n-1}-\lambda}{n^{2n}} \prod_{i=1,2,\cdots,n-2} (\lambda_{n,2i-1}-\lambda)(\lambda_{n,2i}-\lambda)$$

$$= \frac{\lambda_{n,2n-1}-\lambda}{n^{2n}}\left[\prod_{i=1,3,\cdots,n-2} \frac{(\lambda_{n,2i-1}-\lambda)}{(\lambda_{n,2i}-\lambda)}\right] \cdot \left[\prod_{i=1,3,\cdots,n-2} (\lambda_{n,2i}-\lambda)\right]^2$$

yields

$$\left[\frac{n^n}{\sqrt{\lambda_{n,2n-1}-\lambda_{n,2k}}\prod_{i=1,3,\cdots,n-2}(\lambda_{n,2i}-\lambda_{n,2k})}\right]^2 \to \frac{1}{\Delta(\lambda_{2k})+2}\prod_{i=1,3,\cdots}\left(\frac{\lambda_{2i-1}-\lambda_{2k}}{\lambda_{2i}-\lambda_{2k}}\right).$$

However,

$$n(\lambda_{n,2n-1}-\lambda_{n,2k})^{-1/2}\to\frac{1}{2} \qquad ((*) \text{ and } (**)),$$

and

$$\frac{1}{\Delta(\lambda_{2k})+2}=\frac{1}{4}\prod_{i=1,3,\cdots}\frac{i^4\pi^4}{(\lambda_{2i-1}-\lambda_{2k})(\lambda_{2i}-\lambda_{2k})},$$

so that

$$\left[\frac{n^{n-1}}{\prod_{i=1,3,\cdots,n-2}(\lambda_{n,2i}-\lambda_{n,2k})}\right]^2\to\left[\prod_{i=1,3\cdots}\frac{i^2\pi^2}{\lambda_{2i}-\lambda_{2k}}\right]^2.$$

It follows that, if $\lambda_{2k-1}=\lambda_{2k}$,

$$\left[\frac{n^{2n-1}\dot{\Delta}_n(\lambda_{n,2k})}{\prod_{i=1,2,\cdots,n-1,i\neq k}(\lambda_{n,2i}-\lambda_{n,2k})}\right]^2\to 0=\varepsilon_k,$$

and, if $\lambda_{2k-1}\neq\lambda_{2k}$ (so that $\dot{\Delta}(\lambda_{2k})\neq 0$),

$$\left[\frac{n^{2n-1}\dot{\Delta}_n(\lambda_{n,2k})}{\prod_{i=1,2,\cdots,n-1,i\neq k}(\lambda_{n,2i}-\lambda_{n,2k})}\right]^2\to\frac{k^4\pi^4[\dot{\Delta}(\lambda_{2k})]^2}{\left[\prod_{i=2,4,\cdots,i\neq k}\frac{(\lambda_{2i-1}-\lambda_{2k})(\lambda_{2i}-\lambda_{2k})}{i^4\pi^4}\right]}$$

$$\cdot\left[\prod_{i=2,4,\cdots,i\neq k}\frac{\lambda_{2i-1}-\lambda_{2k}}{\lambda_{2i}-\lambda_{2k}}\right]\left[\prod_{i=1,3,\cdots}\frac{i^2\pi^2}{\lambda_{2i}-\lambda_{2k}}\right]^2$$

$$=\left[\frac{k^2\pi^2\dot{\Delta}(\lambda_{2k})}{\prod_{i=1,2,\cdots,i\neq k}\left[\frac{\lambda_{2i}-\lambda_{2k}}{i^2\pi^2}\right]}\right]^2=\varepsilon_k^2.$$

This completes the proof of step 3.

Step 4. We complete the proof of Theorem 3.1. Write $f_{n,2k}(l)=a_{n,k}y_{n,1}(l,\lambda_{n,2k})+b_{n,k}y_{n,2}(l,\lambda_{n,2k})$, where $a_{n,k}$ and $b_{n,k}$ are real. Then

$$1=n^{-1}\sum_{l=1}^{n}f_{n,2k}^2(l)$$

$$=(a_{n,k},\,b_{n,k})$$

$$\cdot\begin{bmatrix} n^{-1}\sum_{l=1}^{n}y_{n,1}^2(l,\lambda_{n,2k}) & n^{-1}\sum_{l=1}^{n}y_{n,1}(l,\lambda_{n,2k})y_{n,2}(l,\lambda_{n,2k}) \\ n^{-1}\sum_{l=1}^{n}y_{n,1}(l,\lambda_{n,2k})y_{n,2}(l,\lambda_{n,2k}) & n^{-1}\sum_{l=1}^{n}y_{n,2}^2(l,\lambda_{n,2k}) \end{bmatrix}$$

$$\cdot\begin{bmatrix} a_{n,k} \\ b_{n,k} \end{bmatrix}.$$

Now,

$$n^{-1}\sum_{l=1}^{n} y_{n,i}(l,\lambda_{n,2k})y_{n,j}(l,\lambda_{n,2k})\to\int_0^1 y_i(x,\lambda_{2k})y_j(x,\lambda_{2k})\,dx,\qquad 1\leqq i,j\leqq 2,$$

and the matrix

$$\begin{bmatrix}\int_0^1 y_1^2(x,\lambda_{2k})\,dx & \int_0^1 y_1(x,\lambda_{2k})y_2(x,\lambda_{2k})\,dx\\ \int_0^1 y_1(x,\lambda_{2k})y_2(x,\lambda_{2k})\,dx & \int_0^1 y_2^2(x,\lambda_{2k})\,dx\end{bmatrix}$$

is strictly positive definite. It follows that for large n, $1\geqq\gamma(a_{n,k}^2+b_{n,k}^2)$ for some $\gamma>0$. In particular, $a_{n,k}$ and $b_{n,k}$ are bounded and, if $\lambda_{2k-1}=\lambda_{2k}$, we have[10] $\varepsilon_{n,k}f_{n,2k}^2([nx])\to 0=\varepsilon_k f_{2k}^2(x)$.

If $\lambda_{2k-1}<\lambda_{2k}$ and λ_{2k} is a Dirichlet eigenvalue, i.e., $y_2(1,\lambda_{2k})=0$, $a_{n,k}$ must converge to 0. If not, there exist $a\neq 0$, b and a subsequence $\{k(m)\}$ such that $a_{n,k(m)}\to a$ and $b_{n,k(m)}\to b$. In particular, $f_{n,2k(m)}([nx])$ and $n(f_{n,2k(m)}([nx]+1)-f_{n,2k(m)}([nx]))$ converge to $ay_1(x,\lambda_{2k})+by_2(x,\lambda_{2k})$, $ay_1'(x,\lambda_{2k})+by_2'(x,\lambda_{2k})$ respectively. As $f_{n,2k(m)}(0)=\pm f_{n,2k(m)}(n)$ and $f_{n,2k(m)}(1)=\pm f_{n,2k(m)}(n+1)$, $ay_1(x,\lambda_{2k})+by_2(x,\lambda_{2k})$ is clearly a periodic/antiperiodic eigenfunction of Hill's equation. But as λ_{2k} is simple and $y_2(1,\lambda_{2k})=0$, the only periodic/antiperiodic eigenfunction is $y_2(x,\lambda_{2k})$, which contradicts $a\neq 0$. Hence, $a_{n,k}\to 0$ and it follows that $b_{n,k}\to[\int_0^1 y_2^2(x,\lambda_{2k})\,dx]^{-1/2}$. Thus,

$$\varepsilon_{n,k}f_{n,2k}^2([nx])\to\varepsilon_k\left[\left(\int_0^1 y_2^2(x,\lambda_{2k})\,dx\right)^{-1}y_2^2(x,\lambda_{2k})\right]=\varepsilon_k f_{2k}^2(x).$$

Finally, suppose $\lambda_{2k-1}<\lambda_{2k}$ and $y_2(1,\lambda_{2k})\neq 0$. Set
(k even)

$$g_{n,k}(l)=y_{n,2}(n,\lambda_{n,2k})y_{n,1}(l,\lambda_{n,2k})+(1-y_{n,1}(n,\lambda_{n,2k}))y_{n,2}(l,\lambda_{n,2k}),$$

(k odd)

$$h_{n,k}(l)=y_{n,2}(n,\lambda_{n,2k})y_{n,1}(l,\lambda_{n,2k})-(1+y_{n,1}(n,\lambda_{n,2k}))y_{n,2}(l,\lambda_{n,2k}).$$

Using $\Delta_n(\lambda_{n,2k})=\pm 2$ and the Wronskian identity

$$y_{n,2}(l+1,\lambda_{n,2k})y_{n,1}(l,\lambda_{n,2k})-y_{n,2}(l,\lambda_{n,2k})y_{n,1}(l+1,\lambda_{n,2k})=\text{const}=n^{-1},$$

it is easily checked that $g_{n,k}$ and $h_{n,k}$ are periodic and antiperiodic eigenvectors, respectively. Moreover,

$$n^{-1}\sum_{l=1}^{n} g_{n,k}^2(l)\to\int_0^1[y_2(1,\lambda_{2k})y_1(x,\lambda_{2k})+(1-y_1(1,\lambda_{2k}))y_2(x,\lambda_{2k})]^2\,dx>0$$

and

$$n^{-1}\sum_{l=1}^{n} h_{n,k}^2(l)\to\int_0^1[y_2(1,\lambda_{2k})y_1(x,\lambda_{2k})-(1+y_1(1,\lambda_{2k}))y_2(x,\lambda_{2k})]^2\,dx>0.$$

[10] If $\lambda_{2k}=\lambda_{2k-1}$ there is ambiguity in the choice of $f_{2k}(x)$. But this is no problem since $\varepsilon_k=0$.

Therefore,

$$f_{n,2k}(l) \equiv \begin{cases} g_{n,k}\left[n^{-1}\sum_{l=1}^{n} g_{n,k}^2(l)\right]^{-1/2} & (k \text{ even}), \\ h_{n,k}\left[n^{-1}\sum_{l=1}^{n} h_{n,k}^2(l)\right]^{-1/2} & (k \text{ odd}) \end{cases}$$

converges and the proof of Theorem 3.1 is finished. □

THEOREM 3.2.[11]

(1) $$\sum_{k\geqq 0} \varepsilon_k = 1.$$

(2) $$\sum_{k\geqq 0} \varepsilon_k f_{2k}^2(x) = 1.$$

Proof. (1) We compute the contour integral

$$\left(\frac{1}{2\pi i}\right)\int_{|\lambda|=(N+\frac{1}{2})^2\pi^2} \frac{\dot{\Delta}(\lambda)}{\left[(\lambda_0-\lambda)\prod_{i\geqq 1}\frac{\lambda_{2i}-\lambda}{i^2\pi^2}\right]}\,d\lambda,$$

as $N\to\infty$. The residue of the integrand at the simple pole $\lambda=\lambda_{2k}$ is ε_k. Also, it is not hard to show that

$$\sup_{|\lambda|=(N+\frac{1}{2})^2\pi^2}\left|\frac{\dot{\Delta}(\lambda)}{\left[\prod_{i\geqq 1}\frac{\lambda_{2i}-\lambda}{i^2\pi^2}\right]}+1\right| = o(1).$$

Therefore, by the residue theorem,

$$\sum_{|\lambda_{2k}|<(N+\frac{1}{2})^2\pi^2} \varepsilon_k \to 1.$$

(2) For any fixed N,

$$\sum_{k=0}^{N} \varepsilon_k f_{2k}^2(x) = \lim_{n\to\infty}\sum_{k=0}^{N} \varepsilon_{n,k} f_{n,2k}^2([nx]) \leqq 1,$$

by Theorem 2.1. Thus, $c(x)\equiv\sum_{k\geqq 0}\varepsilon_k f_{2k}^2(x)\leqq 1$ and, by Fubini,

$$\int_0^1 c(x)\,dx = \sum_{k\geqq 0}\varepsilon_k\int_0^1 f_{2k}^2(x)\,dx = \sum_{k\geqq 0}\varepsilon_k = 1.$$

This is only possible if $c(x)=1$ a.e., and the result will follow if we show that $c(x)$ is continuous. As $\sum_{k\geqq 0}\varepsilon_k<\infty$, it suffices to prove that $f_{2k}(x)$ is uniformly bounded in x and in k. Write $f_{2k}(x)=a_k y_1(x,\lambda_{2k})+b_k\sqrt{\lambda_{2k}}y_2(x,\lambda_{2k})$. From the estimates

$$y_1(x,\lambda)=\cos\sqrt{\lambda}x+O(\lambda^{-1/2}), \qquad y_2(x,\lambda)=\frac{\sin\sqrt{\lambda}x}{\sqrt{\lambda}}+O(\lambda^{-1}),$$

we have

$$1=\int_0^1 f_{2k}^2(x)\,dx \geqq \frac{1}{4}(a_k^2+b_k^2)$$

for k large. Clearly, $f_{2k}(x)$ is bounded and the proof is finished. □

[11] For another proof of these formulae, see Deift and Trubowitz [1980].

4. Some remarks on the continuum limit. The difficulty in taking the continuum limit of the standard Lanczos algorithm is clearly seen by considering the formula

$$\sum_{k=1}^{n} \left(\frac{1}{\sqrt{n}} g_k(l)\right)\left(\frac{1}{\sqrt{n}} g_k(m)\right) = \delta_{lm}, \qquad 1 \leqq l, m \leqq n,$$

which follows from the orthonormality of the eigenfunctions $\{g_k\}$. The continuum limit is

$$\sum_{k=1}^{\infty} g_k(x) g_k(y) = \delta(x-y)$$

(convergence is in the sense of distributions or as a quadratic form in $L^2(0, 1)$). The difficulty with this formula, from the inverse point of view, is that it cannot be evaluated along the diagonal $x = y$. By contrast, in the weighted scheme we have

$$\sum_{k=0}^{\infty} \varepsilon_k f_{2k}^2(x) = 1, \qquad 0 \leqq x \leqq 1,$$

which is basic and leads, in particular, to an interpretation of inverse spectral theory as an integrable system of constrained harmonic oscillators (see Moser and Trubowitz; see also Deift, Lund and Trubowitz [1980]).

In § 3 we showed that $\varepsilon_{n,k} f_{n,2k}^2([nx]) \to \varepsilon_k f_{2k}^2(x)$. As $\sum_{k=0}^{n-1} \varepsilon_{n,k} f_{n,2k}^2(l) = 1$, we can immediately conclude that

$$\sum_{k=0}^{\infty} \varepsilon_k f_{2k}^2(x) = \lim_{N\to\infty} \lim_{n\to\infty} \sum_{k=0}^{N} \varepsilon_{n,k} f_{n,2k}^2([nx]) \leqq 1, \qquad 0 \leqq x \leqq 1.$$

In § 3, we proved equality by a contour integral; to prove equality directly, however, turns out to be a surprisingly intricate problem in rates of convergence (see Appendix for details). A related and somewhat puzzling question of convergence occurs in the formula

$$q(x) = \sum_{k=0}^{\infty} \varepsilon_k \lambda_{2k} f_{2k}^2(x) - \sum_{k=0}^{\infty} \varepsilon_k (f_{2k}'(x))^2$$

and its companion

$$(***) \qquad q(x) = 2 \sum_{k=0}^{\infty} \varepsilon_k \lambda_{2k} f_{2k}^2(x) + \text{const.}$$

$$\left(\frac{d}{dx}\left(\sum_{k=0}^{\infty} \varepsilon_k \lambda_k f_{2k}^2(x) + \sum_{k=0}^{\infty} \varepsilon_k (f_{2k}'(x))^2\right) = 0\right).$$

The point is that the expression $\sum_{k=0}^{n-1} \lambda_{n,2k} \varepsilon_{n,k} f_{n,2k}^2([nx])$ *cannot* converge to $\sum_{k=0}^{\infty} \varepsilon_k \lambda_{2k} f_{2k}^2(x)$. Indeed, from the formula

$$q\left(\frac{l}{n}\right) = \text{const.}\,(n) + \sum_{k=0}^{n-1} \lambda_{n,2k} \varepsilon_{n,k} f_{n,2k}^2(l)$$

on p. 805, we would conclude that

$$q(x) = \text{const.} + \sum_{k=0}^{\infty} \lambda_{2k} \varepsilon_k f_{2k}^2(x),$$

which can be reconciled with (***) only if $\sum_{k=0}^{\infty} \varepsilon_k \lambda_{2k} f_{2k}^2(x)$, and hence $q(x)$ is constant. In other words, the convergence of $\sum_{k=0}^{n-1} \varepsilon_{n,k} \lambda_{n,2k} f_{n,2k}^2([nx])$ involves an infinity which

is precisely cancelled by an infinity in the convergence of $n^2 \sum_{k=0}^{n-1} \varepsilon_{n,k}(f_{n,2k}([nx]+1) - f_{n,2k}([nx]))^2$ to give

$$q(x) = \sum_{k=0}^{\infty} \varepsilon_k \lambda_{2k} f_{2k}^2(x) - \sum_{k=0}^{\infty} \varepsilon_k (f'_{2k}(x))^2,$$

but we do not present any details. Beyond the technicalities (see also the remark in the Appendix), the reason why

$$1 = \sum_{k \geqq 0} \varepsilon_k f_{2k}^2(x)$$

and

$$q(x) = \sum_{k \geqq 0} \varepsilon_k \lambda_{2k} f_{2k}^2(x) - \sum_{k \geqq 0} \varepsilon_k (f'_{2k}(x))^2$$

can be interpreted as continuum limits, but

$$q(x) = 2 \sum_{k \geqq 0} \varepsilon_k \lambda_{2k} f_{2k}^2(x) + \text{const.}$$

cannot, remains unclear.

Appendix. Uniform convergence. Here we show $\sum_{k \geqq 0} \varepsilon_k = 1$ by direct estimation. As we have seen in § 3, this is enough to prove $\sum_{k \geqq 0} \varepsilon_k f_{2k}^2(x) = 1$.

LEMMA A.1. *Fix $0 < \delta < \frac{1}{2}$. Then there exists a number k_0, depending only on q and δ, for which*

$$\lambda_{n,2m} - \lambda_{n,2k} \geqq 3(m-k)k$$

whenever $k_0 \leqq k < m \leqq n^{1/2+\delta}$.

Proof.

$$\begin{aligned}
\lambda_{n,2m} - \lambda_{n,2k} &\geqq 4n^2\left(\sin^2\frac{m\pi}{2n} - \sin^2\frac{k\pi}{2n}\right) - 2K \\
&= 2n^2\left(\cos\frac{k\pi}{n} - \cos\frac{m\pi}{n}\right) - 2K \\
&= 2n(m-k)\sin\frac{\theta\pi}{n} - 2K \qquad (k < \theta < m) \\
&> 2n(m-k)\left|\left(\frac{\theta\pi}{n}\right)\frac{2}{\pi}\right| - 2K > 4(m-k)k - 2K,
\end{aligned}$$

as $\sin y / y \geqq (2/\pi)$ for $y \leqq (\pi/2)$ and $\theta\pi/n \leqq \pi/n^{1/2-\delta} \leqq \pi(k_0^{-(1-2\delta)/(1+2\delta)}) \leqq \pi/2$ for k_0 large enough. The result follows if, in addition, k_0 is chosen $\geqq 2K$. □

LEMMA A.2. *Fix $0 < \delta < \frac{1}{2}$. Then there exists a number $\lambda^0 > 0$, depending only on q and δ, for which*

$$|\dot{\Delta}_n(\lambda)| \leqq \frac{K}{\sqrt{\lambda}}$$

whenever $\lambda^0 \leqq \lambda \leqq n^{1+2\delta}$.

Proof. $\dot{\Delta}_n(\lambda) = \dot{y}_{n,1}(n,\lambda) + n[\dot{y}_{n,2}(n+1,\lambda) - \dot{y}_{n,2}(n,\lambda)]$. We will only show that $|\dot{y}_{n,1}(n,\lambda)| \leqq K/\sqrt{\lambda}$; the second term is treated similarly. As in the Hill's case, $y_{n,1}(l,\lambda)$

satisfies an "integral" equation

$$y_{n,1}(l,\lambda)=\frac{\cos((l-\frac{1}{2})\alpha)}{\cos(\alpha/2)}+\sum_{m=1}^{l}\frac{\sin((l-m)\alpha)}{n^2\sin\alpha}q\left(\frac{m}{n}\right)y_{n,1}(m,\lambda),$$

where $\sin\alpha=(\sqrt{\lambda}/n)(1-\lambda/4n^2)^{1/2}$. Clearly, $\alpha=\sqrt{\lambda}/n+O((\sqrt{\lambda}/n)^3)$. Now, for λ^0 large enough so that $\sqrt{\lambda}/n\leqq n^{-(1/2-\delta)}\leqq(\lambda^0)^{-(1-2\delta)/2(1+2\delta)}$ is small enough, we have

$$n\sin\alpha\geqq\frac{\sqrt{\lambda}}{2}\quad\text{and}\quad\cos\left(\frac{\alpha}{2}\right)\geqq\frac{1}{2}.$$

Thus

$$\max_{1\leqq l\leqq n}|y_{n,1}(l,\lambda)|\leqq 2+\frac{2}{\sqrt{\lambda}}\left[n^{-1}\sum_{m=1}^{n}\left|q\left(\frac{m}{n}\right)\right|\right]\left[\max_{1\leqq l\leqq n}y_{n,1}(l,\lambda)\right].$$

As $n^{-1}\sum_{m=1}^{n}|q(m/n)|\to\int_0^1|q(x)|\,dx<\infty$, we conclude that $\max_{1\leqq l\leqq n}|y_{n,1}(l,\lambda)|\leqq K_1$ for $\sqrt{\lambda}>\sqrt{\lambda^0}$ sufficiently large. Differentiating with respect to λ, we find

$$\begin{aligned}\dot{y}_{n,1}(l,\lambda)=&-\left(\cos\left(\frac{\alpha}{2}\right)\sin\left(\left(l-\frac{1}{2}\right)\alpha\right)\left(l-\frac{1}{2}\right)\right.\\&\left.+\frac{1}{2}\cos\left(\left(l-\frac{1}{2}\right)\alpha\right)\sin\left(\frac{\alpha}{2}\right)\right)\dot{\alpha}\Big/\cos^2\left(\frac{\alpha}{2}\right)\\&+\sum_{m=1}^{l}\frac{\sin((l-m)\alpha)}{n^2\sin\alpha}q\left(\frac{m}{n}\right)\dot{y}_{n,1}(m,\lambda)\\&+\dot{\alpha}\sum_{m=1}^{l}\left[\frac{(l-m)\sin\alpha\cos((l-m)\alpha)-\sin((l-m)\alpha)\cos\alpha}{n^2(\sin\alpha)^2}\right]\\&\cdot q\left(\frac{m}{n}\right)y_{n,1}(m,\lambda).\end{aligned}$$

The estimate $|\dot{\alpha}|=|2n^2\sin\alpha|^{-1}\leqq 1/n\sqrt{\lambda}$ implies

$$\left|-\left(\cos\left(\frac{\alpha}{2}\right)\sin\left(\left(l-\frac{1}{2}\right)\alpha\right)\left(l-\frac{1}{2}\right)+\frac{1}{2}\cos\left(\left(l-\frac{1}{2}\right)\alpha\right)\sin\left(\frac{\alpha}{2}\right)\right)\dot{\alpha}\Big/\cos^2\left(\frac{\alpha}{2}\right)\right|\leqq 4\lambda^{-1/2}$$

and

$$\left|\dot{\alpha}\left[\frac{(l-m)\sin\alpha\cos((l-m)\alpha)-\sin((l-m)\alpha)\cos\alpha}{n(\sin\alpha)^2}\right]\right|\leqq 2\lambda^{-1}+4\lambda^{-3/2}.$$

It follows as above that

$$\max_{1\leqq l\leqq n}|\dot{y}_{n,1}(l,\lambda)|\leqq\frac{K_2}{\sqrt{\lambda}}$$

which proves the lemma. □

Now write

$$n^{4n}(\Delta_n^2(\lambda)-4)=(\lambda_{n,0}-\lambda)(\lambda_{n,2n-1}-\lambda)\prod_{i=1}^{n-1}(\lambda_{n,2i-1}-\lambda)(\lambda_{n,2i}-\lambda)$$

and differentiate to obtain

$$\frac{n^{4n}\dot{\Delta}_n(\lambda_{n,2k})}{[\prod_{i=0,1,\cdots,n-1,i\neq k}(\lambda_{n,2i}-\lambda_{n,2k})]^2}$$
$$= \pm\frac{(\lambda_{n,2n-1}-\lambda_{n,2k})}{4}\left[\frac{\lambda_{n,2k-1}-\lambda_{n,2k}}{\lambda_{n,0}-\lambda_{n,2k}}\right]\prod_{\substack{i=1\\ i\neq k}}^{n-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right] \qquad \begin{pmatrix}+, k \text{ odd}\\ -, k \text{ even}\end{pmatrix}.$$

Therefore,

$$(\varepsilon_{n,k})^2 = \left[\frac{n^{4n-2}\dot{\Delta}_n(\lambda_{n,2k})}{\prod_{i=0,1,\cdots,n-1}(\lambda_{n,2i}-\lambda_{n,2k})^2}\right]\dot{\Delta}_n(\lambda_{n,2k})$$
$$= \pm\left[\frac{\lambda_{n,2n-1}-\lambda_{n,2k}}{4n^2}\right]\left[\frac{\lambda_{n,2k-1}-\lambda_{n,2k}}{\lambda_{n,0}-\lambda_{n,2k}}\right]\dot{\Delta}_n(\lambda_{n,2k})$$
$$\cdot\prod_{\substack{i=1\\ i\neq k}}^{n-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right].$$

As $(\lambda_{n,2i-1}-\lambda_{n,2k})/(\lambda_{n,2i}-\lambda_{n,2k})\leqq 1$ for $k<i\leqq n-1$,

$$\prod_{\substack{i=1\\ i\neq k}}^{n-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right]\leqq\prod_{i=1}^{k-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right].$$

From Lemma A.1,

$$\prod_{i=k_0}^{k-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right]=\prod_{i=k_0}^{k-1}\left[1+\frac{\lambda_{n,2i-1}-\lambda_{n,2i}}{\lambda_{n,2i}-\lambda_{n,2k}}\right]$$
$$\leqq\prod_{i=k_0}^{k-1}\left[1+\frac{K}{3(i-k)k}\right]$$
$$\leqq\exp\frac{K}{3}\sum_{i=k_0}^{k-1}\frac{1}{(k-i)i}$$
$$\leqq\exp\frac{K}{3}\int_{k_0-1}^{k-1}\frac{dx}{(k-x)x}$$
$$=\exp\frac{K}{3k}\log\left[\frac{(k-1)(k-k_0+1)}{k_0-1}\right]\to 1 \quad \text{as } k\to\infty.$$

Moreover the finite product

$$\prod_{i=1}^{k_0-1}\left[1+\frac{\lambda_{n,2i-1}-\lambda_{n,2i}}{\lambda_{n,2i}-\lambda_{n,2k}}\right]\leqq\left[1+\frac{K}{4k^2-\pi^2(k_0-1)^2-2K}\right]^{k_0-1},$$

which is bounded for k large enough, say $k>k_1>k_0$; it follows that

$$\prod_{\substack{i=1\\ i\neq k}}^{n-1}\left[\frac{\lambda_{n,2i-1}-\lambda_{n,2k}}{\lambda_{n,2i}-\lambda_{n,2k}}\right]$$

is bounded in $k>k_1$, independent of n. Using $\lambda_{n,2n-1}/4n^2\leqq K_1$, $\lambda_{n,2k}\geqq 4k^2-K_2$ and Lemma A.2, it is now easy to see that for $k_2>k_1$ sufficiently large,

$$\varepsilon_{n,k}\leqq\frac{K_3}{(\lambda_{n,2k})^{3/4}}\leqq\frac{K_4}{k^{3/2}},$$

whenever $n^{(1/2)+\delta}>k>k_2$.

We also have the estimate

$$\sum_{k \geqq n^{(1/2)+\delta}}^{n-1} \varepsilon_{n,k} \leqq K n^{-2\delta}.$$

To see this, sum the formula for $q(l/n)$, derived in § 2, to obtain

$$\sum_{k=0}^{n-1} \lambda_{n,2k}\varepsilon_{n,k} = \left[n^{-1} \sum_{l=1}^{n} q\left(\frac{l}{n}\right)\right] + \frac{(4n^2 - \lambda_{n,2n-1}) + \lambda_{n,0}}{2} + \frac{1}{2}\sum_{k=1}^{n-1} (\lambda_{n,2k} - \lambda_{n,2k-1}).$$

But $\lambda_{n,2k} - \lambda_{n,2k-1} \leqq K$ and $|4n^2 - \lambda_{n,2n-1}| \leqq K$. Thus $\sum_{k=0}^{n-1} \lambda_{n,2k}\varepsilon_{n,k}$ is $O(n)$ and

$$\sum_{k \geqq n^{(1/2)+\delta}}^{n-1} \varepsilon_{n,k} \leqq [4n^{1+2\delta} - K]^{-1}\left[\sum_{k \geqq n^{1/2+\delta}}^{n-1} \lambda_{n,2k}\varepsilon_{n,k}\right] = O(n^{-2\delta})$$

as $\lambda_{n,2k} \geqq 4k^2 - K$.

If we write

$$1 = \sum_{k=0}^{n-1} \varepsilon_{n,k} = \sum_{k=0}^{k_2} \varepsilon_{n,k} + \sum_{k=k_2+1}^{[n^{(1/2)+\delta}]} \varepsilon_{n,k} + \sum_{k=[n^{(1/2)+\delta}]+1}^{n-1} \varepsilon_{n,k}$$

and use the above estimates, the identity $\sum_{k \geqq 0} \varepsilon_k = 1$ is immediate and we are done.

Remark. We are able to prove convergence for $\sum_{k=0}^{n-1} \varepsilon_{n,k} = 1$, essentially, by controlling $\sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k} f_{n,2k}^2([nx])$ through the formula

$$q\left(\frac{l}{n}\right) = -2n^2 + \frac{(\lambda_{n,2n-1} - \lambda_{n,0})}{2} + \frac{1}{2}\sum_{k=1}^{n-1} (\lambda_{n,2k-1} - \lambda_{n,2k}) + \sum_{k=0}^{n-1} \lambda_{n,2k}\varepsilon_{n,k} f_{n,2k}^2(l)$$

of § 2. Could we obtain a similar formula for $\sum_{k=0}^{n-1} \lambda_{n,2k}^2 \varepsilon_{n,k}$ and use this to control $\sum_{k=0}^{n-1} \lambda_{n,2k}\varepsilon_{n,k} f_{n,2k}^2([nx])$ (cf., § 4, where we show that this quantity *must* diverge)? What goes wrong?

Indeed, by the methods of § 2, we can derive the formula

$$\begin{aligned} q\left(\frac{l}{n}\right)\left[-\frac{\lambda_{n,2n-1}}{2} + \frac{\lambda_{n,0}}{2} - \frac{1}{2}\sum_{j=0}^{n-1} (\lambda_{n,2j-1} - \lambda_{n,2j})\right] + q^2\left(\frac{l}{n}\right) + 4n^2 q\left(\frac{l}{n}\right) \\ = 6n^4 + \left[\sum_{0 \leqq i < k \leqq n-1} \lambda_{n,2i}\lambda_{n,2k} - \sum_{1 \leqq k < l \leqq n} \lambda_{n,k}^{(P)} \lambda_{n,l}^{(P)}\right] \\ - \left[2n^2 + \sum_{j=0}^{n-1} \lambda_{n,2j}\right]\left[\frac{\lambda_{n,2n-1}}{2} - \lambda_{n,0} + \frac{1}{2}\sum_{j=1}^{n-1} (\lambda_{n,2j-1} - \lambda_{n,2j})\right] \\ + \sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2 f_{n,2k}^2\left(\frac{l}{n}\right), \end{aligned} \tag{X}$$

where $\lambda_{n,k}^{(P)}$, $k = 1, \cdots, n$, are the periodic eigenvalues. Now it is clear from the preceding estimates that we would indeed be able to prove convergence for $\sum_{k=0}^{n-1} \lambda_{n,2k}\varepsilon_{n,k} f_{n,2k}^2([nx])$ if we could show that

$$\sum_{k \geqq n^{1/2+\delta}}^{n-1} \varepsilon_{n,k}\lambda_{n,2k} \to 0 \quad \text{as } n \to \infty$$

for some $0<\delta<\frac{1}{2}$. But

$$\sum_{k\geqq n^{1/2+\delta}}^{n-1} \varepsilon_{n,k}\lambda_{n,2k} \leqq \frac{\text{const.}}{n^{1+2\delta}} \sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2,$$

so that we need $\sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2 = o(n^{1+2\delta})$ for some $0<\delta<\frac{1}{2}$. We will show that this cannot happen in general: choose $h_{n,i}$, $0\leqq i\leqq n-1$, such that

$$\sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} = 0, \qquad \sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} q\left(\frac{i}{n}\right) \geqq \text{const.} = \gamma > 0,$$

$$|h_{n,i}| \leqq 1, \qquad 0\leqq i\leqq n-1$$

(for example, if $q(x)$ is not even, i.e., $q(x)\not\equiv q(1-x)$, set $h_{n,i} = \pm[(1/(n-1))i - \frac{1}{2}]$, $0\leqq i\leqq n-1$). Multiplying (X) by $(1/n)h_{n,i}$ and summing, we get

$$\left[\sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} q\left(\frac{i}{n}\right)\right]\left[4n^2 - \frac{\lambda_{n,2n-1}}{2} + \frac{\lambda_{n,0}}{2} + \frac{1}{2}\sum_{j=0}^{n-1} (\lambda_{n,2j-1} - \lambda_{n,2j})\right]$$

$$+ \sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} q^2\left(\frac{i}{n}\right)$$

$$= \sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2 \left(\sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} f_{n,2k}^2(i)\right).$$

Now

$$\left|\sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} f_{n,2k}^2(i)\right| \leqq \sum_{i=0}^{n-1} \frac{1}{n} f_{n,2k}^2(i) = 1$$

and

$$\left|\sum_{i=0}^{n-1} \frac{1}{n} h_{n,i} q^2\left(\frac{i}{n}\right)\right| \leqq \sup_{0\leqq x\leqq 1} |q(x)|^2.$$

Thus, if $\sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2 = o(n^{1+2\delta})(0<\delta<\frac{1}{2})$, then the right-hand side would be $o(n^{1+2\delta})$, but the left-hand side grows like $2n^2\gamma$ ($\lambda_{n,2n-1}/2 \sim 2n^2$ and $|\sum_{j=0}^{n-1} (\lambda_{n,2j-1} - \lambda_{n,2j})| \leqq \text{const. } n$). This gives a contradiction, and we see that the higher moment $\sum_{k=0}^{n-1} \varepsilon_{n,k}\lambda_{n,2k}^2$ provides no information.

Acknowledgment. The authors gratefully acknowledge useful and stimulating discussions with Gene Golub. Also they would like to express their gratitude for the hospitality of I.H.E.S. in Paris where the final version of this paper was written and typed.

REFERENCES

H. P. McKean and E. Trubowitz, [1976], *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., 29, pp. 143–226.

P. Deift and E. Trubowitz, [1981], *An identity among squares of eigenfunctions*, Comm. Pure Appl. Math., to appear.

———, [1979], *Inverse scattering on the line*, Comm. Pure Appl. Math., 32, pp. 121–251.

P. van Moerbeke, [1976], *The spectrum of Jacobi matrices*, Inventiones Math., 37, pp. 45–81.

E. Isaacson and H. Keller, [1966], *Analysis of Numerical Methods*, John Wiley, New York.

D. Boley and G. H. Golub, [1977], *Inverse eigenvalue problems for band matrices*, Proc. of Dundee Conference on Numerical Analysis.

D. BOLEY AND G. H. GOLUB, [1978], *The matrix inverse eigenvalue problem for periodic Jacobi matrices*, Fourth Conference on Basic Problems of Numerical Analysis, (Liblice IV), Pilsen, Czechoslovakia.

J. MOSER AND E. TRUBOWITZ, (to appear), *Harmonic oscillators and hyperelliptic curves.*

P. DEIFT, F. LUND AND E. TRUBOWITZ, [1980], *Nonlinear waves and constrained harmonic motion*, Comm. Math. Phys., 74, pp. 141–188.

H. HOCHSTADT, [1974], *On the construction of a Jacobi matrix from spectral data*, Linear Algebra Appl. 8, pp. 435–446.

K. CASE AND M. KAC, [1973], *On discrete inverse scattering problems, Part I*, J. Math. Phys., 14, pp. 595 ff.

K. CASE, [1973], *On discrete inverse scattering problems, Part II*, J. Math. Phys., 14, pp. 916–920.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0002 $01.00/0

THE JABOTINSKY MATRIX OF A POWER SERIES*

J. L. LAVOIE† AND R. TREMBLAY‡

Abstract. We consider the inversion of formal power series and related results, in terms of matrices introduced by Jabotinsky.

1. Introduction. The theory of umbral operators introduced by Mullin and Rota in [16] and developed further by Garcia in [6] (see also [4], [7], [13]) furnishes a natural setting for the study of the inversion of formal power series in any number of variables.

The purpose of this paper is to point out that the important one-dimensional case can be elegantly and effectively set in terms of matrices introduced by Jabotinsky in [11] and [12].

2. The Jabotinsky matrix. Let

$$F(t)=\sum_{i=0}^{\infty} q_i t^i \in \mathscr{L}(\mathscr{F}) \quad \text{with } q_0=q\neq 0 \tag{2.1}$$

be a given formal power series (fps). The set $\mathscr{L}(\mathscr{F})$ is the totality of fps with coefficients $q_i \in \mathscr{F}$, a field of characteristic zero. Also, it is well known that $\mathscr{L}(\mathscr{F})$ itself becomes a field if we define the addition in $\mathscr{L}(\mathscr{F})$ by term-by-term sums and the multiplication by Cauchy products (see [9] and [10] for more details).

In [11], Jabotinsky associates with $F(t)$ an infinite dimensional matrix A whose elements, denoted by A_{ij}, are generated by

$$f^j(t)=\sum_{i=j}^{\infty} A_{ij}t^i, \qquad j=0,\pm 1,\pm 2,\cdots, \tag{2.2}$$

with $f(t)=tF(t)$. This is a lower triangular matrix:

$$A=\begin{bmatrix} & & \vdots & & \\ A_{-1-1} & 0 & 0 & 0 & 0 \\ A_{0-1} & A_{00} & 0 & 0 & 0 \\ \cdots A_{1-1} & A_{10} & A_{11} & 0 & 0\cdots \\ A_{2-1} & A_{20} & A_{21} & A_{22} & 0 \\ A_{3-1} & A_{30} & A_{31} & A_{32} & A_{33} \\ & & \vdots & & \end{bmatrix},$$

and the formal operational representation

$$A_{ij}=\frac{1}{(i-j)!}D^{i-j}F^j(t)\Big|_{t=0}, \qquad i\geqq j, \tag{2.3}$$

with $D=d/dt$, follows from Taylor's theorem. Of course, DF is the formal derivative of F defined by the fps

$$DF=\sum_{i=0}^{\infty}(i+1)q_{i+1}t^i \in \mathscr{L}(\mathscr{F}).$$

* Received by the editors April 26, 1979, and in revised form December 16, 1980.
† Département de mathématiques, Université Laval, Québec, Canada, G1K 7P4.
‡ Mathematics Department, Royal Military College of Canada, Kingston, Ontario, Canada, K7L 2W3.

We have

$$A_{i+1,1} = q_i, \quad A_{i0} = \delta_{i0}, \quad \delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}, \qquad i = 0, 1, 2, \cdots.$$

Also,

$$A_{ii} = q^i$$

$$A_{i+1,i} = iq_1 q^{i-1}$$

$$A_{i+2,i} = iq_2 q^{i-i} + \frac{i(i-1)}{2} q_1^2 q^{i-2}, \cdots, \text{etc.}$$

for $i = 0, \pm 1, \pm 2, \cdots$.

If $F(t) = 1$ then $A = I$, the unit matrix of infinite order. In table 1 we show a number of other special cases, with $i \geqq j = 1, 2, \cdots$.

TABLE 1

$F(t)$	q_i	A_{ij}
$1 + xt$	—	$x^{i-j}\binom{j}{i-j}$
$(1+xt)^{-1}$	$(-1)^i x^i$	$(-1)^{i-j} x^{i-j}\binom{i-1}{j-1}$
$\frac{\ln(1+t)}{t}$	$\frac{(-1)^i}{i+1}$	$\frac{j!}{i!} S_1(i, j)$
e^{xt}	$\frac{x^i}{i!}$	$\frac{(jx)^{i-j}}{(i-j)!}$
$\frac{t}{e^t - 1}$	$\frac{B_i}{i!}$	$\frac{1}{(i-j)!} B_{i-j}^{(j)}$
$(1 - 2xt + t^2)^{-1/2}$	$P_i(x)$	$C_{i-j}^{j/2}(x)$
$(1-t)^{-\alpha-1} \exp\left(\frac{-xt}{1-t}\right)$	$L_i^{(\alpha)}(x)$	$L_{i-j}^{[(\alpha+1)j-1]}(x)$

The $S_1(i, j)$ represents the Stirling numbers of the first kind, B_i and $B_i^{(j)}$ are, respectively, the Bernoulli numbers and the Bernoulli numbers of order j [15, p. 127]. The $P_i(x)$ and $C_i^{j/2}(x)$ are the Legendre and Gegenbauer polynomials while $L_i^{(\alpha)}(x)$ is the generalized Laguerre polynomial.

3. Pairs of inverse formal power series. Let $G(t)$ be another fps. Let $g(t) = tG(t)$ and let B be its Jabotinsky matrix. We have

$$f^j(g(t)) = \sum_{i=j}^{\infty} \sum_{r=j}^{i} B_{ir} A_{rj} t^i = \sum_{i=j}^{\infty} (BA)_{ij} t^i. \tag{3.1}$$

Thus the composition of two fps involves the ordinary (Cayley) product of their Jabotinsky matrix ((3.1) can also be denoted $f^j \circ g$, [9], [10]).

It is convenient to call $F(t)$ and $G(t)$ a pair of inverse fps provided that

$$f(g(t)) = t = g(f(t)). \tag{3.2}$$

In the light of (3.1), with $F(t)$ given by (2.1), this means that $G(t)$ must be such that

$$g^j(t) = \sum_{i=j}^{\infty} A_{ij}^{-1} t^i, \tag{3.3}$$

where A_{ij}^{-1} is the (i, j)-element of the inverse of A. These components are obtained from the Jabotinsky theorem which states that they are generated by

$$f^{-i-1}(t)Df(t) = \sum_{j=-\infty}^{i} A_{ij}^{-1} t^{-j-1}. \tag{3.4}$$

Proof. Using (2.2) and (3.4), it is found that $(A^{-1}A)_{ij}$ is the coefficient of t^{-1} in $f^{j-i-1}(t)Df(t)$. But, when $i > j$,

$$f^{j-i-1}(t)Df(t) = \frac{-1}{i-j} Df^{-(i-j)}(t) = \frac{-1}{i-j} \sum_{r=-(i-j)}^{\infty} rA_{r-(i-j)} t^{r-1}$$

and the coefficient of t^{-1} is zero, whereas for $i = j$ it is readily seen to be equal to one. The above argument is similar to Jabotinsky's proof [11].

Returning to (3.4), we obtain the remarkable explicit expression ([9], [10], [11]),

$$A_{ij}^{-1} = \frac{j}{i} A_{-j-i}, \qquad i \neq 0, \tag{3.5}$$

and, from (2.3), the formal operational representation

$$A_{ij}^{-1} = \frac{j}{i(i-j)!} D^{i-j} F^{-i}(t)\Big|_{t=0}, \qquad i \neq 0. \tag{3.6}$$

Thus,

$$G(t) = \sum_{i=0}^{\infty} p_i t^i \quad \text{with } p_i = \frac{1}{(i+1)!} D^i F^{-i-1}(t)\Big|_{t=0}$$

forms a pair of inverse fps with $F(t)$. Some explicit values of p_i are shown in Table 2.

TABLE 2

i	p_i
0	q^{-1}
1	$-q^{-3}q_1$
2	$q^{-5}(2q_1^2 - q)$
3	$-q^{-7}(5q_1^3 - 5qq_1q_2 + q^2q_3)$

Formal power series forming inverse pairs with, respectively, the first three entries of Table 1 are shown in Table 3, where $S_2(i, j)$ represents the Stirling numbers of the second kind.

TABLE 3

$G(t)$	p_i	A_{ij}^{-1}
$\dfrac{2}{1+(1+4xt)^{1/2}}$	$\dfrac{(-1)^i x^i}{i+1}\binom{2i}{i}$	$(-1)^{i-j} x^{i-j} \dfrac{j}{i}\binom{2i-j-1}{i-1}$
$(1-xt)^{-1}$	x^i	$x^{i-j}\binom{i-1}{j-1}$
$\dfrac{e^t-1}{t}$	$\dfrac{1}{(i+1)!}$	$\dfrac{j!}{i!} S_2(i, j)$

Let $F(t)$, $G(t)$ and $P(t)$, $Q(t)$ be two pairs of inverse fps. Then, clearly,

$$P(t)F(tP(t)), \qquad G(t)Q(tG(t))$$

is also a pair of inverses. The special case

$$P(t)=(1+kt)^{-1}, \qquad Q(t)=(1-kt)^{-1}$$

yields

$$\frac{1}{1+kt}F\left(\frac{t}{1+kt}\right), \qquad \frac{G(t)}{1-ktG(t)}.$$

This pair of inverse series has been considered in [3], where a number of interesting special cases are given. The first is the kth Euler transformation of a series while the second is the kth star transformation. Any number of pairs of inverses can be obtained in this way. For instance, using the first entries in Tables 1 and 3, for $P(t)$ and $Q(t)$, we have the pair

$$(1+kt)F(t(1+kt)), \qquad \frac{2G(t)}{1+[1+4ktG(t)]^{1/2}}.$$

4. The Bürmann-Lagrange expansion. An easy formal derivation of the Bürmann-Lagrange expansion can be found in [9], [10]. The relation

$$t^i=\sum_{i=j}^{\infty} A_{ij}^{-1}f^i(t),$$

obtained by replacing t by $f(t)$ in (3.3), is used in the fps $H(t)=\sum_{i=0}^{\infty} h_i t^i$ to obtain

$$H(t)=h_0+\sum_{i=1}^{\infty}\sum_{j=1}^{i} A_{ij}^{-1}h_j f^i(t).$$

But, from (3.6) and Taylor's theorem,

$$\sum_{j=1}^{i} A_{ij}^{-1}h_j=\frac{1}{i!}\sum_{j=0}^{i-1}\binom{i-1}{j}D^j[DH(t)]D^{i-j-1}F^{-i}(t)\Big|_{t=0}$$

$$=\frac{1}{i!}D^{i-1}[F^{-i}(t)DH(t)]_{t=0},$$

using Leibniz's formula for the nth derivative of a product. Therefore, the desired expansion is

$$H(t)=H(0)+\sum_{i=1}^{\infty} D^{i-1}\left[\left(\frac{t}{f(t)}\right)^i DH(t)\right]_{t=0}\frac{f^i(t)}{i!}.$$

5. The kth iterate of a formal power series. Let C be an $n+1$-square segment of A, the Jabotinsky matrix associated with $F(t)$. The columns $j=1,2,\cdots,n+1$ are chosen so that C is a lower triangular matrix whose eigenvalues $q, q^2, \cdots, q^{n+1}$ appear in the main diagonal.

For any $n+1$-square matrix having these eigenvalues, the Lagrange-Sylvester interpolation polynomial [5, pp. 101–103] can be used to obtain the interesting relation

$$C^m=\sum_{k=0}^{n} q^{(m-k)(n-k+1)}\begin{bmatrix} m \\ k \end{bmatrix}\begin{bmatrix} n-m \\ n-k \end{bmatrix}C^k, \tag{5.1}$$

where $m=0,\pm1,\pm2,\cdots$ and $\begin{bmatrix} x \\ k \end{bmatrix}$ is a q-binomial coefficient. There is an extensive

literature related to these coefficients, which satisfy relations somewhat similar to the relations satisfied by the ordinary binomial coefficients and reduce to them when $q \to 1$. For more details see [8].

A direct proof of (5.1) involves the q-analog of Saalchütz's theorem [17, p. 48]. However, with the result in hand, it is much simpler to use induction. Note first that both members of (5.1) are identical for $m = 0, 1, \cdots, n$. For $m = n+1$, we have

$$C^{n+1} = \sum_{k=0}^{n} (-1)^{n-k} q^{1/2(n-k+1)(n-k+2)} \begin{bmatrix} n+1 \\ k \end{bmatrix} C^k, \tag{5.2}$$

which can be obtained directly from the characteristic polynomial of C. Indeed, from the q-binomial theorem,

$$\prod_{k=0}^{n} (\lambda - q^{k+1}) = \sum_{k=0}^{n+1} (-1)^k q^{1/2k(k+1)} \begin{bmatrix} n+1 \\ k \end{bmatrix} \lambda^{n-k-1}.$$

(5.2) follows from an application of the Cayley–Hamilton theorem. Assuming the validity of (5.1), we multiply both side by C and use (5.2) on the right side. This member then reduces to an expression identical with the right-hand side of (5.1), but with m replaced by $m+1$. Since $q \neq 0$, C is nonsingular and (5.1) is still valid when m is a negative integer.

We are dealing with lower triangular matrices so that the (i, j) component of C^r is identical with the element in the same position in A^r, not only for $r = 1$ but for r equal to any positive or negative integer. Hence, equating the elements at the bottom of the j-column on each side of (5.1) yields, after a simple change of variable,

$$A_{ij}^{m} = \sum_{k=0}^{i-1} q^{(m-k)(i-k)} \begin{bmatrix} m \\ k \end{bmatrix} \begin{bmatrix} i-m-1 \\ i-k-1 \end{bmatrix} A_{ij}^{k}. \tag{5.3}$$

Here, $i \geqq j = 1, 2, \cdots$ and $m = 0, \pm 1, \pm 2, \cdots$, and we have written A_{ij}^{r} for $(A^r)_{ij}$. In virtue of the relation

$$\begin{bmatrix} -x \\ k \end{bmatrix} = (-1)^k q^{-kx-1/2k(k-1)} \begin{bmatrix} x+k-1 \\ k \end{bmatrix},$$

(5.3) can be written

$$A_{ij}^{m} = \sum_{k=0}^{i-1} (-1)^{i-k-1} q^{m-k+1/2(i-k)(i-k-1)} \begin{bmatrix} m \\ k \end{bmatrix} \begin{bmatrix} m-k-1 \\ i-k-1 \end{bmatrix} A_{ij}^{k}. \tag{5.4}$$

This is Tambs Lyche's formula [18], obtained more than fifty years ago from different considerations. For $m = -1$, (5.3) reduces to

$$A_{ij}^{-1} = \frac{j}{i} A_{-j-i} = \sum_{k=0}^{i-1} (-1)^k q^{-1/2(k+1)(2i-k)} \begin{bmatrix} i \\ k+1 \end{bmatrix} A_{ij}^{k}. \tag{5.5}$$

In particular, by noticing that $A_{ii}^{r} = q^{ir}$, we have from (5.3) and (5.5)

$$\sum_{k=0}^{i-1} q^{-k(m-k)} \begin{bmatrix} m \\ k \end{bmatrix} \begin{bmatrix} i-m-1 \\ i-k-1 \end{bmatrix} = 1$$

and

$$\sum_{k=0}^{i-1} (-1)^k q^{1/2k(k+1)} \begin{bmatrix} i \\ k+1 \end{bmatrix} = 1.$$

Also for $F(t)=(1+t)^{-1}$, we have $q=1$, $f_k(t)=t(1+kt)^{-1}$ and hence

$$A_{ij}^{k}=(-k)^{i-j}\binom{i-1}{j-1}.$$

This is replaced in (5.3) to obtain

$$\sum_{k=0}^{i-1}(-1)^{i-k-1}\binom{m}{k}\binom{m-k-1}{i-k-1}k^{i-j}=m^{i-j}.$$

Finally, we should perhaps remark that in view of (5.5) it is tempting to look for a new relation between the pair of inverse fps $f(t)$ and $g(t)$. We begin with the expansion

$$f_k^i(t)=\sum_{i=j}^{\infty}A_{ij}^{k}t^i,$$

where $f_0(t)=t$, $f_1(t)=f(t)$, $f_2(t)=f(f(t))$ and $f_k(t)$ is the k-fold composition of $f(t)$; and note that $i!A_{i1}^{k}=D^i f_k(t)|_{t=0}\equiv f_k^{(i)}(0)$. Hence, with $q=1$,

$$\begin{aligned}g(t)&=\sum_{i=1}^{\infty}A_{i1}^{-1}t^i\\&=\sum_{i=1}^{\infty}\sum_{k=0}^{i-1}(-1)^k\binom{i}{k+1}\frac{1}{i!}f_k^{(i)}(0)t^i\\&=\sum_{k=0}^{\infty}\frac{(-1)^k t^{k+1}}{(k+1)!}\sum_{i=0}^{\infty}\frac{1}{i!}D^i\{f_k^{(k+1)}(t)\}_{t=0}t^i,\end{aligned}$$

which we may write as

$$g(t)=\sum_{k=1}^{\infty}(-1)^{k+1}f_{k-1}^{(k)}(0)\frac{t^k}{k!}.$$

This result, given by Brun [1], should be handled with great care because of the absence of a complementary term which, in general, is not zero. The papers [14] and [2] should be consulted.

Acknowledgments. The authors would like to thank Professor Henrici for many valuable suggestions and helpful criticisms.

REFERENCES

[1] V. BRUN, *Sur la formule d'inversion de M. Tambs Lyche*, C.R. Acad. Sci. Paris, 194 (1932), pp. 2276–2277.

[2] ———, *Une formule d'inversion corrigée*, Math. Scand., 3 (1955), pp. 224–228.

[3] R. DONAGHEY, *Two transformations of series that commute with compositional inversion*, J. Combinatorial Theory, Ser. A, 27 (1979), pp. 360–364.

[4] J. P. FILLMORE AND S. G. WILLIAMSON, *A linear algebra setting for the Rota-Mullin theory of polynomials of binomial type*, J. Linear and Multilinear Algebra, 1 (1973), pp. 67–80.

[5] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York.

[6] A. M. GARCIA, *An exposé of the Mullin-Rota theory of polynomials of binomial type*, J. Linear and Multilinear Algebra, 1 (1973), pp. 47–65.

[7] A. M. GARCIA AND S. A. JONI, *A new expression for the umbral operators and power series inversion*, Proc. Amer. Math. Soc., 64 (1977), pp. 179–185.

[8] H. W. GOULD, *The operator* $(a^x\Delta)^n$ *and Stirling numbers of the first kind*, Amer. Math. Month., 71 (1964), pp. 850–858.

[9] P. HENRICI, *An algebraic proof of the Lagrange-Bürmann formula*, J. Math. Anal. Appl. 8 (1964), pp. 218–225.
[10] ———, *Applied and Computational Complex Analysis*, vol. 1, Wiley and Sons, N.Y., 1974.
[11] E. JABOTINSKY, *Representation of functions by matrices. Application to Faber polynomials*, Proc. Amer. Math. Soc., 4 (1953), pp. 546–553.
[12] ———, *Analytic iteration*, Trans. Amer. Math. Soc., 108 (1963), pp. 457–477.
[13] S. A. JONI, *Lagrange inversion in higher dimensions and umbral operators*, J. Linear and Multilinear Algebra, 6 (1978), pp. 111–121.
[14] O. KOLBERG, *Über ein Problem von Viggo Brun*, Math. Scand. 3 (1955), pp. 221–223.
[15] L. M. MILNE-THOMSON, *The Calculus of Finite Difference*, Macmillan, London, 1933.
[16] G.-C. ROTA AND R. MULLIN, *On the foundation of combinatorial theory*, Graph Theory and its Applications, Academic Press, New York, 1970, pp. 167–213.
[17] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.
[18] R. TAMBS LYCHE, *Une formule d'iteration*, Bull. Soc. Math. France, 55 (1927), pp. 102–113.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0003 $01.00/0

A GENERALIZATION OF THE KREISS MATRIX THEOREM*

SHMUEL FRIEDLAND†

Abstract. Let $\mathcal{A}$ be a set of $n \times n$ complex matrices A which satisfy the condition $\|(I - zA)^{-1}\| \leqq K/(1-|z|)^{\alpha+1}$ for some $\alpha \geqq 0$ and all $|z| < 1$. Then it is shown here that there exists a constant $\rho(\alpha, n)$ such that $\|A^{\nu}\| \leqq K\rho(\alpha, n)\nu^{\alpha}$, $\nu = 0, 1, \cdots$. This forms a generalization of the Kreiss resolvent condition (where $\alpha = 0$).

1. Introduction.

In various instances one deals with iterative systems of equations

$$x^{(i+1)} = Ax^{(i)}, \qquad i = 0, 1, 2, \cdots. \tag{1.1}$$

Here $x^{(i)} \in C^n$, $A \in M_n$, where C^n is the set of n column complex vectors and M_n is the set of $n \times n$ complex matrices. Clearly

$$x^{(i)} = A^i x^{(0)}, \tag{1.2}$$

and thus in order to investigate the behavior of $x^{(i)}$ for large i one needs to study the powers A^i, $i = 0, 1, \cdots$. Let $\mathcal{A}$ be a set of $n \times n$ matrices. $\mathcal{A}$ is called an α-stable set if

$$\|A^{\nu}\| \leqq K\nu^{\alpha}, \qquad \nu = 0, 1, 2, \cdots. \tag{1.3}$$

Here α is a nonnegative number and $\|\cdot\|$ is a norm on M_n. The concept of stability of the numerical schemes for solutions of partial differential equations is intimately connected with the notion of stable sets. Consult for example Kreiss [1962] or Richtmyer and Morton [1967]. It seems that α-stable sets are related to the concept of weakly stable numerical schemes for partial differential equations. See Kreiss [1962] and Forsyth and Wasow [1960]. The stable sets were characterized completely by Kreiss [1962]. In this paper we generalize the Kreiss result to α-stable sets.

THEOREM 1. *Let α be a nonnegative number and $\mathcal{A}$ be a set of $n \times n$ complex valued matrices. Then the following two conditions are equivalent.*

(A) *There exists a constant $K(\geqq 1)$ such that for all $A \in \mathcal{A}$ (1.3) holds.*

(R) *There exists a constant $K(\geqq 1)$ such that for all $A \in \mathcal{A}$*

$$\|(I - zA)^{-1}\| \leqq K(1-|z|)^{-(\alpha+1)}, \qquad |z| < 1. \tag{1.4}$$

The implication (A) $\Rightarrow$ (R) is obvious. The implication (R) $\Rightarrow$ (A) is a consequence of Theorem 2 which estimates the Maclaurin coefficients of a certain family of rational functions in terms of the growth of their moduli. We were not able to give conditions analogous to the conditions (S) and (H) of Kreiss [1962].

2. Coefficient estimates for certain analyic functions.

Let D be a unit disk $|z| < 1$. Suppose that $f(z)$ is an analytic function in D. Consider the Maclaurin expansion of f,

$$f(z) = \sum_{\nu=0}^{\infty} a_{\nu} z^{\nu}, \qquad |z| < 1. \tag{2.1}$$

Suppose that

$$|a_{\nu}| \leqq K\nu^{\alpha}, \qquad \alpha = 0, 1, 2, \cdots \tag{2.2}$$

for $\alpha > -1$. It is a standard result in theory of special functions (e.g., Olver [1974,

* Received by the editors, September 15, 1980. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and appeared as Technical Summary Report 2108, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin.

† Institute of Mathematics, Hebrew University, Jerusalem, Israel.

p. 119]) that

$$\nu^{\alpha} \approx (-1)^{\nu}\binom{-(\alpha-1)}{\nu} = \frac{\Gamma(\alpha+\nu+1)}{\Gamma(\nu+1)\Gamma(\alpha+1)}. \tag{2.3}$$

Here two positive sequences $\{u_m\}$ and $\{v_m\}$ are called equivalent $u_m \approx v_m$ if

$$\lim_{m\to\infty} \frac{u_m}{v_m} = \beta, \qquad 0 < \beta < \infty.$$

Thus (2.3) implies

$$|f(z)| \leqq K\rho(\alpha)(1-|z|)^{-(\alpha+1)} \tag{2.4}$$

for some positive constant $\rho(\alpha)$ with $\alpha > -1$. Conversely, we have a weaker result.

LEMMA 1. *Let $f(z)$ be analytic in D. Assume that*

$$|f(z)| \leqq K(1-|z|)^{-\alpha}, \tag{2.5}$$

for some $\alpha > 0$ and all $|z| < 1$. Then

$$|a_\nu| \leqq K\left(1+\frac{\alpha}{\nu}\right)^{\nu}\left(\frac{\nu+\alpha}{\alpha}\right)^{\alpha} < Ke(\nu+1)^{\alpha}, \tag{2.6}$$

and this inequality is sharp.

Proof. As

$$a_\nu = (2\pi i)^{-1}\int_{|z|=r<1} f(z)z^{-(\nu+1)}\,dz, \tag{2.7}$$

we get

$$|a_\nu| \leqq [\max_{|z|=r}|f(z)|]r^{-\nu} \leqq K(1-r)^{-\alpha}r^{-\nu}. \tag{2.8}$$

Note that

$$\min_{0\leqq r\leqq 1}(1-r)^{-\alpha}r^{-\nu} = (1-r)^{-\alpha}r^{-\nu}|_{r=\nu/(\nu+\alpha)} = \left(1+\frac{\alpha}{\nu}\right)^{\nu}\left(\frac{\nu+\alpha}{\alpha}\right)^{\alpha}.$$

This establishes the first inequality in (2.6). To obtain the second inequality in (2.6) choose, in (2.8), $r = \nu/(\nu+1)$ and use the well-known fact that $(1+1/\nu)^{\nu} < e$. To see that (2.6) is sharp for each ν, consider the polynomial

$$p(z) = K\left(1+\frac{\alpha}{\nu}\right)^{\nu}\left(\frac{\nu+\alpha}{\alpha}\right)^{\nu} z^{\nu}. \tag{2.9}$$

Let B be a Banach space with a norm $\|\cdot\|$. Assume that $A : B \to B$ is a bounded linear operator. Suppose that the spectrum of A lies in the unit disk. Then expanding $(I-zA)^{-1}$ in power series

$$(I-zA)^{-1} = \sum_{\nu=0}^{\infty} z^{\nu}A^{\nu}, \tag{2.10}$$

we get

$$A^{\nu} = (2\pi i)^{-1}\int_{|z|=r<1} z^{-(\nu+1)}(I-zA)^{-1}\,dz. \tag{2.11}$$

Thus if

$$\|(I-zA)^{-1}\| \leqq K(1-|z|)^{-\alpha}, \qquad |z|<1, \tag{2.12}$$

then, applying the results of Lemma 1, we obtain

$$\|A^{\nu}\| < Ke(\nu+1)^{\alpha}. \tag{2.13}$$

□

It is an open problem whether the estimate (2.13) is sharp in some infinite dimensional Banach space. The following result enables one to improve the inequality (2.13) for matrices (i.e., B is finite dimensional).

THEOREM 2. *Consider all polynomials $p(z)$ and $q(z)$ of degrees m and n respectively such that the function $f(z)=p(z)/q(z)$ satisfies* (2.5). *Suppose that $\alpha \geqq 1$. Then there exists a positive constant $\rho(\alpha, m, n)$ such that*

$$|a_{\nu}| \leqq K\rho(\alpha, m, n)\nu^{(\alpha-1)}. \tag{2.14}$$

To prove this theorem we need the following lemma.

LEMMA 2. *Let $p(z)$ be a polynomial of degree m. Then there exists a constant $K(m)$ such that*

$$\max_{|z|=r} |p(z)| \leqq K(m)\left(\max_{|\theta|\leqq \pi/4} |p(re^{i\theta})|\right). \tag{2.15}$$

Proof. It is enough to consider the case $r=1$ with $p(z)$ of the form

$$p(z)=\prod_{i=1}^{m}(z-\zeta_i), \qquad |\zeta_1| \leqq |\zeta_2| \cdots \leqq |\zeta_m|.$$

For $m=1$, it suffices to chose $K(1)=5$. Let $m>1$. Define

$$K'(m)=\max_{0\leqq|\zeta_1|\leqq\cdots\leqq|\zeta_m|\leqq 3}\left(\frac{\max_{|z|=1}|p(z)|}{\max_{|\theta|\leqq\pi/4}|p(e^{i\theta})|}\right).$$

In the case where $|\zeta_m|>3$ let $q(z)=\prod_{i=1}^{m-1}(z-\zeta_i)$. Then

$$\max_{|z|=1}|p(z)| \leqq (|\zeta_m|+1)\max_{|z|=1}|q(z)|$$

$$\leqq 2(|\zeta_m|-1)K(m-1)\max_{|\theta|\leqq\pi/4}|q(e^{i\theta})| \leqq 2K(m-1)\max_{|\theta|\leqq\pi/4}|p(e^{i\theta})|.$$

Put

$$K(m)=\max\,(K'(m),\, 2K(m-1))$$

and the lemma follows. □

Proof of Theorem 2. Without loss of generality, we may assume that $p(z)$ and $q(z)$ do not have common zeros. Also, it is enough to consider the case $K=1$. The inequality (2.5) implies that we can choose q and p of the form

$$p(z)=z^{l}A\prod_{i=1}^{m-l}(1-z\omega_i), \qquad q(z)=\prod_{i=1}^{n}(1-z\zeta_i). \tag{2.16}$$

The inequality (2.5) yields $|\zeta_i| \leqq 1$, $i=1,\cdots,n$. Put

$$g(z)=\frac{A\prod_{i=1}^{m-l}(z-\omega_i)}{\prod_{i=1}^{n}(z-\zeta_i)}, \tag{2.17}$$

$$|g(z)| \leqq \frac{|z|^{m-n+\alpha}}{(|z|-1)^{\alpha}}, \qquad |z|>1. \tag{2.18}$$

Also,

$$g(z) = \sum_{\nu=0}^{\infty} a_{\nu} z^{-(\nu+n-m)}, \qquad |z|>1. \tag{2.19}$$

Note that

$$a_{\nu} = (2\pi i)^{-1} \int_{|z|=R>1} g(z) z^{(\nu+n-m-1)}\, dz.$$

Let $D_1, \cdots, D_p$ be p-mutually disjoint, open and bounded domains with the boundary $\Gamma_1, \cdots, \Gamma_p$ respectively. Assume that $\zeta_i \in \cup_{j=1}^{p} D_j$, $i = 1, \cdots, n$. Then we obtain

$$a_{\nu} = \sum_{j=1}^{p} (2\pi i)^{-1} \int_{\Gamma_j} g(z) z^{(\nu+n-m-1)}\, dz. \tag{2.20}$$

To obtain the estimate (2.14) we are going to choose the domains $D_1, \cdots, D_p$ according to the configuration of $\zeta_1, \cdots, \zeta_n$ and the value of ν. First we group the points $S_1, \cdots, S_s$, following Morton [1964]. Let ζ_{i_1} be one of the points with the largest modulus, $|\zeta_{i_1}| = 1-\delta_1 \geqq |\zeta_i|$, $i = 1, \cdots, n$. Then we form S_1 from all those points which can be joined to ζ_{i_1} by a chain of points, each link of which has length $\leqq \delta_1$.

In the same way S_2 is formed from the remaining points, and so on until all the points have been included in some S_β. For each S_β we denote by $1-\delta_\beta$ and $1-\varepsilon_\beta$ the modulus of the largest and the smallest $|\zeta_i|$, $\zeta_i \in S_\beta$. We rename $\zeta_1, \cdots, \zeta_n$ so that

$$0 \leqq \delta_1 \leqq \cdots \leqq \delta_s. \tag{2.21}$$

Consider any particular S_β and denote its members by λ_i, $i = 1, 2, \cdots, k$, where $1-\varepsilon_\beta \leqq |\lambda_i| \leqq 1-\delta_\beta$, $i = 1, \cdots, k$. Also denote the points not in S_β by μ_j, $j = 1, 2, \cdots, n-k$. We claim

$$\delta_\beta \leqq 1-|\lambda_i| \leqq k\delta_\beta, \qquad |\lambda_i - \lambda_j| \leqq (k-1)\delta_\beta, \qquad |\lambda_i - \mu_j| > \delta_\beta. \tag{2.22}$$

Indeed, the first two inequalities follow immediately from the assumption that there exists a chain of at most k points between λ_i and λ_j such that the distance of any link $\leqq \delta_\beta$. The last inequality is a consequence of μ_j not being in S_β. Let

$$h(z) = A \prod_{i=1}^{m-l} (z-\omega_i). \tag{2.23}$$

For $\lambda_t \in S_\beta$ put

$$\eta = (1+2\delta_\beta)\frac{\lambda_t}{|\lambda_t|}. \tag{2.24}$$

Then

$$h(z) = \sum_{j=0}^{m-l} h_j (z-\eta)^j. \tag{2.25}$$

We now estimate h_j. Let Γ be a circle $|z-\eta| = \delta_\beta$. Then

$$\begin{aligned} |z-\zeta_i| &\leqq |z-\lambda_t| + |\lambda_t - \zeta_i| \leqq |z-\eta| + |\eta - \lambda_t| + |\lambda_t - \zeta_i| \\ &= \delta_\beta + 1 + 2\delta_\beta - |\lambda_t| + |\lambda_t - \zeta_i| \leqq (k+3)\delta_\beta + |\lambda_t - \zeta_i|, \end{aligned}$$

where the last inequality follows from (2.22). In particular

$$|z-\lambda_j| \leqq 2(k+1)\delta_\beta,$$

in view of (2.22). Apply the Cauchy formula for h_j and use (2.18) to get

$$
\begin{aligned}
|h_j| &= (2\pi)^{-1}\left|\int_\Gamma h(z)\,dz(z-\eta)^{-(j+1)}\right| \\
&\leqq \delta_\beta^{-(j+\alpha)} 4^{m+\alpha} \prod_{i=1}^{n} [(k+3)\delta_\beta + |\lambda_t - \zeta_i|] \\
&\leqq [2(k+1)]^k 4^{m+\alpha}\delta_\beta^{-(j+\alpha)+k} \prod_{i=1}^{n-k} [(k+3)\delta_\beta + |\lambda_t - \mu_i|] \\
&\leqq [2(k+2)]^n 4^{m+\alpha}\delta_\beta^{-(j+\alpha)+k} \prod_{i=1}^{n-k} |\lambda_t - \mu_i|.
\end{aligned} \tag{2.26}
$$

We now consider the following three cases:

(i) $\delta_\beta \geqq 1/(2^{n+2}n\nu)$,

(ii) $\delta_\beta \leqq 1/(4n\nu)$,

(iii) neither (i) nor (ii) holds.

Here ν is a positive integer and $\nu \geqq m-n+\alpha$.

Case (i). Let C_i be a disk $|z-\zeta_i| < \delta_\beta/2$ for $\zeta_i \in S_\beta$. Then

$$
D = \bigcup_{i=1}^{n} C_i = \bigcup_{j=1}^{p} D_j, \tag{2.27}
$$

where each D_j contains a subset of some S_β and $D_j \cap D_k = \emptyset$ for $j \neq k$. Let Γ_j be the boundary of D_j. Then $l(\Gamma_j)$–the length of Γ_j–satisfies the inequality

$$
l(\Gamma_j) \leqq 2\pi n(D_j)\delta_\beta, \tag{2.28}
$$

where $n(D_j)$ is the number of points $\zeta_1, \cdots, \zeta_n$ in D_j. Let $z \in \Gamma_j$. Clearly, $z = \lambda_t + \rho$, $|\rho| = \delta_\beta/2$, $S_\beta = \{\lambda_1, \cdots, \lambda_k\}$. By the definition of D_j, $|z-\lambda_j| \geqq \delta_\beta/2$ for $1 \leqq j \leqq k$. Also

$$
|z-\mu_j| = |\lambda_t - \mu_j + \rho| \geqq |\lambda_t - \mu_j| - \frac{\delta_\beta}{2} \geqq \frac{1}{2}|\lambda_t - \mu_j|.
$$

Thus

$$
\left|\prod_{i=1}^{n} (z-\zeta_i)^{-1}\right| \leqq 2^n \delta_\beta^{-k} \prod_{j=1}^{n-k} |\lambda_t - \mu_j|^{-1}. \tag{2.29}
$$

Also for η of the form (2.24) we have

$$
|z-\eta| \leqq |z-\lambda_t| + |\lambda_t - \eta| \leqq \frac{\delta_\beta}{2} + 1 + 2\delta_\beta - |\lambda_t| < (k+3)\delta_\beta.
$$

Combine (2.25)–(2.26) with the above equality to deduce

$$
|h(z)| \leqq [2(k+2)]^{n+m} m 4^{m+\alpha}\delta_\beta^{k-\alpha} \prod_{i=1}^{n-k} |\lambda_t - \mu_i|. \tag{2.30}
$$

Finally we deduce

$$
|g(z)| \leqq [16(n+2)]^{n+m+\alpha}\delta_b^{-\alpha}, \qquad z \in \Gamma_j. \tag{2.31}
$$

Using the equality (2.20) and the inequalities (2.28), (2.31) for $\nu > m-n$, we get

$$
\begin{aligned}
|a_\nu| \leqq \sum_{j=1}^{p} (2\pi)^{-1} \int_{\Gamma_j} |g(z)||z|^{(\nu+m-n-1)}|dz| &\leqq n[16(n+2)]^{n+m+\alpha}(\min_{1\leqq\beta\leqq s} \delta_\beta)^{-\alpha+1} \\
&\leqq n^\alpha [16(n+2)]^{n+m+\alpha} 2^{(n+2)(\alpha-1)} \nu^{\alpha-1},
\end{aligned}
$$

where $\alpha \geqq 1$. Thus we have shown (2.14) ($K = 1$).

Case (ii). Let C_i be an open disk with center at $\zeta_i/|\zeta_i|$ and radius $1/2\nu$. Form D by (2.27). Assume that $z \in \Gamma_j$. So

$$z = \frac{\zeta_i}{|\zeta_i|} + \rho, \qquad |\rho| = \frac{1}{2\nu}. \tag{2.32}$$

We now estimate

$$K(\Gamma) = \max_{z\in\Gamma} |h(z)|, \qquad \Gamma = \left\{ z,\ z = \frac{\zeta_i}{|\zeta_i|}\left(1 + \frac{1}{2\nu} e^{i\theta}\right),\ |\theta| \leqq \frac{\pi}{4} \right\}.$$

According to (2.18)

$$K(\Gamma) \leqq e(4\nu)^{\alpha} \left[\max_{z\in\Gamma} \prod_{t=1}^{n} |z - \zeta_t| \right]$$

for $\nu \geqq m - n + \alpha$. Let $\eta_i = (1 + 1/2\nu)\zeta_i/|\zeta_i|$. Clearly, $\eta_i \in \Gamma_j$. We claim that for $z \in \Gamma$ or z of the form (2.32) which is in Γ_j, we have

$$\frac{|\eta_i - \zeta_t|}{5} \leqq |z - \zeta_t| \leqq 3|\eta_i - \zeta_t|.$$

Indeed it is easy to see that for such z the following inequalities hold

$$|z - \zeta_t| \geqq \frac{1}{4\nu}, \quad |\eta_i - \zeta_t| \geqq \frac{1}{2\nu}, \quad |z - \eta_i| \leqq \frac{1}{\nu}.$$

So,

$$|\eta_i - \zeta_t| \leqq |z - \eta_i| + |z - \zeta_t| \leqq 4|z - \zeta_t| + |z - \zeta_t| = 5|z - \zeta_t|,$$
$$|z - \zeta_t| \leqq |z - \eta_i| + |\eta_i - \zeta_t| \leqq 2|\eta_i - \zeta_t| + |\eta_i - \zeta_t| = 3|\eta_i - \zeta_t|.$$

Therefore

$$K(\Gamma) \leqq e3^n(4\nu)^{\alpha} \prod_{t=1}^{n} |\eta_i - \zeta_t|.$$

Let $z = \zeta_i/|\zeta_i| + \rho \in \Gamma_j$, $|\rho| = 1/2\nu$. Then by Lemma 2 and the above inequalities

$$|g(z)| \leqq \frac{K(\Gamma)K(m-l)}{\prod_{t=1}^{n} |z - \zeta_t|} \leqq K(m-l)(15)^n(4\nu)^{\alpha} e,$$

and

$$|g(z)z^{\nu+m-n-1}| \leqq K(m-l)(15)^n(4\nu)^{\alpha} e \left(1 + \frac{1}{2\nu}\right)^{\nu+m-n-1} \leqq K(m-l)(15)^n(4\nu)^{\alpha} e^2$$

for $\nu \geqq m - n + \alpha$. As the length of the boundary of D does not exceed $\pi n/\nu$, from (2.20), we get

$$|a_\nu| \leqq K(m-l)n(15)^n 4^{\alpha} e^2 \nu^{\alpha-1}.$$

Case (iii). In this case we claim that there exists $1 < \gamma < s$ such that

$$\delta_{\beta+1} < \frac{1}{2^{n+2}\nu n} + \max_{0\leqq j\leqq\beta} \varepsilon_j, \qquad \beta = 0, \cdots, \gamma - 1, \qquad \varepsilon_0 = 0 \tag{2.33}$$

and

$$\delta_{\gamma+1} \geqq \frac{1}{2^{n+2}\nu n} + \max_{0 \leqq j \leqq \gamma} \varepsilon_j, \tag{2.34}$$

otherwise either (i) or (ii) hold. (Note the inequality (2.21).) Put

$$r = (\max_{0 \leqq j \leqq \gamma} \varepsilon_j) + \frac{1}{2^{n+3}\nu n}. \tag{2.35}$$

It is not difficult to show that $r \leqq 1/2\nu n$. Let $\zeta_i \in S_\beta$. For $\beta \leqq \gamma$, denote by C_i a disk with center at $\zeta_i/|\zeta_i|$ and radius r. For $\beta > \gamma$, let C_i be a disk with center at ζ_i and radius $\delta_\beta/2$. As before, define D by (2.27). Now estimate a_ν from (2.20), using the arguments of the Cases (i) and (ii), in accordance with $\beta > \gamma$ or $\beta \leqq \gamma$, to deduce (2.14). This concludes the proof of Theorem 2. □

Remark 1. A special case of Theorem 2, namely, $\alpha = 1$ and $m = n - 1$, was established in Morton [1964].

Proof of Theorem 1.

(A) $\Rightarrow$ (R). This follows immediately from (2.3).

(R) $\Rightarrow$ (A). Let $(I - zA)^{-1} = (f_{ij}(z))_1^n$.

Then $f_{ij}(z) = p_{ij}(z)/q_{ij}(z)$, where the degrees of p_{ij} and q_{ij} are $n-1$ and n, respectively. Now (1.3) follows from Theorem 2.

Acknowledgment. I would like to thank S. Parter for stimulating discussions we had together.

Note added in proof. E. Tadmov in a recent paper [*The equivalence of L_2-stability, the resolvent condition and H-stability*, preprint, California Institute of Technology, 1981] used Laptev's arguments [*Conditions for the uniform well-posedness of the Cauchy problem for systems of equations*, Soviet Math. Dokl., 16 (1975), pp. 65–69] to prove that the inequality (1.4) implies

$$\|A^{\nu-1}\| \leqq \frac{32enK}{\pi} \nu^\alpha.$$

REFERENCES

G. E. FORSYTH AND W. R. WASOW [1960], *Finite Difference Methods for Partial Differential Equations*, John Wiley, New York.

H. O. KREISS [1962], *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren*, BIT 2, pp. 153–181.

K. W. MORTON [1964], *On a matrix theorem due to H. O. Kreiss*, Comm. Pure Appl. Math., 17, pp. 375–379.

F. W. J. OLVER [1974], *Asymptotic and Special Functions*, Academic Press, New York.

R. D. RICHTMYER AND K. W. MORTON [1967], *Difference Methods for Initial Value Problems*, (second edition), Interscience, New York.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0004 $01.00/0

ON THE SOLVABILITY OF CERTAIN SYSTEMS OF LINEAR DIFFERENCE EQUATIONS*

A. S. CAVARETTA, JR.,† W. DAHMEN,‡ C. A. MICCHELLI§ AND P W. SMITH‖

Abstract. For a certain class of block Toeplitz matrices, we identify the smallest sector containing the zeros of the determinant for the corresponding symbol.

0. We have been concerned with the inversion of infinite linear systems when the coefficient matrix enjoys certain positivity assumptions, for instance, multiple positivity in the sense of Fekete. Such systems arise in many contexts, one notable case being cardinal spline interpolation. When the cofficient matrix has the Toeplitz structure $a_{i+1,j+1} = a_{ij}$, the inversion is well understood. Our purpose here is to analyze the inversion of banded totally positive systems when the Toeplitz structure is replaced by the weaker hypothesis that for some fixed natural number N, $a_{i+N,j+N} = a_{ij}$. This leads us to a question concerning the location of the zeros of a certain determinant. When N and the multiple positivity is prescribed, we obtain the smallest sector containing these zeros; our theorem extends a result of Schoenberg [2] for the Toeplitz case.

1. Block Toeplitz systems. Fix any summable bi-infinite sequence $\{a_i\}$. Associated with such a sequence is the discrete convolution given by

$$\sum_{j=-\infty}^{\infty} a_{i-j}\, x_j = y_i, \qquad -\infty < i < \infty. \tag{1.1}$$

The bounded inversion of such systems is a fundamental and well understood problem. For example, we encounter (1.1) in cardinal spline interpolation; in this case all but finitely many of the a_i vanish so that the associated Toeplitz matrix is banded. If we investigate the cardinal spline interpolation problem further and allow for general periodically distributed nodes, (1.1) is replaced with

$$\begin{aligned} &\sum_{j=-\infty}^{\infty} a_{i-j}^{(n)} x_j = y_i, \qquad i \equiv n \bmod N, \quad -\infty < i < \infty, \\ &\sum_{i=-\infty}^{\infty} |a_i^{(n)}| < \infty, \qquad n = 1, \cdots, N. \end{aligned} \tag{1.2}$$

Clearly (1.2) reduces to (1.1) when $N = 1$. System (1.2) is easily characterized in terms of the coefficient matrix $A = (a_{ij})$ for the general system $\sum_{-\infty}^{\infty} a_{ij} x_j = y_i$: we obtain (1.2) if $a_{ij} = a_{i-j}^{(i)}$, $1 \leqq i \leqq N$, $-\infty < j < \infty$ and $a_{i+N,j+N} = a_{i,j}$. Such matrices will be called *block Toeplitz* (of order N).

It is convenient to rewrite the system (1.2) in matrix notation. In the obvious way we group the sequences $\{x_j\}$ and $\{y_i\}$ into sequences $\{\bar{x}_i\}$, $\{\bar{y}_i\}$ of N-tuples by

* Received by the editors August 4, 1980, and in revised form February 12, 1981.

† Department of Mathematics, Kent State University, Kent, Ohio 44242.

‡ Universität Bonn, Institut für Angewandte Mathematik, Bonn, West Germany.

§ Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

‖ Old Dominion University, Norfolk, Virginia 23508. The research of this author was partially supported by the U.S. Army Research Office under Grant #DAHC04-75-0816.

setting for each integer j

$$\bar{x}_j = (x_{jN}, \cdots, x_{jN+N-1}),$$
$$\bar{y}_j = (y_{jN}, \cdots, y_{jN+N-1}).$$

For each integer k define

$$(A_k)_{ij} = a^{(i)}_{i+k,N-j}, \qquad 1 \leqq i,j \leqq N.$$

With this notation, (1.2) takes the suggestive form

$$\sum_{j=-\infty}^{\infty} A_{i-j}\, \bar{x}_j = \bar{y}_i. \tag{1.3}$$

The inversion of such block Toeplitz systems has been previously investigated by several authors; see, for example, I. I. Hirschman, Jr. [1]. From general principles, we knows that (1.3) is invertible provided its *symbol*

$$A(z) = \sum_{n=-\infty}^{\infty} A_n z^n \tag{1.4}$$

is invertible for $|z| = 1$. This in turn requires an investigation of the zeros of $\det A(z)$.

Set

$$A = \begin{bmatrix} \cdots & \cdot & \cdot & \cdot & \cdots \\ \cdots & A_0 & A_1 & A_2 & \cdots \\ \cdots & A_{-1} & A_0 & A_1 & \cdots \\ \cdots & A_{-2} & A_{-1} & A_0 & \cdots \\ \cdots & \cdot & \cdot & \cdot & \cdots \end{bmatrix} \tag{1.5}$$

so that A is the coefficient matrix for the block Toeplitz system (1.3). We say that the matrix sequence $\{A_n\}$ is *totally positive* provided the matrix A is totally positive, i.e., has no minor which is negative. More generally, given a natural integer k, $\{A_n\}$ is said to be *k-positive* provided the matrix A of (1.5) has no negative minor of order $\leqq k$. In § 2, we locate the zeros of $\det A(z)$ when the matrix A is k-positive and strictly banded. Our methods are refinements and extensions of Schoenberg's elegant analysis for the Toeplitz case [2]. In § 3 we extend our results to the case A totally positive and derive consequences for the inversion of A. Section 4 finishes the paper with applications to cardinal spline interpolation.

2. Multiply positive matrix sequences. Let the sequences of $N \times N$ matrices

$$\cdots O, O, A_0, A_1, \cdots, A_q, O, O, \cdots, \det A_0 \neq 0 \tag{2.1}$$

be k-positive, as defined in § 1. We will assume that the matrix A is p-banded, i.e., $a_{ij} = 0$ if $j < i$ or if $j > i+p$. So in (2.1) A_0 is assumed to be nonsingular and upper triangular, and q is the smallest integer greater than or equal to p/N. Then the symbol (1.4) becomes

$$A(z) = A_0 + A_1 z + \cdots + A_q z^q \tag{2.2}$$

and we put

$$P(z) = \det A(z). \tag{2.3}$$

THEOREM 1. *Assume the sequence* (2.1) *is k-positive and the matrix* (1.5) *is p-banded, If N is even, then all zeros of* $P(z)$ *lie in the sector*

$$|\arg z| \leqq \pi N\left(\frac{p-1}{p+k-1}\right).$$

If N is odd, than all zeros of $P(z)$ *lie in the sector*

$$|\arg z - \pi| \leqq \pi N\left(\frac{p-1}{p+k-1}\right).$$

Note that when $N = 1$, our result reduced to that of Schoenberg. The result only has content when the right-hand side of the inequalities is less than π, as is certainly the case when k is sufficiently large. Using the central Gaussian coefficients [2], we will show that these two inequalities are sharp.

Our proof of Theorem 1 depends on the known result that a linear transformation

$$y_i = \sum_{j=1}^{n} m_{ij} x_j, \qquad i = 1, \cdots, m$$

is variation diminishing whenever $M = (m_{ij})$ is totally positive (see Schoenberg [2]). More precisely, if we denote by $v(\bar{x})$ the number of variations of signs in $\bar{x} = (x_1, \cdots, x_n)$ and by $v(\bar{y})$ the corresponding number for the sequence $y_1, \cdots, y_m$, then the inequality

$$v(\bar{y}) \leqq v(\bar{x})$$

always holds provided M is totally positive.

Proof of Theorem 1. Suppose $z = \alpha$ is a zero for the polynomial in (2.3). Clearly, since A_0 is nonsingular, $\alpha \neq 0$ and we may assume $\alpha = \rho e^{i\theta}$, $\rho > 0$ and $0 \leqq \theta \leqq \pi$. Since $A(\alpha)$ is singular, there exists some nonzero vector $\bar{x} = (x_1, \cdots, x_N)$ in the null space of $A(\alpha)$. It follows that $(\cdots \alpha^{-2}\bar{x}, \alpha^{-1}\bar{x}, \bar{x}, \alpha\bar{x}, \alpha^2\bar{x}, \cdots)$ is in the null space of the matrix A of (1.5). Indeed, up to certain common factors of α^l, each equation in the infinite system of (1.5) is equivalent to one of the equations

$$A(\alpha) \cdot \bar{x} = 0. \tag{2.4}$$

From the matrix A of (1.5), which we have assumed to be k-positive, we now identify N different totally positive submatrices. Each of these is formed from k consecutive rows and $p+k$ consecutive columns of A. The N diagonal elements of A_0 supply the N different first row, first column positions for each of these $k+(p+k)$ matrices. Label these matrices $B_1, \cdots, B_N$. Each of these matrices is k-positive, hence totally positive. Note also that since A_0 is invertible each B_i is of full rank k; indeed, the minor of B_i which consists of the first k columns of B_i is upper triangular with nonzero diagonal entries.

Using the B_i as building blocks, we will now construct an auxiliary totally positive, full rank matrix M. Let h be a fixed positive integer still to be disposed of later; set $n = hN$. We set up a linear transformation whose matrix M has the form

$$M = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_{i_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & B_{i_n} \end{bmatrix}. \tag{2.5}$$

The schematic form representing M in (2.5) fails to reveal one essential feature of this

matrix: the last element of each $B_{i_{j-1}}$ is immediately above and in the same column as the leading element of the next B_{i_j}. Thus M has nk rows and $n(p+k)-(n-1)=n(p+k-1)+1$ columns. This overlapping of the columns in the structure of M means that the last equation of $B_{i_{j-1}}$ and the first equation of B_{i_j} share a common variable. So in order to make the equations determined by M consistent with those of (2.4), we determine the i_j recursively by

$$i_j \equiv i_{j-1}+p+k-1 \qquad (\mathrm{mod}\ N). \tag{2.6}$$

Now for $l=h(p+k-1)$ set

$$\bar{y}=(\bar{x}, \alpha\bar{x}, \cdots, \alpha^{l-1}\bar{x}, \alpha^{l}x_1).$$

Then it is easily seen from (2.4), (2.5) and (2.6) that $M\bar{y}=0$. Setting $\bar{y}_1=\mathrm{Im}\,\bar{y}$, $\bar{y}_2=\mathrm{Re}\,\bar{y}$, we have $M\bar{y}_1=M\bar{y}_2=0$. In the following argument, we assume the real vector $\bar{y}_1\neq 0$ (otherwise we would use $\bar{y}_2$).

The matrix M has two important properties for us.

1. M is totally positive. This follows as each B_i is totally positive and the compounding used to construct M preserves total positivity (see Schoenberg [2]).

2. M has full rank. Indeed the minor obtained from M by selecting all columns corresponding to the first k columns of each B_i is upper triangular with all nonzero diagonal elements.

As M has full rank, we can select $\bar{v}$ such that $M\bar{v}=\bar{e}$, where $e_i=(-1)^i$. Then, for any $\varepsilon>0$, $M(\bar{y}_1+\varepsilon\bar{v})=\varepsilon\bar{e}$, and so by that variation diminishing property of the linear transformation induced by M, we can conclude that

$$v(\bar{y}_1+\varepsilon\bar{v})\geqq nk-1. \tag{2.7}$$

An upper bound on $v(\bar{y}_1+\varepsilon\bar{v})$ is obviously $n(p+k-1)$.

To improve on this upper bound, pick i such that $\sin\theta_i\neq 0$.
The vector $\bar{y}_1$ we write as

$$\begin{aligned}[\rho_1\sin\theta_1, \cdots, \rho_N\sin\theta_N, \rho\rho_1\sin(\theta_1+\theta), \cdots, \rho\rho_N\sin(\theta_N+\theta),\\ \rho^2\rho_1\sin(\theta_1+2\theta), \cdots, \rho^{l-1}\rho_N\sin(\theta_N+l\theta-\theta), \rho^{l}\rho_1\sin(\theta_1+l\theta)].\end{aligned} \tag{2.8}$$

In the case N even and $\theta\neq 0$, a sign change between $\sin(\theta_i+(j-1)\theta)$ and $\sin(\theta_i+j\theta)$ would allow us to reduce by one our upper bound for $v(\bar{y}_1+\varepsilon\bar{v})$. It follows that

$$\begin{aligned}v(\bar{y}_1+\varepsilon\bar{v}) &\leqq n(p+k-1)-v(\sin\theta_i, \sin(\theta_i+\theta), \cdots, \sin(\theta_i+l\theta-\theta))\\ &= n(p+k-1)-\left(\left[\frac{(l-1)\theta}{\pi}\right]+\sigma\right), \qquad \sigma=-1 \text{ or } 0.\end{aligned} \tag{2.9}$$

Combining (2.9) with (2.7), dividing by l and simplifying yields

$$\frac{1}{l}(nk-1)\leqq N-\frac{1}{l}\left(\left[\frac{(l-1)\theta}{\pi}\right]+\sigma\right).$$

Letting l (hence n) tend to infinity, we obtain

$$\theta\leqq\pi N\left(\frac{p-1}{p+k-1}\right).$$

This settles the case N even.

For N odd we argue similarly on the basis that *no* sign change between $\sin(\theta_i+(j-1)\theta)$ and $\sin(\theta_i+j\theta)$ will reduce by one our upper bound on $v(\bar{y}_1+\varepsilon\bar{v})$. The

resulting inequality is

$$nk-1 \leqq n(p+k-1)-l+1+\left[\frac{(l-1)\theta}{\pi}\right]+\sigma.$$

Dividing through by l which we then let tend to infinity yields, after simplification,

$$\pi-\theta \leqq \pi N\left(\frac{p-1}{p+k-1}\right).$$

This completes the proof of Theorem 1.

A few remarks will suffice to show that the factor $N(p-1)/(p+k-1)$ of π which appears on the right-hand side of the inequalities in Theorem 1 cannot be replaced by any smaller quantity. For the case $N=1$, Schoenberg showed that his inequality is best possible by introducing the *central Gaussian coefficients*, defined by

$$\begin{Bmatrix} p \\ \nu \end{Bmatrix} = \frac{\sin p\theta \sin (p-1)\theta \cdots \sin (p-\nu+1)\theta}{\sin \theta \sin 2\theta \cdots \sin \nu\theta} \quad \text{if } 0<\nu \leqq p,$$

$$\begin{Bmatrix} p \\ 0 \end{Bmatrix} = 1.$$

As Schoenberg shows, the sequence

$$\cdots 0, 0, \begin{Bmatrix} p \\ 0 \end{Bmatrix}, \begin{Bmatrix} p \\ 1 \end{Bmatrix}, \cdots, \begin{Bmatrix} p \\ p \end{Bmatrix}, 0, 0, \cdots \tag{2.10}$$

is k times positive if θ is in the range $0 \leqq \theta \leqq \pi/(p+k-1)$, and if $\theta = \pi/(p+k-1)$ then the symbol

$$q(z) = \sum_{\nu=0}^{p} \begin{Bmatrix} p \\ \nu \end{Bmatrix} z^{\nu} \tag{2.11}$$

has zeros at

$$z_1 = \exp\left\{\frac{k\pi i}{p+k-1}\right\}$$

and

$$z_2 = \exp\left\{\left(\frac{(p-1)\pi}{p+k-1}+\tau\right)i\right\}.$$

This sequence provides a Toeplitz matrix which we view as N-blocked. We wish to compute its symbol.

The symbol $A(z)$ of (1.4) can be calculated in terms of generating functions of the rows of A:

$$f_i(\tau) = \sum_{\nu=-\infty}^{\infty} a_{i\nu}\tau^{\nu}, \qquad i=0, \cdots, N-1.$$

Put $z=\tau^N$ and let ω be a primitive Nth root of unity. Then

$$A(z) = \begin{bmatrix} f_0(\tau) & f_0(\omega\tau) & \cdots & f_0(\omega^{N-1}\tau) \\ f_1(\tau) & f_1(\omega\tau) & \cdots & f_1(\omega^{N-1}\tau) \\ \vdots & \vdots & & \vdots \\ f_{N-1}(\tau) & f_{N-1}(\omega\tau) & \cdots & f_{N-1}(\omega^{N-1}\tau) \end{bmatrix} \cdot V^{-1}, \tag{2.12}$$

where

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \tau & \omega\tau & \cdots & \omega^{N-1}\tau \\ \vdots & \vdots & & \vdots \\ \tau^{N-1} & (\omega\tau)^{N-1} & \cdots & (\omega^{N-1}\tau)^{N-1} \end{bmatrix}.$$

If we specialize to the Toeplitz matrix determined by the sequence (2.10), (2.12) becomes

$$A(z) = V \cdot \begin{bmatrix} q(\tau) & 0 & \cdots & 0 \\ 0 & q(\omega\tau) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & q(\omega^{N-1}\tau) \end{bmatrix} V^{-1}.$$

It follows that if $q(\tau) = 0$, then $\det A(\tau^N) = 0$. From this we easily conclude that if $\theta = \pi/(p+k-1)$ the sequence (2.10) provides a Toeplitz matrix which, when viewed as block Toeplitz, has a singular symbol $A(z)$ when

$$z = \exp\left\{\pm\frac{N(p-1)\pi}{p+k-1}\right\} \qquad (N \text{ even})$$

or

$$z = \exp\left\{\pi \pm \frac{N(p-1)\pi}{p+k-1}\right\} \qquad (N \text{ odd}).$$

So we see that the zero sector given in Theorem 1 for the zeros of $P(z) = \det A(z)$ cannot be made smaller.

3. Totally positive matrix sequences. Suppose the sequence

$$\cdots O, O, A_0, A_1, \cdots, A_q, O, O, \cdots \tag{3.1}$$

is totally positive. Then the matrix A of (1.5) has only nonnegative minors and the result of Theorem 1 applies for every k provided (3.1) corresponds to a p-banded system and $\det A_0 \neq 0$. We conclude that $P(z) = \det A(z)$ has only real zeros with signs $(-1)^N$. Actually more can be said:

THEOREM 2. *Suppose $A(z)$ is the symbol of the totally positive sequence* (3.1). *Let $\bar{x} = (x_1, \cdots, x_N)$ satisfy $A(\alpha)\cdot\bar{x} = \bar{0}$, $\bar{x} \neq \bar{0}$. Then, provided* $\det A(z) \not\equiv 0$,

(i) $(-1)^N \cdot \alpha \geqq 0$

(ii) $x_j \cdot x_{j+1} \leqq 0, j = 1, \cdots, N-1$.

We remark that in Theorem 2 there is no need to assume any special band structure on the matrix A generated by (3.1).

Proof. We use the same arguments of Theorem 1, with the simplification that the matrix M can be constructed from a single $nN \times (n+p)N$ block of A. Set

$$\bar{y} = (\bar{x}, \alpha\bar{x}, \alpha^2\bar{x}, \cdots, \alpha^{n+p-1}\bar{x}). \tag{3.2}$$

Then $M\bar{y} = \bar{0}$. As we assume $\det A(z) \not\equiv 0$, we may choose β such that $(-1)^N\beta > 0$ and $\det A(\beta) \neq 0$. Fix $\bar{e} = (e_\nu)_{\nu=1}^N$, $e_\nu = (-1)^\nu$ and then define $\bar{w}$ by $A(\beta)\bar{w} = \bar{e}$. Let $\bar{v} = (\bar{w}, \beta\bar{w}, \cdots, \beta^{n+q-1}\bar{w})$. Then, for any $\varepsilon > 0$, $M(\bar{y} + \varepsilon\bar{v})$ has a full complement of sign changes. If we now use the variation diminishing arguments as before we obtain

$$nN-1 \leqq v(\bar{y}). \tag{3.3}$$

From (3.3) we obtain (i) arguing exactly as in Theorem 1. Once the sign of α is determined, (ii) follows, since (3.3) holds for n arbitrarily large.

COROLLARY 1. *For α as in Theorem 2, the null space of $A(\alpha)$ is one-dimensional.*

Proof. Indeed, any solution $A(\alpha)\cdot \bar{x} = \bar{0}$ must satisfy (ii) and from this the assertion easily follows.

COROLLARY 2. *A polynomial $P(z)$ equals* $\det A(z)$ *for some totally positive sequence* (3.1) *with* $\det A_0 \neq 0$ *if and only if*

$$P(z) = p(\tau)p(\omega\tau)\cdots p(\omega^{N-1}\tau),$$

where $p(\tau)$ has only negative zeros, $z=\tau^N$, and ω is a primitive Nth root of unity.

Proof. The fact that P is $\det A(z)$ for some totally positive block Toeplitz matrix follows from (2.13). The converse follows from (i) of Theorem 2. Indeed,

$$P(z) = a\prod_{j=1}^{s}(1+(-1)^{N+1}\gamma_j z),$$

where a, $\gamma_j > 0$. Then the polynomial

$$p(\tau) = a^{1/N}\prod_{j=1}^{s}(1+\gamma_j^{1/N}\tau)$$

gives the desired factorization of $P(z)$.

A simple but instructive example of Theorem 2 is the symmetric Jacobi matrix generated by $A_0 = \left[\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right]$ and $A_1 = \left[\begin{smallmatrix} c & 0 \\ d & a \end{smallmatrix}\right]$:

$$A = \begin{bmatrix} \cdot & \cdot & \cdot & & & & & & \\ \cdot & 0 & a & b & c & 0 & \cdot & \cdot & \\ \cdot & \cdot & 0 & c & d & a & 0 & \cdot & \\ & & & 0 & a & b & c & \cdot & \cdot \\ & & & & & & \cdot & \cdot & \cdot \end{bmatrix}.$$

Using recurrence relations for the principal minors of A, one finds that A is totally postive if and only if

$$bd - a^2 - b^2 \geqq 2ac.$$

As for the symbol

$$A(z) = \begin{bmatrix} a+cz & b \\ dz & c+az \end{bmatrix}$$

its determinant is a quadratic whose discriminant is positive precisely when a, b, c, and d satisfy the above inequality. Thus $\det A(z)$ has only real zeros if A is totally positive. That these zeros are in fact positive is a consequence of Descartes' rule of signs.

One application of Theorem 2 is the following:

THEOREM 3. *Let A of* (1.5) *be determined by the totally positive sequence* (3.1). *Then the following are equivalent*:

(1) *$A\bar{x} = \bar{y}$ is invertible on l^∞.*
(2) $\det A((-1)^N) \neq 0$.
(3) *For $\bar{e} = \{(-1)^\nu\}_{\nu=-\infty}^{+\infty}$, the system $A\bar{x} = \bar{e}$ has a solution $\bar{x}_0 \in l^\infty$.*

Remark. Recently, C. de Boor has proved that for *arbitrary* strictly banded totally positive matrices A 1) and 3) are equivalent if $\bar{x}_0$ in 3) is assumed to be unique.

Proof. From general principles the block Toeplitz system, $A\bar{x} = \bar{y}$, is invertible if its symbol $A(z)$ is invertible for $|z| = 1$. Since by Theorem 2, the only possible zero of modulus one for $\det A(z)$ is $(-1)^N$, 2) guarantees that $A(z)$ is invertible for $|z| = 1$. Thus 2) implies 1).

Clearly, 1) implies 3); we need only show 3) implies 2). This implication depends on Theorem 2 applied to A^T which is itself totally positive. As is easily observed, $A^T(z) = (A(z^{-1}))^T$. Assume 2) fails so that $A((-1)^N)$ is singular. Then $A^T((-1)^N)$ is also singular and, according to Theorem 2, there exists $\bar{y} = (y_1, \cdots, y_N)$ such that $A^T((-1)^N)\bar{y} = \bar{0}$, and $y_j y_{j+1} \leqq 0$, $j = 1, \cdots, N-1$. By 3), there is an $\bar{x} \in l^\infty$ satisfying $A\bar{x} = \bar{e}$. Let S be the forward shift operator on l^∞ (i.e., $(S\bar{x})_i = x_{i-1}$) and define

$$\bar{w} = \lim_{L\to\infty} \frac{1}{2L+1} \sum_{j=-L}^{L} (-1)^{Nj} S^{Nj} \bar{x};$$

here the limit may have to be taken over some suitable subsequence as L tends to infinity. It is easy to see that $S_{\bar{w}}^N = (-1)^N \bar{w}$ and that $A\bar{w} = \bar{e}$. Setting $\bar{z} = (w_1, \cdots, w_N)$ we see that $A((-1)^N)\bar{z} = (-1, 1, \cdots, (-1)^N)$, and this yields the contradiction

$$0 = (A^T((-1)^N)\bar{y}, \bar{z}) = (\bar{y}, A((-1)^N)\bar{z}) \neq 0.$$

Hence, $A((-1)^N)$ must be nonsingular.

4. Cardinal spline interpolation. Following Schoenberg [3], we denote by $M(x)$ the forward B-spline of order m based on the integer knots $0, 1, \cdots, m$. The cardinal spline functions consist of all linear combinations $f(x) = \sum_{\nu=-\infty}^{\infty} c_\nu M(x-\nu)$. For a fixed positive integer N, we choose N nodes

$$0 \leqq \alpha_1 < \cdots < \alpha_N < N \tag{4.1}$$

and extend this collection periodically by

$$\alpha_{\nu+N} = \alpha_\nu, \qquad \nu \in Z. \tag{4.2}$$

Putting $a_{ij} = M(\alpha_i - j)$ we obtain a banded, block Toeplitz totally positive matrix A to which we apply Theorem 3. If we assume that the symbol is not identically zero (which amounts to certain interlacing conditions between the integer knots and the α_ν) we obtain

THEOREM 4. *The cardinal interpolation problem $f(a_\nu) = y_\nu$, $\nu \in Z$, is solvable for any l^∞ sequence $\{y_\nu\}$ at the nodes* (4.1), (4.2) *if and only if either of the following conditions is satisfied*:

(i) *There exists no nontrivial null spline $f(x)$, $f(\alpha_\nu) = 0$, $\nu \in Z$, satisfying $f(x+N) = (-1)^N f(x)$.*

(ii) *There exists a bounded spline function f such that $f(\alpha_\nu) = (-1)^\nu$ for all integers ν,*

$$\det(\Phi_m(\alpha_j, \omega^l)) \neq 0,$$

where ω is a primitive Nth root of unity and $\Phi_m(x; t)$ is the exponential Euler spline of [3].

Proof. (i) is a direct translation of (2) in Theorem 3. Moreover, (ii) is equivalent to (i) since the N Euler splines $\Phi_m(x; \omega^l)$, $l = 0, \cdots, N-1$, provide a basis for the N-periodic (or anti-periodic) cardinal splines.

REFERENCES

[1] I. I. HIRSCHMAN, JR., *Matrix-valued Toeplitz operators*, Duke Math. J., 34 (1967), pp. 403–415.

[2] I. J. SCHOENBERG, *On the zeros of the generating functions of multiply positive sequences and functions*, Ann. Math., 62 (1955), pp. 447–470.

[3] I. J. SCHOENBERG, *Cardinal Spline Interpolation*, CBMS Regional Conference Series in Applied Mathematics, 12, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0005 $01,00/0

THE PERRON CONDITION FOR DIFFERENTIAL-DIFFERENCE EQUATIONS IN A HILBERT SPACE*

RICHARD DATKO†

Abstract. The Perron condition, or bounded input-bounded output criterion for linear systems, is extended to a class of linear differential-difference equations in a Hilbert space.

1. Introduction. In this paper, a version of the Perron condition is developed for differential-difference equations in a Hilbert space. In engineering parlance this condition is frequently referred to as the bounded input-bounded output criterion for linear systems (see, e.g., [2]). The condition was originally given by O. Perron [9] for linear systems in R^n and has since been extended to a variety of problems (see, e.g., [1], [3], [4] and [6]). For linear systems of the form

$$\dot{x}(t) = A(t)x(t) + f(t), \tag{1.1}$$

where suitable restrictions are placed on $A(t)$, and $x(t)$ is in X, a Banach space, one statement of the Perron condition is as follows (see, e.g., [4]). Consider the system

$$\dot{y}(t) = A(t)y(t), \tag{1.2}$$

Then (1.2) is uniformly exponentially stable if and only if all solutions of (1.1), with initial values $|x(t_0)| = 0$ and forcing term f such that $|f(\cdot)| \in L_q$, $1 \leqq q \leqq \infty$, satisfy an inequality of the form

$$\int_{t_0}^{\infty} |x(t, t_0, 0, f)|^p \, dt \leqq M(t) < \infty, \tag{1.3}$$

where $M(f)$ does not depend on t_0 and $1 < p < \infty$.

The main purpose of this paper is to obtain a similar condition for systems defined on a real Hilbert space H which may, with a certain license, be written in the form

$$\frac{d}{dt}\left[y(t) - \sum_{j=1}^{m} B_j(t)y(t-h_j) \right] = A(t)y(t) + \sum_{j=1}^{m} A_j(t)y(t-h_j) + f(t). \tag{1.4}$$

In (1.4), $0 < h_1 < \cdots < h_m = h$, $\{A_j(t)\}$ and $\{B_j(t)\}$ are uniformly bounded linear mappings from H into itself defined for all $t \geqq 0$, and $A(t)$ is a (possibly unbounded) linear mapping from a dense set D in H into H and satisfies certain other conditions. Actually, we consider an integrated or weak version of (1.4) and show in Theorem 3.2 that the homogeneous equation related to the weakened form of (1.4) is uniformly exponentially stable if and only if the weak solutions of (1.4), with zero initial data, satisfy an inequality of the form

$$|y(t, t_0, 0, f)| \leqq M(f) \tag{1.5}$$

for all essentially bounded strongly measurable $f : [0, \infty) \to H$. To obtain this condition it was found expedient to make an assumption on the difference equation

$$Z(t) - \sum_{j=1}^{m} B_j(t)Z(t-h_j) = 0 \tag{1.6}$$

which is similar to the D-stable condition of Cruz and Hale (see, e.g., [7, Chap. 12]).

* Received by the editors December 20, 1978, and in final revised form February 11, 1981.
† Department of Mathematics, Georgetown University, Washington, DC 20057.

Using this assumption, we transform the original problem into one of the form

$$\frac{d}{dt}(x(t)) = A(t)x(t) + \sum_{j=1}^{\infty} H_j(t)x(t-w_j) + f(t), \tag{1.7}$$

where $h_1 = w_1 < w_2 \cdots$ and $\lim_{n\to\infty} w_n = \infty$, and establish the Perron condition (Theorem 3.1) for the system (1.7). Then, using results from [5], we apply Theorem 3.1 to obtain the main result, Theorem 3.2, for (1.4). Section 2 consists of preliminaries and the statement of results from [5] which are necessary to the development of § 3.

A few general comments will be made on the methodology used in this paper. First of all, the term *measurable* is always used in the sense of strongly measurable (see, e.g., [8]). Moreover, whenever the order of integration and summation or the order of integration in multiple integrals is changed it will not be explicitly justified since the nature of the systems under consideration permits these operations and their verification can be done in a straightforward manner. More importantly, the results in this paper can be easily extended to Banach spaces, since the major proofs do not intrinsically depend on the Hilbert space structure. The reason this is not done is that heavy reliance is placed on [5, § 3], which is developed for a Hilbert space, and it is felt that extension of the results of that section to Banach spaces would obfuscate a basically simple treatment of the Perron condition for delay differential equations.

The core of this paper is Theorem 3.1. In § 2, it is shown that there exists a piecewise strongly continuous family of mappings, $S(t, \sigma)$, from $[0, \infty) \times [0, \infty)$ into the space of continuous linear mappings on H such that solutions of (1.7), with initial data zero, can be written in the form

$$x(t, t_0, 0, f) = \int_{t_0}^{t} S(t, \sigma)f(\sigma)\, d\sigma. \tag{1.8}$$

Then it is shown, in the proof of Theorem 3.1, that for any x_0 in H and all $t \geqq t_0 \geqq 0$ there exists an essentially bounded measurable mapping $f: [t_0, \infty) \to H$ such that

$$(t-t_0)S(t, t_0)x_0 = \int_{t_0}^{t} S(t, \sigma)f(\sigma)\, d\sigma. \tag{1.9}$$

Using results from [5] and (1.9), the Perron condition is proved for (1.7) and then extended in Theorem 3.2 to (1.4).

2. Preliminaries.

DEFINITION 2.1. Let H stand for a real Hilbert space. The bounded linear mappings from H into itself will be denoted by $[H]$. The identity mapping in $[H]$ will be denoted by I.

DEFINITION 2.2. The norm in any Hilbert space will be denoted by $|\cdot|$.

DEFINITION 2.3. The direct sum of two Hilbert spaces H_1 and H_2 will be denoted by $H_1 + H_2 = \{(x_1, x_2): x_1 \in H_1 \text{ and } x_2 \in H_2\}$; $|x|^2 = |(x_1, x_2)|^2 = |x_1|^2 + |x_2|^2$.

DEFINITION 2.4. Let $[a, b) \subset R$, $-\infty \leqq a$, $b \leqq \infty$. Let H be a Hilbert space and $f: [a, b] \to H$ Bochner square integrable. The equivalence classes of such mappings f will be denoted by $L_2[[a, b), H]$. The space $C[[-h, 0], H] = C_h$ will denote the space of continuous mappings from $[-h, 0]$ in H. The norm on C_h is

$$\phi = \sup\{|\phi(\sigma)|: \sigma \in [-h, 0]\}.$$

The equivalence classes of measurable mappings f from $[0, \infty)$ into H such that

$$\operatorname*{ess\,sup}_{0 \leqq t < \infty} |f(t)| < \infty,$$

will be denoted by $L[R^+, H]$.

Notational convention. The symbol Δ will denote the set in $R^+ \times R^+$ defined by

$$\Delta = \{(t, t_0) : 0 \leqq t_0 \leqq t \leqq \infty\}.$$

DEFINITION 2.5. A family of mappings $U(t, t_0)$ in $[H]$ with $(t, t_0) \in \Delta$ will be called a *strongly continuous evolutionary process with exponential growth* if, for all (s, t_0) and (t, s) in Δ with $t_0 \leqq s \leqq t$ and $x \in H$:

(i) $U(t, s)U(s, t_0) = U(t, t_0)x$;

(ii) there exists constants $M_1 \geqq 1$ and $\omega > 0$ such that

$$|U(t, t_0)| \leqq M_1 e^{\omega(t - t_0)}$$

for all $(t, t_0) \in \Delta$;

(iii) $U(\cdot, t_0)$ is strongly continuous for $t \geqq t_0$ and $|U(t, t_0)| = 0$ for $t < t_0$;

(iv) $\lim_{t \to t_0^+} U(t, t_0)x = x$.

A family satisfying (i)–(iv) will be called an *evolutionary process* of class $C(0, e)$.

Observe that if $U(t, t_0)$ satisfies, in addition to Definition 2.5, the condition $U(t, t_0) = U(t - t_0, 0)$, then the evolutionary process is a semigroup of class C_0 (see, e.g., [8]).

Let $0 < h_1 < h_2 \cdots < h_m = h$. Let $T(t, t_0) \subset [H]$ be an evolutionary process of class $C(0, e)$, i.e., a family of mappings which satisfies Definition 2.5. Let $\{B_j(t)\}$ and $\{A_j(t)\}$, $1 \leqq j \leqq m$, be two families of strongly continuous mappings from $[0, \infty) \to [H]$ satisfying, for all j, the conditions $|A_j(t)| \leqq M_1$, $|B_j(t)| \leqq M_1$ and $|B_j(t)| = 0$ if $t < 0$. We consider the system of delay equations defined as follows:

$$\begin{aligned} &y(t, t_0, \phi) - \sum_{j=1}^{m} B_j(t) y(t - h_j, t_0, \phi) \\ &\quad = T(t, t_0)\left[\phi(0) - \sum_{j=1}^{m} B_j(t_0)\phi(-h_j)\right] \\ &\qquad + \int_{t_0}^{t} \left[T(t, \sigma) \sum_{j=1}^{m} A_j(\sigma) y(\sigma - h_j, t_0, \phi)\right] d\sigma \end{aligned} \tag{2.1a}$$

if $t \geqq t_0 \geqq 0$, and

$$y(t, t_0, \phi) = \phi(t - t_0), \qquad \phi \in C_h \tag{2.1b}$$

if $t \in [t_0 - h, t_0]$.

The primary goal of this paper is to develop a Perron condition for systems of type (2.1). To do this we shall find it convenient to first develop the Perron condition for systems of the type (1.7). To do this it is necessary to make the following transformation.

Let $t_0 \geqq 0$ and $\phi \in C_h$. Define

$$\psi(0) = \phi(0) - \sum_{j=1}^{m} B_j(t_0)\phi(-h_j), \tag{2.2a}$$

$$\psi(\sigma) = \phi(\sigma) - \sum_{j=1}^{m} B_j(t_0 + \sigma)\phi(\sigma - h_j) \quad \text{if } \sigma \in [-h, 0), \tag{2.2b}$$

$$\|\psi(\sigma)\| = 0 \quad \text{if } \sigma > -h. \tag{2.2c}$$

Clearly, (2.2) defines a linear mapping on C_h.

Applying (2.2) to (2.1), we obtain

$$\begin{aligned} x(t, t_0, \hat{\psi}) &= y(t, t_0, \phi) - \sum_{j=1}^{m} B_j(t) y(t-h_j, y_0, \phi) \\ &= T(t, t_0)\psi(0) + \int_{t_0}^{t} \left[T(t, \sigma) \sum_{j=1}^{m} A_j(\sigma) y(\sigma - h_j, y_0, \phi) \right] d\sigma. \end{aligned} \tag{2.3}$$

The y terms under the integral in (2.3) may be replaced by x terms if we formally invert

$$x(t) = y(t) - \sum_{j=1}^{m} B_j(t) y(t-h_j) \tag{2.4}$$

to obtain

$$y(t) = x(t) + \sum_{j=1}^{\infty} \hat{H}_j(t) x(t-\omega_j), \tag{2.5}$$

where $h_1 = \omega_1 < \omega_2 \cdots$ is the ordered sequence of real numbers consisting of all possible linear combinations of the $\{h_j\}$ over the semigroup of positive integers and zero, i.e., $\{\omega_j\}$ are of the form $\omega_j = n_1 h_1 + \cdots + n_m h_m$ where $\{n_i\}$ are positive integers or zero. If, for some $\delta > 0$, the formal series in (2.5) $\{\hat{H}_j(t)\}$ satisfies the conditions

$$|\hat{H}_j(t)| \leqq \hat{r}_j, \tag{2.6}$$

$$\sum_{j=1}^{\infty} \hat{r}_j e^{2\delta\omega_j} < \infty, \tag{2.7}$$

then (2.4) and (2.5) are the inverses of one another in the sense that any mapping $y: (-\infty, \infty) \to H$ which satisfies

$$|y(t)| = 0 \quad \text{if } t < -h, \qquad |y(t)| \leqq M e^{kt} \quad \text{if } t \geqq -h$$

defines $x(t)$, and conversely $x(t)$, satisfying (2.4), can be converted to $y(t)$ by (2.5). In [5], it is shown that a system of the form (2.1) satisfying (2.6) can be converted via (2.4) into a system of the following type:

$$x(t, t_0, \hat{\phi}) = T(t, t_0)\phi(0) + \int_{t_0}^{t} T(t, \sigma) \sum_{j=1}^{\infty} H_j(\sigma) x(\sigma - \omega_j, t_0, \hat{\phi})\, d\sigma \tag{2.8a}$$

if $t \geqq t_0$,

$$x(t, t_0, \phi) = \phi(t - t_0) \tag{2.8b}$$

if $t \in (-\infty, t_0)$ and

$$x(t_0, t_0, \hat{\phi}) = \phi(0). \tag{2.8c}$$

Here

$$|H_j(t)| \leqq r_j, \tag{2.9}$$

and for the same δ as in (2.7),

$$\sum_{j=1}^{\infty} r_j e^{2\delta\omega_j} < \infty \tag{2.10}$$

(see [5, pp. 118–119]). However, not every solution of (2.8) can be reduced to a solution of (2.1). In any case, there is a strong connection between the two systems if we make the following assumption which will hold from now on.

Assumption 2.1. We shall assume that (2.6) and (2.7) hold and hence systems (2.1) and (2.8) are related by (2.4).

Let $\tilde{L}_2[(-\infty, 0), H]$ denote the set of equivalence classes of all measurable mappings ϕ from $(-\infty, 0) \to H$ such that

$$|\phi|^2 = \sum_{j=1}^{\infty} r_j^2 \, e^{2\delta\omega_j} \int_{-\omega_j}^{0} |\phi(\sigma)|^2 \, d\sigma < \infty. \tag{2.11}$$

Let

$$\mathcal{H} = H + \tilde{L}_2[(-\infty, 0), H]. \tag{2.12}$$

DEFINITION 2.6. System (2.1), ((2.8)) is uniformly L_2 stable if, for all $\phi \in C_h$ ($\hat{\phi} \in \mathcal{H}$), there exists a finite constant $M(\phi)$ ($M(\hat{\phi})$), independent of t_0, such that

$$\int_{t_0}^{\infty} |y(t, t_0, \phi)|^2 \, dt \leqq M(\phi) \qquad \left(\int_{t_0}^{\infty} |x(t, t_0, \hat{\phi})|^2 \, dt \leqq M(\hat{\phi}) \right).$$

THEOREM 2.1. *If systems* (2.1) *and* (2.8) *are connected by* (2.4) *and* (2.5), *then* (2.1) *is uniformly* L_2 *stable if and only if* (2.8) *is uniformly* L_2 *stable.*

Proof. See [5, Theorem 3.1].

COROLLARY. *If systems* (2.1) *and* (2.8) *are connected by* (2.4) *and* (2.5) *they are uniformly* L_2 *stable if and only if there exist constants* M_1 *and* $\alpha > 0$, *independent of* t_0, *such that, for all* $t \geqq t_0$,

$$|y(t, t_0, \phi)| \leqq M_1 \, e^{-\alpha(t-t_0)} |\phi|$$

and

$$|x(t, t_0, \hat{\phi})| \leqq M_1 \, e^{-\alpha(t-t_0)} |\hat{\phi}|.$$

Hence, systems (2.1) *and* (2.8) *are uniformly* L_2 *stable if and only if there exists* M_2, *independent of* t_0, *such that*

$$\int_{t_0}^{\infty} |y(\alpha, t_0, \phi)|^2 \, d\alpha \leqq M_2 |\phi|^2$$

and

$$\int_{t_0}^{\infty} |x(\alpha, t_0, \hat{\phi})|^2 \, d\alpha \leqq M_2 |\hat{\phi}|^2.$$

Proof. See [5, Thm. 3.4 and its corollary].

In conjunction with systems (2.1) and (2.8), we shall consider their non-homogeneous versions. They are defined, in the case of (2.1), by the equation

$$\begin{aligned} & y(t, t_0, \phi, f) - \sum_{j=1}^{m} B_j(t) y(t - h_j, t_0, \phi, f) \\ & = T(t, t_0) \left[\phi(0) - \sum_{j=1}^{m} B_j(t_0) \phi(-h_j) \right] \\ & \quad + \int_{t_0}^{t} \left[T(t, \sigma) \sum_{j=1}^{m} A_j(\sigma) y(\sigma - h_j, t_0, \phi, f) \right] d\sigma + \int_{t_0}^{t} T(t, \sigma) f(\sigma) \, d\sigma \end{aligned} \tag{2.13}$$

if $t \geqq t_0$, $f \in L_\infty[R^+, H]$ and $\phi \in C_h$. In the case of (2.8), the corresponding non-homogeneous equation has, by [5, Thms. 2.9, 2.10], the unique form

$$x(t, t_0, \hat{\phi}, f) = x(t, t_0, \hat{\phi}) + \int_{t_0}^{t} S(t, \sigma) f(\sigma)\, d\sigma \tag{2.14}$$

where $S: \Delta \to [H]$ satisfies

$$S(t, t_0)x_0 = T(t, t_0)x_0 + \int_{t_0}^{t} T(t, \sigma) \sum_{j=1}^{\infty} H_j(\sigma) S(\sigma - \omega_j, t_0)x_0\, d\sigma \tag{2.15a}$$

and

$$|S(t, t_0)| = 0, \tag{2.15b}$$

if $t < t_0$.

Thus, the existence of solutions of (2.14) is obvious. To prove the existence of solutions of (2.13), we let $\phi \in C_h$ and $f \in L_\infty[R^+, H]$ be given and define $\hat{\psi}$ in $\mathcal{H}$ by means of (2.2). Then, [5, Thm. 2.9], (2.2) and the definition of the $\{H_j(t)\}$ given in [5, pp. 118, 119], we know there exists $y(t)$ such that

$$\begin{aligned}
x(t, t_0, \hat{\psi}, f) &= x(t, t_0, \hat{\psi}) + \int_{t_0}^{t} S(t, \sigma) f(\sigma)\, d\sigma \\
&= T(t, t_0)\psi(\sigma) + \int_{t_0}^{t} \left[\sum_{j=1}^{\infty} T(t, \sigma) H_j(\sigma) x(\sigma - \omega_j, t_0, \hat{\psi}, f) \right] d\sigma \\
&\quad + \int_{t_0}^{t} T(t, \sigma) f(\sigma)\, d\sigma \\
&= T(t, t_0)\left[\phi(\sigma) - \sum_{j=1}^{m} B_j(t_0)\phi(-h_j) \right] \\
&\quad + \int_{t_0}^{t} T(t, \sigma) \sum_{j=1}^{m} A_j(\sigma) \Big[x(\sigma - h_j, t_0, \hat{\psi}, f) \\
&\qquad + \sum_{k=1}^{\infty} \hat{H}_k(\sigma) x(\sigma - h_j - \omega_k, t_0, \hat{\psi}, f) \Big] d\sigma \\
&\quad + \int_{t_0}^{t} T(t, \sigma) f(\sigma)\, d\sigma \\
&= T(t, t_0)\left[\phi(0) - \sum_{j=1}^{\infty} B_j(t_0)\phi(-h_j) \right] \\
&\quad + \int_{t_0}^{t} T(t, \sigma) \left[\sum_{j=1}^{m} A_j(\sigma) y(\sigma - h_j) \right] d\sigma + \int_{t_0}^{t} T(t, \sigma) f(\sigma)\, d\sigma \\
&= y(t) - \sum_{j=1}^{m} B_j(t) y(t - h_j).
\end{aligned}$$

Clearly $y(t)$ above satisfies (2.13). Thus, we have:

THEOREM 2.2. *For given $\phi \in C_h$ and $f \in L_\infty[R^+, H]$, (2.13) has a unique solution.*

Proof. The above discussion establishes existence. Uniqueness follows from the uniqueness of solutions of (2.14) which is obvious.

We now define the piecewise strongly continuous mapping $Y: \Delta \to [H]$ for system (2.13), which is given by the equation

$$Y(t, \sigma) = S(t, \sigma) + \sum_{j=1}^{\infty} \hat{H}_j(t) S(t - \omega_j, \sigma). \tag{2.16}$$

THEOREM 2.3. *If* $|\phi| = 0$ *on* $[-h, 0]$ *and* $f \in L_\infty[R^+, H]$, *then the solution of* (2.13) *can be written in the form*

$$y(t, t_0, 0, f) = \int_{t_0}^{t} Y(t, \sigma) f(\sigma)\, d\sigma. \tag{2.17}$$

Proof. The solution of the system (2.14), related to system (2.13), is

$$x(t, t_0, 0, f) = \int_{t_0}^{t} S(t, \sigma) f(\sigma)\, d\sigma.$$

However, using (2.4) and the properties of $S(t, \sigma)$ given by (2.15), we have

$$\begin{aligned} y(t, t_0, 0, f) &= x(t, t_0, 0, f) + \sum_{j=1}^{\infty} \hat{H}_j(t) x(t - \omega_j, t_0, 0, f) \\ &= \int_{t_0}^{t} S(t, \sigma) f(\sigma)\, d\sigma + \sum_{j=1}^{\infty} \hat{H}_j(t) \int_{t_0}^{t-\omega_j} S(t - \omega_j, \sigma) f(\sigma)\, d\sigma \\ &= \int_{t_0}^{t} [S(t, \sigma) + \hat{H}_j(t) S(t - \omega_j, \sigma)] f(\sigma)\, d\sigma = \int_{t_0}^{t} Y(t, \sigma) f(\sigma)\, d\sigma. \end{aligned}$$

3. The Perron condition.

LEMMA 3.1. *If* $\int_{t_0}^{\infty} |S(t, t_0) x_0|^2\, dt \leqq M_3 |x_0|^2$, *for all* $t_0 \geqq 0$ *and* $x_0 \in H$, *where* M_3 *does not depend on* t_0 *and* x_0, *then system* (2.8) *and, consequently system* (2.1), *is uniformly* L_2 *stable.*

Proof. The proof is given in [5, proof of Thm. 3.1, eq. (3.18)ff].

THEOREM 3.1. *Suppose that for every* $f \in L_\infty[R^+, H]$ *there exists a finite constant* $M(f)$ *independent of* t_0 *such that the solution of* (2.14) $x(t, t_0, 0, f)$ *satisfies*

$$|x(t, t_0, 0, f)| = \left| \int_{t_0}^{t} S(t, \sigma) f(\sigma)\, d\sigma \right| \leqq M(f); \tag{3.1}$$

then system (2.8) *is uniformly* L_2 *stable.*

Proof. It can easily be shown, using the principle of uniform boundedness, that (3.1) implies the existence of $M_4 < \infty$ and independent of t_0 such that

$$|x(t, t_0, 0, f)| \leqq M_4 |f| \tag{3.2}$$

(see, e.g., [4, Thm. 6]).

Let $T > 0$ be fixed and define

$$S_T(\sigma, t_0) = S(\sigma, t_0), \quad \sigma \leqq T + t_0, \qquad \text{and} \qquad |S_T(\sigma, t_0)| = 0 \quad \text{if } \sigma > T + t_0. \tag{3.3}$$

Let $\hat{\phi} = (x_0, 0)$. By [5, eq. (2.52)],

$$\int_{t_0}^{t_0+T} x(t, t_0, \hat{\phi})\, d\sigma = T x(t, t_0, \hat{\phi}) = T S(t, t_0) x_0.$$

If $t \geqq t_0 + T$, we use (2.3), letting $\hat{\phi} = (S_T(\sigma, t_0) x_0, S_T(\cdot, t_0) x_0)$ and the representation [5, eq. (2.52)], which states that

$$x(t, t_0, \hat{\phi}) = S(t, t_0) \phi(0) + \sum_{j=1}^{\infty} \int_{t_0-\omega_j}^{t_0} S(t, \sigma + \omega_j) H_j(\sigma + \omega_j) \phi(\sigma)\, d\sigma.$$

We obtain

$$\begin{aligned}
&\int_{t_0}^{t_0+T} x(t, t_0, \hat{\phi})\, d\sigma \\
&\quad = Tx(t, t_0, \hat{\phi}) \\
&\quad = \int_{t_0}^{t_0+T} S(t, t_0)x_0\, dt \\
&\quad = \int_{t_0}^{t_0+T} \left[S(t, \sigma)S_T(\sigma, t_0)x_0 + \sum_{j=1}^{\infty} \int_{\alpha-\omega_j}^{\sigma} S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha \right] d\alpha \\
&\quad = \int_{t_0}^{t_0+T} S(t, \sigma)S_T(\sigma, t_0)x_0 \qquad (3.4) \\
&\qquad + \sum_{j=1}^{\infty} \int_{t_0}^{t_0+T} \int_{\sigma-\omega_j}^{\sigma} S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha\, d\sigma \\
&\quad = \int_{t_0}^{t} S(t, \sigma)S_T(\sigma, t_0)x_0\, d\sigma \\
&\qquad + \sum_{j=1}^{\infty} \int_{t_0+T-\omega_j}^{t_0+T} \int_{\alpha}^{t_0+T} S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\sigma\, d\alpha \\
&\qquad + \sum_{j=1}^{\infty} \int_{t_0}^{t_0+T-\omega_j} \int_{\alpha}^{\alpha+\omega_j} S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\sigma\, d\alpha \\
&\qquad + \sum_{j=1}^{\infty} \int_{t_0-\omega_j}^{t_0} \int_{t_0}^{\alpha+\omega_j} S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\sigma\, d\alpha.
\end{aligned}$$

Since $|S_T(\alpha, t_0)x_0| = 0$ if $t_0 - \omega_j \leqq \alpha < t_0$, (3.4) can be simplified to

$$\begin{aligned}
Tx(t, t_0, \hat{\phi}) &= \int_{t_0}^{t} S(t, \sigma)S_T(\sigma, t_0)x_0\, d\sigma \\
&\quad + \sum_{j=1}^{\infty} \int_{t_0+t-\omega_j}^{t_0+T} (t_0+T-\alpha)S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha \\
&\quad + \sum_{j=1}^{\infty} \int_{t_0}^{t_0+T-\omega_j} \omega_j S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha \qquad (3.5) \\
&= \int_{t_0}^{t} S(t, \sigma)S_T(\sigma, t_0)x_0\, d\sigma \\
&\quad + \sum_{j=1}^{\infty} \int_{t_0}^{t} \psi_j(\alpha)S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha,
\end{aligned}$$

where the scalar function ψ_j, on the right side of (3.5) is defined

$$\psi_j(\alpha) = \begin{cases} t_0+T-\alpha, & t_0+T-\omega_j \leqq \alpha \leqq t_0+T, \\ \omega_j, & t_0 \leqq \alpha \leqq t_0+T-\omega_j, \\ 0 & \text{otherwise.} \end{cases} \qquad (3.6)$$

Notice that for each j,

$$0 \leqq \psi_j(\alpha) \leqq \omega_j. \qquad (3.7)$$

Furthermore, there exist constants M_5 and k such that, for all α,

$$|S_T(\alpha, t_0)x_0| \leqq M_5 e^{kT}|x_0|. \tag{3.8}$$

(See, e.g., [5, eq. (2.31), p. 124].) Thus for each j the inequality

$$|\psi_j(\alpha)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0| \leqq r_j\omega_j M_5 e^{kT}|x_0| \tag{3.9}$$

holds.

By [5, Lemma 1.2] and (2.9),

$$\sum_{j=1}^{\infty} r_j\omega_j < \infty. \tag{3.10}$$

Moreover, for each j on the right side of (3.5) we can write

$$\begin{aligned}
&\int_{t_0}^{t} \psi_j(\alpha)S(t, \alpha+\omega_j)H_j(\alpha+\omega_j)S_T(\alpha, t_0)x_0\, d\alpha \\
&= \int_{t_0+\omega_j}^{t+\omega_j} \psi(\beta-\omega_j)S(t, \beta)H_j(\beta)S_T(\beta-\omega_j, t_0)x_0\, d\beta \\
&\quad - \int_{t_0}^{t} \psi_j(\beta-\omega_j)S(t, \beta)H_j(\beta)S_T(\beta-\omega_j, t_0)x_0\, d\beta \\
&= \int_{t_0}^{t} S(t, \beta)f_j(\beta)\, d\beta.
\end{aligned} \tag{3.11}$$

In (3.11),

$$f_j(\beta) = \begin{cases} \psi_j(\beta-\psi_j)H_j(\beta)S_T(\beta-\omega_j, t_0)x_0 & \text{if } t_0+\omega_j \leqq \beta \leqq t \leqq T+t_0+\omega_j, \\ 0 & \text{if } t_0 \leqq \beta \leqq t_0+\omega_j. \end{cases}$$

Thus, from (3.9) and the definition of f_j,

$$|f_j(\beta)| \leqq f_j\omega_j M_5 e^{kT}|x_0| \tag{3.12}$$

for all β in $[0, \infty)$. From the above, we can write (3.4) as

$$TS(t, t_0)x_0 = \int_{t_0}^{t} S(t, \sigma)\left[S_T(\sigma, t_0)x_0 + \sum_{j=1}^{\infty} f_j(\sigma)\right] d\sigma.$$

Thus from (3.2) we obtain for $t \geqq t_0+T$

$$|TS(t, t_0)x_0| \leqq M_4M_5 e^{kT}|x_0|\left[1+\sum_{j=1}^{\infty} r_j\omega_j\right]. \tag{3.13}$$

By (3.1), inequality (3.13) can be rewritten, if $t \geqq t_0+T$,

$$|TS(t, t_0)x_0| \leqq M_6|x_0|\, e^{kT}. \tag{3.14}$$

Hence, since $T>0$ was arbitrary, we choose $T=1$. Then for $t \geqq t_0$

$$|S(t, t_0)x_0| \leqq M_7|x_0|, \tag{3.15}$$

where $M_7 = \max[M_6 e^k, M_5 e^k]$ does not depend on t_0.

We now repeat the above argument replacing $S_T(\sigma, t_0)$ by $S(\sigma, t_0)$ and t_0+T by $t \geqq t_0$ and observe that inequality (3.8) becomes (3.15). Inequality (3.13) becomes

$$|(t-t_0)S(t, t_0)x_0| \leqq M_4M_7|x_0|\left[1+\sum_{j=1}^{\infty} r_j\omega_j\right], \tag{3.16}$$

which if $t > t_0$ yields

$$|S(t, t_0)x_0| \leqq \frac{M_8}{t-t_0}|x_0| \tag{3.17}$$

for all x_0 in X. Since (3.17) implies that

$$\int_{t_0}^{\infty} |S(t, t_0)x_0|^2 \, dt \leqq M_9|x_0|^2, \tag{3.18}$$

where M_9 does not depend on t_0, it follows from Lemma 3.1 that system (2.8) is uniformly L_2 stable.

THEOREM 3.2. *System* (2.1) *is uniformly* L_2 *stable if and only if the solutions of its nonhomogeneous version* (2.13) *satisfy, for all* $t \geqq t_0$ *and for all* $f \in L_\infty[R^+, H]$, *an inequality of the type*

$$|y(t, t_0, 0, f)| = \left| \int_{t_0}^{t} Y(t, \sigma)f(\sigma) \, d\sigma \right| \leqq M(f) < \infty, \tag{3.19}$$

where $M(f)$ *does not depend on* t_0.

Proof. (i) If (2.1) is uniformly L_2 stable, then by Theorem 2.1 so is (2.8). Hence, by (2.16), (2.6), and Theorem 2.1 and its corollary,

$$\begin{aligned} |y(t, t_0, 0, f)| &= \left| \int_{t_0}^{t} Y(t, \sigma)f(\sigma) \, d\sigma \right| \\ &\leqq \left| \int_{t_0}^{t} \left[S(t, \sigma) + \sum_{j=1}^{\infty} \hat{H}_j(t)S(t-\omega_j, \sigma) \right] f(\sigma) \, d\sigma \right| \\ &\leqq \left[\frac{M_1}{\alpha} + \frac{M_1}{\alpha} \sum_{j=1}^{\infty} \hat{r}_j \right] |f| = M(f). \end{aligned}$$

(ii) If (3.19) holds, then the inversion of (2.16) and the bounds on the $|B_j(t)|$ yield, for all f in $L_\infty[R^+, H]$, the inequality

$$\begin{aligned} \left| \int_{t_0}^{t} S(t, \sigma)f(\sigma) \, d\sigma \right| &= \left| \int_{t_0}^{t} Y(t, \sigma)f(\sigma) \, d\sigma - \sum_{j=1}^{m} B_j(t) \int_{t_0}^{t-h_j} Y(t-h_j, \sigma)f(\sigma) \, d\sigma \right| \\ &\leqq M_1 M(f)m. \end{aligned}$$

Hence, Theorem 3.1 is satisfied, and, by Theorem 2.1, system (2.1) is uniformly L_2 stable.

Acknowledgment. I would like to thank the referee for suggesting the elimination of much chaff in § 2 of this paper.

REFERENCES

[1] R. BELLMAN, *On an application of a Banach–Steinhaus theorem to the study of the boundedness of solutions of nonlinear differential and difference equations*, Ann. Math., 49 (1948), pp. 515–522.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] J. L. DALETSKII AND M. G. KREIN, *Stability of Solutions of Differential Equations in a Banach Space*, Izdat. Nauka, Moscow, 1970 (in Russian).

[4] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, this Journal, 3 (1973), pp. 428–445.

[5] ———, *Representation of solutions and stability of linear differential-difference equations in a Banach space*, J. Differential Eqs., 29 (1978), pp. 105–166.

[6] A. HALANAY, *Differential Equations Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.
[7] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
[8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, Colloquium Publications, Amer. Math. Soc., Providence, RI, 1957.
[9] O. PERRON, *Die Stabilitätsfrage bei Differentialgleichungen*, Math. Z., 32 (1930), pp. 703–728.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0006 $01,00/0

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF $y''=\phi(t)f(y)$*

STEVEN D. TALIAFERRO†

Abstract. Explicit formulas are obtained for the asymptotic behavior of solutions of the three problems

$$y''=\phi(t)f(y), \qquad \lim_{t\to\infty} y(t)=\lim_{t\to\infty} y'(t)=0,$$

$$y''+\phi(t)f(y)=0, \qquad \lim_{t\to 0^+} y(t)=0, \qquad \lim_{t\to 0^+} y'(t)=\infty,$$

$$y''+\phi(t)f(y)=0, \qquad \lim_{t\to\infty} y(t)=\infty, \qquad \lim_{t\to\infty} y'(t)=0,$$

as t tends to ∞, 0, and ∞ respectively. In all three problems $\phi(t)$ and $f(y)$ are assumed to be positive and continuous. Necessary conditions for existence of solutions to these problems are also given.

1. We will consider the following "initial value" problems, where a and b are positive constants:

(A) $$y''=\phi(t)f(y), \qquad (t, y)\in(a,\infty)\times(0, b)=I\times J,$$
$$\lim_{t\to\infty} y(t)=\lim_{t\to\infty} y'(t)=0;$$

(B) $$y''+\phi(t)f(y)=0, \qquad (t, y)\in(0, a)\times(0, b)=I\times J,$$
$$\lim_{t\to 0^+} y(t)=0, \qquad \lim_{t\to 0^+} y'(t)=\infty;$$

(C) $$y''+\phi(t)f(y)=0, \qquad (t, y)\in(a,\infty)\times(b,\infty)=I\times J,$$
$$\lim_{t\to\infty} y(t)=\infty, \qquad \lim_{t\to\infty} y'(t)=0.$$

In all three problems, $f: J\to(0,\infty)$ and $\phi: I\to[0,\infty)$ are continuous. We will mainly be concerned with the asymptotic behavior of solutions of problems (A), (B) and (C) as t tends to ∞, 0 and ∞ respectively; however we will also discuss existence of solutions to these problems. We group these problems together because they can all be handled using the same techniques.

The Thomas–Fermi problem,

(1) $$y''=t^{-1/2}y^{3/2}, \qquad y(\infty)=0, \qquad y'(\infty)=0,$$

which arises in nuclear physics [2], [9], is a special case of problem (A), and so is the Emden–Fowler equation $y''=t^{\sigma}y^{\lambda}$, $\lambda>1$, which is dealt with in [1]. The problem

(2) $$y''+\frac{t}{y^{\lambda}}=0, \qquad y(0)=0, \qquad y'(0)=\infty, \qquad \lambda>0,$$

which arises in fluid mechanics [5], is a special case of problem (B).

Problem (A) with $f(y)=y^{\lambda}$, $\lambda>1$, is discussed in [6], and Problems (B) and (C) with $f(y)=y^{-\lambda}$, $\lambda>0$, are dealt with in [7] and [8], respectively. The results in these papers rely heavily on the monotonicity of $f(y)$. In this paper we make no monotonicity assumptions. Also results on existence of solutions of the more general problem $y''=f(t, y, y')$, $y(\infty)=y'(\infty)=0$, can be found in [3], [10].

* Received by the editors May 14, 1980, and in final form January 29, 1981.
† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

The recent paper [4] contains results for problem (A) when $\phi(t)$ and $f(y)$ belong to the class of o-regularly varying functions. The definition of o-regularly varying functions partially motivates the following definition which will be used in stating the theorems.

DEFINITION. For $c, d > 0$, let

$$A(d) = \overline{\lim_{y\to 0^+}} \sup_{0<\alpha<d} \frac{f(\alpha y)}{\alpha f(y)},$$

$$B(d) = \underline{\lim_{y\to 0^+}} \inf_{0<\alpha<d} \frac{f(\alpha y)}{\alpha f(y)},$$

$$C(d) = \overline{\lim_{y\to \infty}} \sup_{\alpha>d} \frac{f(\alpha y)}{\alpha f(y)};$$

$$A^{-1}(c) = \sup[\{d>0: A(d)<c\}\cup\{0\}],$$

$$B^{-1}(c) = \sup[\{d>0: B(d)>c\}\cup\{0\}],$$

$$C^{-1}(c) = \inf\{d>0: C(d)<c\};$$

$$\beta(d) = \underline{\lim_{y\to 0^+}} \inf_{0<\alpha<d} \frac{f(\alpha y)}{f(y)},$$

$$\gamma(d) = \overline{\lim_{y\to \infty}} \sup_{\alpha>d} \frac{f(\alpha y)}{f(y)}.$$

The three sets of functions $\{A(d), A^{-1}(c)\}$, $\{B(d), B^{-1}(c), \beta(d)\}$ and $\{C(d), C^{-1}(c), \gamma(d)\}$ will be used in the analysis of problems (A), (B) and (C), respectively. Since $f(y)$ is positive, we have that $A(d)$ and $A^{-1}(c)$ are nonnegative and increasing and the other 6 functions are nonnegative and decreasing. It follows directly from the definition that $A(d^2) \leqq A(d)^2$, $dB(d) \geqq \beta(d)$ and $dC(d) \leqq \gamma(d)$. Thus, $A(d)<1$ for some $0<d<1$ implies $A(d)\to 0$ as $d\to 0$ and $A^{-1}(c)$ is positive for $c>0$; $\lim_{d\to 0^+}\beta(d)>0$ implies $B(d)\to\infty$ as $d\to 0^+$ and $B^{-1}(c)$ is positive for $c>0$; and $\lim_{d\to\infty}\gamma(d)<\infty$ implies $C(d)\to 0$ as $d\to\infty$ and $C^{-1}(c)<\infty$ for $c>0$. It is only under these conditions the following asymptotic results are useful, but this includes a large class of nonlinear differential equations; in particular, problems (1) and (2) and the Emden–Fowler equation are in this class.

For the following four theorems we use the conventions $1/0 = +\infty$ and $(1/+\infty) = 0$.

THEOREM 1A. *Let I and J be as in problem* (A), $t_0 \in I$, $y_0 \in J$, $f: J \to (0,\infty)$ *be continuously differentiable, and $\phi: I \to (0,\infty)$ be continuous. Suppose for some $d>0$ we have $A(d)<1$. If for some continuously differentiable function $\psi: I\to(0,\infty)$ we have*

$$0 < c_1 = \underline{\lim_{t\to\infty}} \frac{\psi(t)}{\phi(t)} \leqq \overline{\lim_{t\to\infty}} \frac{\psi(t)}{\phi(t)} = c_2 < \infty \tag{3}$$

and

$$0 < c_3 = \underline{\lim_{t\to\infty}} 1 - \left(\frac{1}{\sqrt{\psi(t)}H'(I(t))}\right)' \leqq \overline{\lim_{t\to\infty}} 1 - \left(\frac{1}{\sqrt{\psi(t)}H'(I(t))}\right)' = c_4 < \infty, \tag{4}$$

then any solution, $y(t)$, of problem (A) *satisfies*

$$A^{-1}(c_1c_3-) \leqq \underline{\lim_{t\to\infty}} \frac{y(t)}{F^{-1}(I(t))} \leqq \overline{\lim_{t\to\infty}} \frac{y(t)}{F^{-1}(I(t))} \leqq \frac{1}{A^{-1}(1/c_2c_4-)},$$

where

$$F(\xi)=\int_{\sqrt{\xi}}^{\sqrt{y_0}}\frac{2\,dx}{\sqrt{f(x^2)}},\qquad H(\zeta)=\log F^{-1}(\zeta),\qquad I(t)=\int_{t_0}^{t}\sqrt{\psi(\tau)}\,d\tau.$$

THEOREM 1B.1. *Let I and J be as in problem* (B), $t_0\in I$, $y_0\in J$, $f:J\to(0,\infty)$ *be continuously differentiable,* $\phi:I\to(0,\infty)$ *be continuous, and* $\int_0^{t_0}\sqrt{\phi(\tau)}\,d\tau<\infty$. *Suppose* $\lim_{d\to 0^+}\beta(d)>0$. *If for some continuously differentiable function* $\psi:I\to(0,\infty)$ *we have*

$$0<c_1=\varliminf_{t\to 0^+}\frac{\psi(t)}{\phi(t)}\leqq\varlimsup_{t\to 0^+}\frac{\psi(t)}{\phi(t)}=c_2<\infty \tag{5}$$

and

$$\begin{aligned}0<c_3&=\varliminf_{t\to 0^+}1+2H(I(t))\left(\frac{1}{\sqrt{\psi(t)}H'(I(t))}\right)'\\&\leqq\varlimsup_{t\to 0^+}1+2H(I(t))\left(\frac{1}{\sqrt{\psi(t)}H'(I(t))}\right)'=c_4<\infty,\end{aligned} \tag{6}$$

then any solution, $y(t)$, *of problem* (B) *satisfies*

$$B^{-1}(c_2c_4+)\leqq\varliminf_{t\to 0^+}\frac{y(t)}{F^{-1}(I(t))}\leqq\varlimsup_{t\to 0^+}\frac{y(t)}{F^{-1}(I(t))}\leqq\frac{1}{B^{-1}\left(\frac{1}{c_1c_3}+\right)},$$

where

$$F(\xi)=\int_0^{\sqrt{\xi}}\frac{2\,dx}{\sqrt{f(x^2)}},\qquad H(\zeta)=[F^{-1}(\zeta)]^2,\qquad I(t)=\int_0^t\sqrt{\psi(\tau)}\,d\tau.$$

The condition in Theorem 1B.1 that $\int_0^{t_0}\sqrt{\phi(\tau)}\,d\tau<\infty$ is not very strong, because, by Theorem 2B, if problem (B) has a solution then $\int_0^{t_0}t\phi(t)\,dt<\infty$; but if we have the slightly stronger condition $\int_0^{t_0}t^{1-\varepsilon}\phi(t)\,dt<\infty$ for some $\varepsilon>0$ then

$$\int_0^{t_0}\sqrt{\phi(t)}\,dt=\int_0^{t_0}\sqrt{t^{1-\varepsilon}\phi(t)}\,t^{-[(1-\varepsilon)/2]}\,dt\leqq\left(\int_0^{t_0}t^{1-\varepsilon}\phi(t)\,dt\right)^{1/2}\left(\int_0^{t_0}t^{-(1-\varepsilon)}\,dt\right)^{1/2}<\infty.$$

In the following theorem we dispense with this condition on ϕ at the expense of giving the asymptotic behavior of $y(t)$ as $t\to 0$ as the solution of a *first* order differential equation rather than giving it by an explicit formula as in Theorem 1B.1. However in many problems, including (2), this first order differential equation can be easily solved. Also, since $\lim_{d\to 0^+}\beta(d)>0$ implies $f(y)$ is bounded away from zero for y near zero, the integral which defines $F(\xi)$ in Theorem 1B.1 converges

THEOREM 1B.2. *Let I and J be as in problem* (B), $f:J\to(0,\infty)$ *be continuously differentiable, and* $\phi:I\to(0,\infty)$ *be continuous. Suppose* $\lim_{d\to 0^+}\beta(d)>0$. *If for some continuously differentiable function* $\psi:I\to(0,\infty)$ *we have*

$$0<c_1=\varliminf_{t\to 0^+}\frac{\psi(t)}{\phi(t)}\leqq\varlimsup_{t\to 0^+}\frac{\psi(t)}{\phi(t)}=c_2<\infty \tag{7}$$

and

$$0<c_3=\varliminf_{t\to 0^+}1+\frac{t\bar{g}''}{2\bar{g}'}\leqq\varlimsup_{t\to 0^+}1+\frac{t\bar{g}''}{2\bar{g}'}=c_4<\infty, \tag{8}$$

then any solution, $y(t)$, of problem (B) *satisfies*

$$B^{-1}(c_2c_4+)\leqq \underline{\lim}_{t\to 0^+}\frac{y(t)}{t\bar{g}(t)}\leqq \overline{\lim}_{t\to 0^+}\frac{y(t)}{t\bar{g}(t)}\leqq \frac{1}{B^{-1}\left(\frac{1}{c_1c_3}+\right)},$$

where $\bar{g}: I\to(0,\infty)$ is a solution of

$$g'=-\tfrac{1}{2}\psi(t)f(tg). \tag{9}$$

THEOREM 1C. *Let I and J be as in problem* (C), $t_0\in I$, $y_0\in J$, $f: J\to(0,\infty)$ *be continuously differentiable, and $\phi: I\to(0,\infty)$ be continuous. Suppose $\lim_{d\to\infty}\gamma(d)<\infty$. If for some continuously differentiable function $\psi: I\to(0,\infty)$ we have*

$$0<c_1=\underline{\lim}_{t\to\infty}\frac{\psi(t)}{\phi(t)}\leqq\overline{\lim}_{t\to\infty}\frac{\psi(t)}{\phi(t)}=c_2<\infty \tag{10}$$

and

$$\begin{aligned}0<c_3&=\underline{\lim}_{t\to\infty}1+2H(I(t))\left(\frac{1}{H'(I(t))\sqrt{\psi(t)}}\right)'\\ &\leqq\overline{\lim}_{t\to\infty}1+2H(I(t))\left(\frac{1}{H'(I(t))\sqrt{\psi(t)}}\right)'=c_4<\infty,\end{aligned} \tag{11}$$

then any solution, $y(t)$, of problem (C) *satisfies*

$$C^{-1}(c_1c_3-)\leqq\overline{\lim}_{t\to\infty}\frac{y(t)}{F^{-1}(I(t))}\leqq\underline{\lim}_{t\to\infty}\frac{y(t)}{F^{-1}(I(t))}\leqq\frac{1}{C^{-1}\left(\frac{1}{c_2c_4}-\right)},$$

where

$$F(\xi)=\int_{\sqrt{y_0}}^{\sqrt{\xi}}\frac{2\,dx}{\sqrt{f(x^2)}},\qquad H(\zeta)=[F^{-1}(\zeta)]^2,\qquad I(t)=\int_{t_0}^{t}\sqrt{\psi(\tau)}\,d\tau.$$

2. The following three theorems are concerned with existence of solutions of problems (A), (B) and (C).

THEOREM 2A. *Let I and J be as in problem* (A), *and $f: J\to(0,\infty)$ and $\phi: I\to[0,\infty)$ be continuous. Suppose $A(d)<1$ for some $d>0$. Then for each $y_0\in J$ and $t_0\in I$ there is a positive decreasing solution of*

$$y''=\phi(t)f(y),\qquad y(t_0)=y_0,\qquad y'(\infty)=0. \tag{12}$$

This solution decreases to zero, and hence is a solution of problem (A), *if and only if*

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

Proof. Let $d_0>0$ be such that $A(d_0)<1$. Then, for some $y_1\in(0,y_0)$ and $c\in(0,1)$, we have for $0<\alpha<d_0$ and $0<y\leqq y_1$ that $f(\alpha y)<c\alpha f(y)$. We claim

$$\lim_{t\to 0^+}\frac{f(y)}{y}=0. \tag{13}$$

Otherwise we would have $\varepsilon > 0$ and a positive sequence $\{y_n\}_{n=1}$ decreasing to zero with $f(y_n) > \varepsilon y_n$, $y_{n+1} = \alpha_n y_n$, and $0 < \alpha_n < d_0$ for $n = 1, 2, 3, \cdots$. Then, for $n = 1, 2, 3, \cdots$,

$$f(y_{n+1}) = f(\alpha_n y_n) < c\alpha_n f(y_n) < c^n \alpha_n \cdots \alpha_1 f(y_1) = c^n \frac{f(y_1)}{y_1} y_{n+1},$$

a contradiction. Hence (13) is true.

A result of Hartman and Wintner [3, Thm. 1] gives the existence of a positive decreasing solution of (12).

Let $y(t)$ be a positive decreasing solution of (12). Integrating the differential equation twice from t to ∞ we get

$$y(t) - y(\infty) = \int_t^\infty (\tau - t)\phi(\tau)f(y(\tau))\, d\tau. \tag{14}$$

Suppose $y(\infty) = 0$. Then, by (13), $\lim_{t\to\infty} f(y(t))/y(t) = 0$, and for sufficiently large values of t we have from (14) that

$$y(t) \leqq \int_t^\infty (\tau - t)\phi(\tau)y(\tau)\, d\tau \leqq y(t) \int_t^\infty (\tau - t)\phi(\tau)\, d\tau.$$

So for large t we have $\int_t^\infty (\tau - t)\phi(\tau)\, d\tau \geqq 1$ and hence $\int_{t_0}^\infty \tau\phi(\tau)\, d\tau = \infty$.

Conversely, suppose $y(\infty) > 0$. Then, for large t, we have from (14) that

$$y(t) - y(\infty) > \tfrac{1}{2} f(y(\infty)) \int_t^\infty (\tau - t)\phi(\tau)\, d\tau,$$

and hence $\int_{t_0}^\infty \tau\phi(\tau)\, d\tau < \infty$.

THEOREM 2B. *Let I and J be as in problem* (B), *$t_0 \in I \cap J$, and $f : J \to (0, \infty)$ and $\phi : I \to [0, \infty)$ be continuous. Suppose $\lim_{d\to 0^+} \beta(d) > 0$. If problem* (B) *has a solution then*

$$\int_0^{t_0} f(t)\phi(t)\, dt = \infty, \qquad \int_0^{t_0} t\phi(t)\, dt < \infty. \tag{15}$$

Remark. It can be shown that if $\int_0^{t_0} t\phi(t)\, dt < \infty$ then for each $y_0 \in J$ there is $t_1 \in I$ such that for each $t_0 \in (0, t_1)$ there is a solution of $y'' + \phi(t)f(y) = 0$, $\lim_{t\to 0^+} y(t) = 0$, $y(t_0) = y_0$. However (15) does not guarantee that $\lim_{t\to 0^+} y'(t) = \infty$.

Proof. Since $\lim_{d\to 0^+} \beta(d) > 0$ there exists $d_0 > 0$ such that $\beta(d_0) > 0$. Hence, for some $y_1 \in J$ we have for $0 < \alpha < d_0$ and $0 < y \leqq y_1$ that

$$f(\alpha y) > \frac{\beta(d_0)}{2} f(y). \tag{16}$$

Suppose $y(t)$ is a solution of problem (B). Then there exists $t_1 \in (0, t_0)$ such that for $0 < t \leqq t_1$ we have that $y(t)$ is defined, $y(t) \leqq y_1$, and $d_0 y(t) > t$. Thus, for each $t \in (0, t_1]$, there is an $\alpha \in (0, d_0)$ such that $\alpha y(t) = t$. Hence, for $0 < t \leqq t_1$, we have by (16) that $f(t) > (\beta(d_0)/2)f(y(t))$. Integrating problem (B) from 0 to t_1, we obtain

$$\infty = -\int_0^{t_1} y''(t)\, dt = \int_0^{t_1} \phi(t)f(y(t))\, dt \leqq \frac{2}{\beta(d_0)} \int_0^{t_1} \phi(t)f(t)\, dt,$$

and hence $\int_0^{t_0} \phi(t)f(t)\, dt = \infty$.

Next, integrating problem (B) twice, we get

$$\int_t^{t_1} (\tau - t)\phi(\tau)f(y(\tau))\, d\tau = y(t_1) - y(t) - (t_1 - t)y'(t_1), \tag{17}$$

and letting $t \to 0^+$ we obtain by the monotone convergence theorem that

$$\int_0^{t_1} \tau\phi(\tau)f(y(\tau))\,d\tau = y(t_1) - t_1 y'(t_1) < \infty. \tag{18}$$

Putting $y = y_1$ in (16) shows $f(y)$ is bounded away from zero on $0 < y \leqq y_1$. Thus from (18) we get $\int_0^{t_0} \tau\phi(\tau)\,d\tau < \infty$.

THEOREM 2C. *Let I and J be as in problem* (C), $t_0 \in I \cap J$, *and* $f : J \to (0, \infty)$ *and* $\phi : I \to [0, \infty)$ *be continuous. Suppose* $\lim_{d\to\infty} \gamma(d) < \infty$. *If problem* (C) *has a solution then*

$$\int_{t_0}^{\infty} t\phi(t)\,dt = \infty, \qquad \int_{t_0}^{\infty} f(t)\phi(t)\,dt < \infty.$$

Proof. Since $\lim_{d\to\infty} \gamma(d) < \infty$ there exists $d_0 > 0$ such that $\gamma(d_0) < \infty$. Hence, for some $y_1 \in J$, we have for $\alpha > d_0$ and $y \geqq y_1$ that

$$f(\alpha y) < (\gamma(d_0) + 1)f(y). \tag{19}$$

Suppose $y(t)$ is a solution of problem (C). Then there exists $t_1 > t_0$ such that for $t > t_1$ we have $y(t)$ is defined, $y(t) \geqq y_1$, and $d_0 y(t) < t$. Thus, for each $t \geqq t_1$ there is an $\alpha > d_0$ such that $\alpha y(t) = t$. Hence, for $t \geqq t_1$, we have by (19) that $f(t) < (\gamma(d_0) + 1)f(y(t))$. Integrating problem (C) from t_1 to ∞ we obtain

$$y'(t_1) = -\int_{t_1}^{\infty} y''(t)\,dt = \int_{t_1}^{\infty} \phi(t)f(y(t))\,dt \geqq \frac{1}{\gamma(d_0)+1} \int_{t_1}^{\infty} \phi(t)f(t)\,dt,$$

and hence $\int_{t_0}^{\infty} \phi(t)f(t)\,dt < \infty$.

Next, integrating $y'(t) = \int_t^{\infty} \phi(\tau)f(y(\tau))\,d\tau$ from t_1 to ∞ and interchanging the order of integration we get

$$\infty = \int_{t_1}^{\infty} (\tau - t_1)\phi(\tau)f(y(\tau))\,d\tau. \tag{20}$$

Putting $y = y_1$ in (19) shows $f(y)$ is bounded on $y \geqq y_1$. Thus from (20) we have $\int_{t_0}^{\infty} \tau\phi(\tau)\,d\tau = \infty$.

3. The following three theorems are needed for the proof of Theorems 1A, 1B and 1C, but, as we will see in § 5, they also can be used to determine the asymptotic behavior of solutions of problems (A), (B) and (C) when Theorems 1A, 1B and 1C fail.

THEOREM 3A. *Let $c > 0$. Let I and J be as in problem* (A). *Suppose* $f : J \to (0, \infty)$ *and* $\phi, \Phi : I \to [0, \infty)$ *are continuous, and* $c\Phi(t) \leqq \phi(t)$ *for* $t \in I$. *Let* $y, Y : I \to J$ *be solutions of* $y'' = \phi(t)f(y)$ *and* $Y'' = \Phi(t)f(Y)$ *respectively, which also satisfy the initial conditions of problem* (A). *Then*

$$A^{-1}(c) \leqq \varliminf_{t\to\infty} \frac{Y(t)}{y(t)}.$$

THEOREM 3B. *Let $c > 0$. Let I and J be as in problem* (B). *Suppose* $f : J \to (0, \infty)$ *and* $\phi, \Phi : I \to [0, \infty)$ *are continuous, and* $c\Phi(t) \geqq \phi(t)$ *for* $t \in I$. *Let* $y, Y : I \to J$ *be solutions of*

$$y'' + \phi(t)f(y) = 0, \qquad \lim_{t\to 0^+} y(t) = 0$$

and

$$Y''+\Phi(t)f(Y)=0, \qquad \lim_{t\to 0^+} Y(t)=0,$$

respectively. Suppose $\lim_{d\to 0^+}\beta(d)>0$. *If* $\lim_{t\to 0^+} y'(t)=\infty$ *then* $\lim_{t\to 0^+} Y'(t)=\infty$ *and*

$$B^{-1}(c)\leqq \underline{\lim}_{t\to 0^+}\frac{Y(t)}{y(t)}. \tag{21}$$

THEOREM 3C. *Let* $c>0$. *Let* I *and* J *be as in problem* (C). *Suppose* $f:J\to(0,\infty)$ *and* $\phi, \Phi: I\to[0,\infty)$ *are continuous, and* $c\Phi(t)\leqq\phi(t)$ *for* $t\in I$. *Let* $y, Y: I\to J$ *be solutions of*

$$y''+\phi(t)f(y)=0, \qquad \lim_{t\to\infty} y'(t)=0,$$

$$Y''+\Phi(t)f(Y)=0, \qquad \lim_{t\to\infty} Y'(t)=0,$$

respectively. Suppose $\lim_{d\to\infty}\gamma(d)<\infty$. *If* $\lim_{t\to\infty} Y(t)=\infty$, *then* $\lim_{t\to\infty} y(t)=\infty$ *and*

$$C^{-1}(c)\geqq \overline{\lim}_{t\to\infty}\frac{Y(t)}{y(t)}. \tag{22}$$

LEMMA 1. *Let* I, $J\subset(0,\infty)$ *be open intervals and* ϕ, $\Phi: I\to[0,\infty)$, y, $Y: I\to J$, $f: J\to R-\{0\}$ *be continuous functions with* y *and* Y *twice continuously differentiable. Let* $t_0\in I$, $s=\int_{t_0}^{t} d\tau/y^2(\tau)$, *and* $v(s)=Y(t)/y(t)$. *If* $Y''(t)=\Phi(t)f(Y(t))$ *and* $y''(t)=\phi(t)f(y(t))$ *then*

$$v''(s)=y^3(t)v(s)f(y(t))\left[\Phi(t)\frac{f(v(s)y(t))}{v(s)f(y(t))}-\phi(t)\right]. \tag{23}$$

Proof. Use the chain rule.

Proof of Theorem 3A. Let $d>0$ be such that $A(d)<c$. Let s and $v(s)$ be as in Lemma 1. Since $y(t)\to 0$ as $t\to\infty$ we have $s\to\infty$ as $t\to\infty$ and there is a subinterval I' of I, with endpoint ∞, such that for $t\in I'$ and $v(s)<d$ we have

$$\Phi(t)\frac{f(v(s)y(t))}{v(s)f(y(t))}\leqq\Phi(t)\frac{c+A(d)}{2}\leqq\Phi(t)c\leqq\phi(t), \tag{24}$$

and thus, by Lemma 1, $v''(s)\leqq 0$.

To prove the theorem, it suffices to show $L\geqq d$, where $L=\underline{\lim}_{s\to\infty} v(s)$. Suppose, to the contrary, $L<d$. We claim $\overline{\lim}_{s\to\infty} v(s)<d$. For otherwise $v(s)$ would cross the horizontal line $v=(d+I)/2$ with nonpositive slope for arbitrarily large values of s; and since $v''(s)\leqq 0$ for $v(s)<d$ and $t\in I'$, we would have $\overline{\lim}_{s\to\infty} v(s)\leqq(d+I)/2$. Thus $\overline{\lim}_{s\to\infty} v(s)<d$, and hence $0<v(s)<d$ for sufficiently large t. Therefore, for sufficiently large t, $v''(s)\leqq 0$ and $v'(s)\geqq 0$. So $\lim_{s\to\infty} v(s)=L$ and $0<L<d$.

Next we claim, for sufficiently large t, that $Y''(t)\leqq[(\gamma+1)/2]Ly''(t)$, where $\gamma=(c+A(d))/2c<1$. This is clearly true if $Y''(t)=0$. We assume $Y''(t)\neq 0$, and thus $\Phi(t)\neq 0$, $\phi(t)\neq 0$ and $y''(t)\neq 0$. For sufficiently large t we have by (24) that

$$c\frac{Y''(t)}{y''(t)}\leqq\frac{\phi(t)}{\Phi(t)}\frac{Y''(t)}{y''(t)}=\frac{f(Y(t))}{f(y(t))}=\frac{f(v(s)y(t))}{f(y(t))}\leqq\frac{c+A(d)}{2}v(s).$$

Since $\lim_{s\to\infty} v(s) = L$ the above claim is established and for sufficiently large t we have

$$Y(t) = \int_t^\infty (\tau - t) Y''(\tau)\, d\tau \leqq \frac{\gamma+1}{2} L \int_t^\infty (\tau - t) y''(\tau)\, d\tau = \frac{\gamma+1}{2} L y(t).$$

Thus,

$$L = \lim_{s\to\infty} v(s) = \lim_{t\to\infty} \frac{Y(t)}{y(t)} \leqq \frac{\gamma+1}{2} L < L,$$

a contradiction which establishes the theorem.

Proof of Theorem 3B. To prove (21), suppose $d > 0$ and $B(d) > c$. Let s and $v(s)$ be as in Lemma 1. Since $y(t) \to 0$ as $t \to 0^+$, there is a subinterval I' of I, with one endpoint zero, such that for $t \in I'$ and $v(s) < d$ we have

$$\Phi(t) \frac{f(v(s)y(t))}{v(s) f(y(t))} \geqq \Phi(t) \frac{B(d)+c}{2} \geqq c\Phi(t) \geqq \phi(t), \tag{25}$$

and thus, by Lemma 1, $v''(s) \leqq 0$.

To prove (21) it suffices to show $L \geqq d$, where $L = \underline{\lim}_{t\to 0^+} v(s)$. Suppose to the contrary $L < d$. Then, as in the proof of Theorem 3A, we have $\overline{\lim}_{t\to 0^+} v(s) < d$ and hence $0 < v(s) < d$ for sufficiently small t. Therefore for sufficiently small t, $v''(s) \leqq 0$. So $\lim_{t\to 0^+} v(s) = L$ and $0 \leqq L < d$. (In contrast to the proof of Theorem 1A, we can't conclude for small t that $v'(s) < 0$ and $L \neq 0$, because s may not tend to $-\infty$ as $t \to 0^+$.)

Also, for $y''(t) \neq 0$, we have $\phi(t) \neq 0$, $\Phi(t) \neq 0$, $Y(t) \neq 0$, and, by (25),

$$c \frac{Y''(t)}{y''(t)} \geqq \frac{\phi(t)}{\Phi(t)} \frac{Y''(t)}{y''(t)} = \frac{f(Y(t))}{f(y(t))} = \frac{f(v(s)y(t))}{f(y(t))} \geqq \frac{B(d)+c}{2} \frac{Y(t)}{y(t)} \tag{26}$$

for sufficiently small t.

We claim $L > 0$. To see this, let $M = \lim_{d\to 0^+} \beta(d) > 0$ and suppose $L = 0$. Then by (26), for sufficiently small t, we have

$$Y''(t) \leqq \frac{1}{c} \frac{f(v(s)y(t))}{f(y(t))} y''(t) \leqq \frac{1}{c} \frac{M}{2} y''(t) \leqq 0, \tag{27}$$

which is clearly valid even if $y''(t) = 0$. Since $\lim_{t\to 0^+} y'(t) = \infty$ we have from (27) that $\lim_{t\to 0^+} Y'(t) = \infty$ and for sufficiently small values of t that $Y'(t) \geqq (1/c)(M/4) y'(t) > 0$. Since $\lim_{t\to 0^+} Y(t) = \lim_{t\to 0^+} y(t) = 0$ we have, for sufficiently small t, that $Y(t) \geqq (1/c)(M/8) y(t)$, which contradicts $L = 0$. Hence $L > 0$.

From (26) we have, for sufficiently small t,

$$Y''(t) \leqq \gamma \frac{Y(t)}{y(t)} y''(t) \leqq \frac{\gamma+1}{2} L y''(t) \leqq 0, \tag{28}$$

where $\gamma = (B(d)+c)/2c > 1$. Since $\lim_{t\to 0^+} y'(t) = \infty$, we conclude from (28), as in the preceding paragraph, that $\lim_{t\to 0^+} Y(t)/y(t) > L$, a contradiction. Hence $L \geqq d$, and (21) is established. Since $\lim_{d\to 0^+} \beta(d) > 0$ implies $B^{-1}(c) > 0$ we have by (21) that $\lim_{t\to 0^+} Y'(t) = \infty$.

Proof of Theorem 3C. Suppose $\lim_{t\to\infty} Y(t) = \infty$. Let $t_0 \in I$. Then by Theorem 2C we have $\int_{t_0}^\infty t\Phi(t)\, dt = \infty$, and since $c\Phi(t) \leqq \phi(t)$ we have $\int_{t_0}^\infty t\phi(t)\, dt = \infty$. Suppose $\lim_{t\to\infty} y(t) = y_\infty < \infty$. Then, for large t, $f(y(t)) > \frac{1}{2} f(y_\infty) > 0$ and

$$y'(t) = \int_t^\infty \phi(\tau) f(y(\tau))\, d\tau \geqq \frac{1}{2} f(y_\infty) \int_t^\infty \phi(\tau)\, d\tau. \tag{29}$$

Integrating (29) we obtain, for large t, that

$$y_\infty - y(t) \geqq \frac{1}{2} f(y_\infty) \int_t^\infty (\tau - t)\phi(\tau)\, d\tau,$$

a contradiction. Thus $\lim_{t\to\infty} y(t) = \infty$.

The rest of the proof of Theorem 3C now proceeds like the proof of Theorem 3B and will be omitted.

4. In this section we prove Theorems 1A, 1B.2 and 1C. The proof of Theorem 1B.1, which is similar to the proof of Theorem 1C, will be omitted.

Proof of Theorem 1A. As in the proof of Theorem 2A, $A(d) < 1$ for some $d > 0$ implies $f(y) < y$ for sufficiently small positive y. Hence $F(\xi) \to \infty$ as $\xi \to 0^+$, and $F^{-1}(\zeta)$ is defined, positive, decreasing and twice continuously differentiable for $\zeta \geqq 0$. For $\zeta \geqq 0$ we have

$$F^{-1\prime}(\zeta) = -\sqrt{f(F^{-1}(\zeta))}\sqrt{F^{-1}(\zeta)}$$

and

$$f(F^{-1}(\zeta)) = H'(\zeta)^2\, e^{H(\zeta)}. \tag{30}$$

For $t \geqq t_0$, let

$$\bar{\psi}(t) = \left[1 - \left(\frac{1}{H'(I(t))}\,\frac{1}{\sqrt{\psi(t)}}\right)'\right]\psi(t) \tag{31}$$

and

$$\bar{z}(t) = F^{-1}(I(t)) = e^{H(I(t))}. \tag{32}$$

For $t \geqq t_0$ we have

$$\frac{\bar{z}'^2}{\bar{z}} = H'(I(t))^2 \psi(t)\, e^{H(I(t))} \tag{33}$$

and, by (30), (31), (32) and (33),

$$\begin{aligned} \bar{\psi}(t) f(\bar{z}(t)) &= \bar{\psi}(t) f(F^{-1}(I(t))) \\ &= \bar{\psi}(t) H'(I(t))^2\, e^{H(I(t))} \\ &= \left[1 - \left(\frac{\bar{z}(t)}{\bar{z}'(t)}\right)'\right]\frac{(\bar{z}')^2}{\bar{z}} \\ &= \bar{z}''(t). \end{aligned} \tag{34}$$

Let $y(t)$ be a solution of problem (A). Let γ_1, γ_2, γ_3 and γ_4 be constants with $0 < \gamma_1 < c_1$, $c_2 < \gamma_2 < \infty$, $0 < \gamma_3 < c_3$ and $c_4 < \gamma_4 < \infty$. By (3) and (4) there is $t_1 > t_0$ such that for $t \geqq t_1$ we have $y(t)$ and $\bar{z}(t)$ are defined and

$$\gamma_1\gamma_3\phi(t) \leqq \bar{\psi}(t) \leqq \gamma_2\gamma_4\phi(t). \tag{35}$$

By Theorem 2A we have $\int_{t_0}^\infty t\phi(t)\, dt = \infty$, and hence by (35) we have $\int_{t_0}^\infty t\bar{\psi}(t)\, dt = \infty$. Thus, since $\bar{z}(t)$ is decreasing, we have by Theorem 2A and (34) that $\bar{z}(t)$ is a solution of problem (A) with $\phi(t)$ replaced with $\bar{\psi}(t)$.

By Theorem 3A and (35) we have

$$A^{-1}(\gamma_1\gamma_3) \leqq \underline{\lim}_{t\to\infty} \frac{y(t)}{\bar{z}(t)} \leqq \overline{\lim}_{t\to\infty} \frac{y(t)}{\bar{z}(t)} \leqq \frac{1}{A^{-1}(1/\gamma_2\gamma_4)}.$$

Letting $\gamma_i \to c_i$, $i = 1, 2, 3, 4$, establishes Theorem 1A.

Proof of Theorem 1B.2. Let

$$\bar{\psi}(t) = \left(1 + \frac{t\bar{g}''}{2\bar{g}'}\right)\psi(t) \quad \text{and} \quad \bar{z}(t) = t\bar{g}(t).$$

Then using (9), we have

$$\bar{z}''(t) = 2\bar{g}'(t)\left(1 + \frac{t\bar{g}''(t)}{2\bar{g}'(t)}\right) = -\psi(t)f(t\bar{g}(t))\frac{\bar{\psi}(t)}{\psi(t)} = -\bar{\psi}(t)f(\bar{z}(t)). \tag{36}$$

By (9), $\bar{g}'(t)$ is negative; hence, by (8), for some $c \in (0, \frac{1}{2})$, we have for sufficiently small t that

$$t\bar{g}'' + 2(1-c)\bar{g}' < 0. \tag{37}$$

Solving (37) for $\bar{g}$ gives, for some $t_0 \in I$,

$$\bar{g}(t) < \bar{g}(t_0) + \frac{t_0^{2(1-c)}\bar{g}'(t_0)}{1-2c}(t_0^{2c-1} - t^{2c-1})$$

for $0 < t \leqq t_0$. Hence $\bar{z}(t) \to 0$ as $t \to 0^+$.

Let $y(t)$ be a solution of problem (B). Let γ_1, γ_2, γ_3, and γ_4 be constants with $0 < \gamma_1 < c_1$, $c_2 < \gamma_2 < \infty$, $0 < \gamma_3 < c_3$, and $c_4 < \gamma_4 < \infty$. By (7) and (8) there is $t_1 > 0$ such that for $0 < t \leqq t_1$ we have $y(t)$ and $\bar{z}(t)$ are defined and

$$\gamma_1\gamma_3\phi(t) \leqq \bar{\psi}(t) \leqq \gamma_2\gamma_4\phi(t). \tag{38}$$

From (38) and Theorem 3B we obtain

$$B^{-1}(\gamma_2\gamma_4) \leqq \underline{\lim}_{t\to 0^+} \frac{y(t)}{t\bar{g}(t)} \leqq \overline{\lim}_{t\to 0^+} \frac{y(t)}{t\bar{g}(t)} \leqq \frac{1}{B^{-1}(1/\gamma_1\gamma_3)}.$$

Letting $\gamma_i \to c_i$, $i = 1, 2, 3, 4$ completes the proof of Theorem 1B.2.

Proof of Theorem 1C. Since $\lim_{d\to\infty} \gamma(d) < \infty$ we have $f(y)$ is bounded on $y_0 \leqq y < \infty$; and hence $F(\xi) \to \infty$ as $\xi \to \infty$. So $F^{-1}(\zeta)$ is defined, positive, increasing, and twice continuously differentiable for $\zeta \geqq 0$. For $\zeta \geqq 0$ we have

$$F^{-1\prime}(\zeta) = \sqrt{f(F^{-1}(\zeta))}\sqrt{F^{-1}(\zeta)}$$

and

$$f(F^{-1}(\zeta)) = \tfrac{1}{4}[H'(\zeta)]^2[H(\zeta)]^{-3/2}. \tag{39}$$

For $t \geqq t_0$, let

$$\bar{\psi}(t) = \left[1 + 2H(I(t))\left(\frac{1}{H'(I(t))\sqrt{\psi(t)}}\right)'\right]\psi(t) \tag{40}$$

and

$$\bar{z}(t) = F^{-1}(I(t)) = \sqrt{H(I(t))}. \tag{41}$$

For $t \geqq t_0$ we have

$$\frac{\bar{z}'^2}{\bar{z}} = \tfrac{1}{4}H'(I(t))^2 H(I(t))^{-3/2}\psi(t), \tag{42}$$

and by (39), (40), (41) and (42),

$$\begin{aligned} \bar{\psi}(t)f(\bar{z}(t)) &= \bar{\psi}(t)f(F^{-1}(I(t))) \\ &= \bar{\psi}(t)\tfrac{1}{4}H'(I(t))^2 H(I(t))^{-3/2} \\ &= \left[1+2\bar{z}^2(t)\left(\frac{1}{2\bar{z}(t)\bar{z}'(t)}\right)'\right]\frac{\bar{z}'(t)^2}{\bar{z}(t)} \\ &= -\bar{z}''(t). \end{aligned} \tag{43}$$

Let $h(t)=H(I(t))$. By (11), for some $c>0$ and for sufficiently large t, we have

$$\left(\frac{1}{h'(t)}\right)' > \frac{c-1}{2h(t)}. \tag{44}$$

Multiplying (44) by $h'(t)$ and integrating twice we obtain for sufficiently large t that $h(t)<k^2 t^{2/(c+1)}$ for some $k>0$; and hence $\bar{z}(t)<kt^{1/(c+1)}$ for large t. Since $c>0$ and $\bar{z}''(t)<0$ we have $\bar{z}'(t)\to 0$ as $t\to\infty$.

Let $y(t)$ be a solution of Problem (C). Let γ_1, γ_2, γ_3, and γ_4 be constants with $0<\gamma_1<c_1$, $c_2<\gamma_2<\infty$, $0<\gamma_3<c_3$ and $c_4<\gamma_4<\infty$. Then by (10) and (11) there is $t_1>t_0$ such that for $t\geqq t_1$ we have $y(t)$ and $\bar{z}(t)$ are defined and $\gamma_1\gamma_3\phi(t)<\bar{\psi}(t)<\gamma_2\gamma_4\phi(t)$. Hence, by Theorem 3C, we have

$$C^{-1}(\gamma_1\gamma_3) \geqq \overline{\lim_{t\to\infty}}\frac{y(t)}{\bar{z}(t)} \geqq \underline{\lim_{t\to\infty}}\frac{y(t)}{\bar{z}(t)} \geqq \frac{1}{C^{-1}(1/\gamma_2\gamma_4)}.$$

Letting $\gamma_i \to c_i$, $i=1, 2, 3, 4$, we obtain Theorem 1C.

5. In this section we give some indication, by way of examples, that when the conditions

(i) $A(d)<1$ for some $d>0$,
(ii) $\lim_{d\to 0^+}\beta(d)>0$,
(iii) $\lim_{d\to\infty}\gamma(d)<\infty$

are satisfied the results of this paper are strong enough to give the asymptotic behavior of solutions of problems (A), (B) and (C) respectively, when such solutions exist.

Example 1. Consider the case $f(y)=y^\lambda$ and $\phi(t)=t^\sigma$, which includes problems (1), (2) and the Emden–Fowler equation. Then (i) holds if and only if $\lambda>1$, and (ii), ((iii)), hold if and only if $\lambda\leqq 0$. And in this case, by § 2, problem (A) has a solution if and only if $\sigma\geqq -2$; problem (B) has a solution only if $-2<\sigma\leqq -\lambda-1$; and problem (C) has a solution only if $-2\leqq\sigma<-\lambda-1$. When these conditions on σ and λ hold, the theorems of sections 1 and 3 give explicit asymptotic formulas for the solutions. A listing of these formulas can be found in [6]–[8].

Example 2. Consider the problem

$$y'' = k_3\, e^{k_1 t^\sigma}\, e^{-k_2 y^{-\lambda}}, \qquad y(\infty)=y'(\infty)=0, \tag{45}$$

which is a special case of problem (A) with $f(y)=e^{-k_2 y^{-\lambda}}$ and $\phi(t)=k_3\, e^{k_1 t^\sigma}$. We assume

λ and k_2 are such that (i) holds. Hence $k_2>0$, $\lambda>0$ and

$$A(d)=\begin{cases}\infty & \text{if } d>1,\\ 1 & \text{if } d=1,\\ 0 & \text{if } 0<d<1,\end{cases}$$

and so $A^{-1}(c)=1$ for all $c>0$.

We also assume $k_3>0$. Since $A^{-1}(c)\equiv 1$, we have by Theorem 3A that the asymptotic behavior of solutions of (45) is not affected by k_3. Hence we can assume $k_3=1$.

Now, as $\xi\to 0^+$,

$$F(\xi)=\int_{\sqrt{\xi}}^{\sqrt{y_0}}\frac{2\,dx}{\sqrt{f(x^2)}}=\frac{1}{\lambda}\left(\frac{k_2}{2}\right)^{1/2\lambda}\int_{k_2/2y_0^\lambda}^{k_2/2\xi^\lambda}\zeta^{-(1/2\lambda)-1}e^{\zeta}\,d\zeta$$

$$\sim\frac{1}{\lambda}\left(\frac{k_2}{2}\right)^{1/2\lambda}\left(\frac{k_2}{2\xi^\lambda}\right)^{-1/(2\lambda)-1}e^{k_2/2\xi^\lambda}=\frac{2}{\lambda k_2}\xi^{(2\lambda+1)/2}e^{k_2/2\xi^\lambda}$$

and hence

$$F^{-1}(\zeta)\sim\left(\frac{k_2}{2\log\zeta}\right)^{1/\lambda}$$

as $\zeta\to\infty$.

Let $y(t)$ be a solution of (45). Then, by Theorem 2A, $\int_1^\infty t\phi(t)\,dt=\infty$ and therefore we have only the following 2 cases to consider:

Case I. Suppose $\sigma>0$ and $k_1>0$. Then

$$I(t)=\int_1^t\sqrt{\phi(\tau)}\,d\tau=\frac{1}{\sigma}\left(\frac{2}{k_1}\right)^{1/\sigma}\int_{(1/2)k_1}^{(1/2)k_1t^\sigma}u^{(1-\sigma/\sigma)}e^u\,du$$

$$\sim\frac{1}{\sigma}\left(\frac{2}{k_1}\right)^{1/\sigma}\left(\frac{1}{2}k_1t^\sigma\right)^{(1-\sigma)/\sigma}e^{(1/2)k_1t^\sigma}=\frac{1}{\sigma}\frac{2}{k_1}t^{1-\sigma}e^{(1/2)k_1t^\sigma},$$

$\log I(t)\sim\frac{1}{2}k_1t^\sigma$, and $F^{-1}(I(t))\sim(k_2/k_1)^{1/\lambda}t^{-\sigma/\lambda}$ as $t\to\infty$. Also

$$\lim_{t\to\infty}1-\left(\frac{1}{\sqrt{\phi(t)}H'(I(t))}\right)'=1+\frac{\lambda}{\sigma}.$$

Hence, by Theorem 1A,

$$y(t)\sim\left(\frac{k_2}{k_1}\right)^{1/\lambda}t^{-\sigma/\lambda}\quad\text{as } t\to\infty.$$

Case II. Suppose $k_1=0$ or $\sigma\leqq 0$. Then $\phi(t)$ approaches a positive constant as $t\to\infty$, and, since $A^{-1}(c)\equiv 1$, we have by Theorem 3A that the asymptotic behavior of $y(t)$ is not affected by the value of this positive constant and we can assume $\phi(t)\equiv 1$. Using the notation of Theorem 1A we have $F^{-1}(I(t))\sim(k_2/2\log t)^{1/\lambda}$, but unfortunately $\lim_{t\to\infty}1-(1/\sqrt{\phi(t)}H'(I(t)))=\infty$. However we still have $y(t)\sim F^{-1}(I(t))$ as $t\to\infty$. To see this, let

$$z(t)=\left(\frac{2}{k_2}\log t+\frac{\lambda+1}{\lambda k_2}\log\log t\right)^{-1/\lambda}.$$

Then

$$e^{k_2 z^{-\lambda}} z'' \to \frac{1}{\lambda}\left(\frac{k_2}{2}\right)^{1/\lambda} \quad \text{as } t \to \infty,$$

and hence, by Theorem 3A, we have

$$y(t) \sim z(t) \sim \left(\frac{k_2}{2 \log t}\right)^{1/\lambda} \quad \text{as } t \to \infty.$$

REFERENCES

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

[2] E. FERMI, *Un metodo statistico par la determinazione di alcune proprietà dell' atome*, Rend. Accad. Naz. Lincei, Cl. Sci. Fis. Mat. Nat., 6 (1927), pp. 602–607.

[3] P. HARTMAN AND A. WINTNER, *On the non-increasing solutions of* $y'' = f(x, y, y')$, American J. Math., 73 (1951), pp. 390–404.

[4] V. MARIĆ AND M. TOMIĆ, *Asymptotics of solutions of a generalized Thomas–Fermi equation*, J. Differential Equations, 35 (1980), pp. 36–44.

[5] A. NACHMAN AND S. TALIAFERRO, *Mass transfer into boundary layers for power law fluids*, Proc. R. Soc. Lond. A., 365 (1979), pp. 313–326.

[6] S. TALIAFERRO, *Asymptotic behavior of solutions of* $y'' = \phi(t)y^{\lambda}$, J. Math. Anal. Appl., 66 (1978), pp. 95–134.

[7] ———, *A nonlinear singular boundary value problem*, Nonlinear Analysis, 3 (1979), pp. 897–904.

[8] ———, *On the positive solutions of* $y'' + \phi(t)y^{-\lambda} = 0$, Nonlinear Analysis, 2 (1978), pp. 437–446.

[9] L. THOMAS, *The calculation of atomic fields*, Proc. Cambridge Philos. Soc., 13 (1927), pp. 542–548.

[10] P. WONG, *Existence and asymptotic behavior of proper solutions of a class of second-order non-linear differential equations*, Pacific J. Math., 13 (1963), pp. 737–760.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0007 $01.00/0

EXISTENCE, OSCILLATION AND EIGENVALUE COMPARISON THEOREMS FOR TWO- AND THREE-POINT FOURTH ORDER PROBLEMS*

JOSEPH DiGIALLONARDO†

Abstract. Oscillation theory has produced results for boundary value problems. We generalize some of the well-known oscillation theory of Leighton and Nehari in order to obtain existence, oscillation and eigenvalue comparison theorems in two- and three-point problems. Our method produces results in nonself-adjoint cases.

In our development a certain relation emerges which is essential for oscillation results in the case of eigenfunctions and hence for existence and eigenvalue comparison.

We also point out generalizations to the present development.

1. Introduction and preliminary considerations. In this paper we will consider the eigenvalue problems

$$(r(x)y'')'' - lp(x)y = 0, \qquad l > 0; \tag{1.1}$$

let

$$L(y(x)) = \gamma_1 y(x) + \gamma_2 y'(x) + \gamma_3 (ry'')(x) + \gamma_4 (ry'')'(x),$$

$$M(y(x)) = \beta_1 y(x) + \beta_2 y'(x) + \beta_3 (ry'')(x) + \beta_4 (ry'')'(x),$$

$$N(y(x)) = \eta_1 y(x) - \eta_2 y'(x) + \eta_3 (ry'')(x) - \eta_4 (ry'')'(x),$$

$$P(y(x)) = \alpha_1 y(x) + (\alpha_1 + \alpha_2) y'(x) + (\alpha_2 + \delta_1)(ry'')(x) + \delta_2 (ry'')'(x);$$

$$y(a) = y'(a) = \alpha_1 y(b) + \alpha_2 y'(b) = L(y(c)) = 0; \tag{1.2}$$

$$y(a) = y'(a) = \alpha_1 (ry'')(b) + \alpha_2 (ry'')'(b) = L(y(c)) = 0; \tag{1.3}$$

$$(ry'')(a) = (ry'')'(a) = \alpha_1 (ry'')(b) + \alpha_2 (ry'')'(b) = L(y(c)) = 0; \tag{1.4}$$

$$y(a) = y'(a) = y(b) = y'(b) = 0; \tag{1.5}$$

$$y(a) = y'(a) = (ry'')(b) = (ry'')'(b) = 0; \tag{1.6}$$

$$(ry'')(a) = (ry'')'(a) = (ry'')(b) = (ry'')'(b) = 0; \tag{1.7}$$

$$N(y(a)) = M(y(a)) = \alpha_1 y(b) + \alpha_2 y'(b) = \alpha_1 y'(b) + \alpha_2 y''(b) = 0; \tag{1.8}$$

$$N(y(a)) = M(y(a)) = \alpha_1 y(b) + \alpha_2 y'(b) = P(y(c)) = 0; \qquad 0 < a < b < c < \infty. \tag{1.9}$$

Three-point problems of this kind have been considered in a third order case [7]. In [1], the method employed by the authors does not include (1.1)–(1.7). Also the problem (1.1)–(1.8) generalizes, in some respects, that in [1].

Our methods enable us to extend a result of Barrett's [2]. Generally speaking, our methods use the two-point problem to obtain information about the three point problem.

In § 5 we point out generalizations and extensions of the present development.

In all that follows we let $r(x) > 0$, $p(x) > 0$, $r(x) \in C^{(2)}$, and $p(x) \in C$, all on $(0, \infty)$; also $l > 0$.

* Received by the editors March 15, 1978, and in final revised form January 12, 1981.

† 1 Mission Street, Gardner, Massachusetts 01440.

Consider the equation

$$[(r_2+e(r_1-r_2))y'']''-l[P_1+h(P_2-P_1)]y=0. \tag{1.10}$$

We let $r_1(x)$ and $r_2(x)$ belong to class $C^{(2)}$ on $(0,\infty)$ and $P_1(x)$, $P_2(x)$ belong to class C on $(0,\infty)$. Also $r_1(x)\geqq r_2(x)>0$ and $P_2(x)\geqq P_1(x)>0$ on $(0,\infty)$ and $e\geqq 0$, $h\geqq 0$, $l>0$.

Let $y_1(x,e,h,l)$, $y_2(x,e,h,l)$, $y_3(x,e,h,l)$ and $y_4(x,e,h,l)$ be the fundamental solutions of (1.10) which satisfy the initial conditions:

$$\begin{aligned} &y_1(a)=1, \qquad y_1'(a)=y_1''(a)=y_1'''(a)=0,\\ &y_2(a)=y_2''(a)=y_2'''(a)=0, \qquad y_2'(a)=1,\\ &y_3(a)=y_3'(a)=y_3'''(a)=0, \qquad y_3''(a)=1,\\ &y_4(a)=y_4'(a)=y_4''(a)=0, \qquad y_4'''(a)=1, \end{aligned}$$

for all (e,h,l), $e\geqq 0$, $h\geqq 0$, $l>0$. We know that y_1, y_2, y_3 and y_4 are analytic functions of e, h, l.

We form the function $(r(x)=r_2+e(r_1-r_2))$

$$F(x,e,h,l)=y_3(x,e,h,l)y_4'(x,e,h,l)-y_4(x,e,h,l)y_3'(x,e,h,l), \tag{1.11}$$

$$G(x,e,h,l)=(ry_3'')(x,e,h,l)(ry_4'')'(x,e,h,l)-(ry_4'')(x,e,h,l)(ry_3'')'(x,e,h,l), \tag{1.12}$$

$$H(x,e,h,l)=(ry_1'')(x,e,h,l)(ry_2'')'(x,e,h,l)-(ry_2'')(x,e,h,l)(ry_1'')'(x,e,h,l). \tag{1.13}$$

Let $y(x,e,h,l)$ be the general solution of (1.10) which satisfies $M(y(a))=N(y(a))=0$. Consider

$$\begin{aligned} &\alpha_1 y(x,e,h,l)+\alpha_2 y'(x,e,h,l)=0,\\ &\alpha_1 y'(x,e,h,l)+\alpha_2 y''(x,e,h,l)=0. \end{aligned} \tag{1.14}$$

Let $C_{ik}(e,h,l)$, $i=1,2,3,\cdots$, $k=1,2,3,4$ be the ith zeros of F, G, H and (1.14) (these points are denumerable by (1.15) and (1.16) below), respectively. C_{11} and C_{12} are considered in [2]. A part of the relationship between C_{11} and C_{12} in [2] will be shown to hold for C_{i1} and C_{i2}. However, our main purpose in introducing these points is to obtain information about our eigenvalue problems. C_{i1} is recognized to be the ith conjugate point of (1.10) (see [5]). When these points exist we have

$$F'(C_{i1},e,h,l),\ G'(C_{i2},e,h,l),\ H'(C_{i3},e,h,l)\neq 0, \tag{1.15}$$

$$\alpha_1 y''(C_{i4})+\alpha_2 y''(C_{i4})\neq 0. \tag{1.16}$$

(1.15) is obvious from theorems in [5]. We establish (1.16) since it generalizes material in [1], [2] and [5].

LEMMA 1.17. *Let $r(x)=1$ in (1.1). Let $\alpha_1>0$, $\alpha_2>0$. If $y(x)$ is a nontrivial solution of (1.1) which satisfies $\alpha_1 y(a)+\alpha_2 y'(a)=\alpha_1 y'(a)+\alpha_2 y''(a)=0$, and $y''(a)\geqq 0$, $y'''(a)\geqq 0$ (but not both zero), then $\alpha_1 y(x)+\alpha_2 y'(x)$, $\alpha_1 y'(x)+\alpha_2 y''(x)$, $y''(x)$ and $y'''(x)$ are greater than zero for $x>a$.*

Proof. We have

$$y(x)=y(a)+y'(a)(x-a)+\frac{y''(a)(x-a)^2}{2}+\frac{y'''(a)(x-a)^3}{6}$$

$$+l\int_a^x dt\int_t^x (u-t)(x-u)p(t)y(t)\,du.$$

If $y''(a)=0$, then [5, Lemma 2.1] proves the theorem. If $y''(a)>0$, then $y(a)>0$. Let c be the first zero of $y(x)$ to the right of a (assuming it exists). Clearly $\alpha_1 y(x)+\alpha_2 y'(x)$ and $\alpha_1 y'(x)+\alpha_2 y''(x)$ are greater than zero on $(a, c]$. It follows that $y'(c)>0$, a contradiction. Hence, we have the lemma.

LEMMA 1.18. *Let* $r(x)=1$ *in* (1.1). *Let* $\alpha_1>0$, $\alpha_2>0$. *Let* $y(x)$ *be a nontrivial solution of* (1.1) *which satisfies* $\alpha_1 y(a)+\alpha_2 y'(a)=\alpha_1 y'(a)+\alpha_2 y''(a)=0$, $y''(a)\geqq 0$ *and* $y'''(a)\leqq 0$, *then* $y(x)$ *and* $y''(x)$ *are positive, and* $y'(x)$ *and* $y'''(x)$ *are negative for* $0<x<a$.

Proof. Clearly, the condition $y''(a)\geqq 0$, implies $y'(a)\leqq 0$ and $y(a)\geqq 0$. Then the lemma is proved by [5, Lemma 2.2].

LEMMA 1.19. *If* $\alpha_1>0, \alpha_2>0, \eta_i\geqq 0, \sum_{i=1}^{4}\eta_i^2\neq 0, r_1=r_2=1$, *then* (1.16) *is true for* (1.10).

Proof. Assume $\alpha_1 y''(x)+\alpha_2 y'''(x)=0$. We may suppose $y'''(x)<0$. Then we have $y(a)>0$, $y''(a)>0$ and $y'(a)<0$, $y'''(a)<0$. This contradicts $N(y(a))=0$, and proves the lemma. $(x=c_{i4})$

Lemmas 1.17–1.19 can also be proved in the case $\alpha_1 y'+\alpha_2 y''$, $\alpha_1 y''+\alpha_2 y'''$. We leave the formulation and proof to the reader.

The implicit function theorem and (1.15) and (1.16) show that C_{ik}, $k=1, 2, 3, 4$ are in C' with respect to (e, h, l). Let $l_{ik}(x, e, h)$, $i=1, 2, 3, \cdots$, $k=1, 2, 3$ be the ith eigenvalue of (1.10)–(1.5), (1.6), (1.7), respectively. Then from the calculus of variations we have (see [1]) the following theorem.

THEOREM 1.20. *Let* (x', e', h') *and* (x'', e'', h'') *be two points such that* $x''>x'>a$, $0\leqq e''<e'$ *and* $h''>h'\geqq 0$. *Then*

$$l_{ik}(x'', e'', h'')<l_{ik}(x', e', h').$$

From (1.10) and the implicit function theorem, $l_{ik}(x, e, h)$ is in C' with respect to (x, e, h).

By methods in [5] we can obtain the following relations, (note also [1]):

$$C_{ik}(e, h, l_{ik}(b, e, h))=b, \text{ and } l_{ik}(C_{ik}(e, h, l), e, h)=l, \quad i=1, 2, 3, \cdots, \quad k=1, 2, 3. \tag{1.21}$$

These relations will be clarified considerably in § 2. The case $k=1$ is proved in [5]. We can also state the following theorem as fact, by [1] and [5].

THEOREM 1.22. *If the point* $C_{ik}(1, 0, l')$, $l'>0$, *of* (1.10) *exists, then* $C_{ik}(0, 1, l'')$, $l''>l'$, *exists and* $C_{ik}(0, 1, l'')<C_{ik}(1, 0, l')$. *The case* $C_{ik}(e, h, l'')<C_{ik}(e, h, l')$ *also holds for* $i=1, 2, 3, \cdots, k=1, 2, 3$.

2. Oscillation theorems. The following theorem generalizes [5, Thm. 3.6].

THEOREM 2.1. *Let* $r(x)=1$ *in* (1.1). *Let* $\alpha_1^2+\alpha_2^2\neq 0, \alpha_1\geqq 0, \alpha_2\geqq 0$. *Let* $\beta_4=-\beta_3\neq 0$, $\eta_4=\eta_3\neq 0$ *and* $\eta_i\geqq 0$, $i=1, 2, 3, 4$. *If* $\alpha_1>0$, $\alpha_2>0$, *we assume* $(\alpha_1/\alpha_2)(\eta_2+\eta_3\beta_3^{-1}\beta_2)+(\eta_1-\eta_3\beta_3^{-1}\beta_1)\neq 0$. *If* $y(x)$ *is a solution of* (1.1) *which satisfies* $M(y(a))=N(y(a))=0$; *and* $\alpha_1 y(x)+\alpha_2 y'(x)$ *has at least* $n+1$ *zeros in* (a, ∞), $n=1, 2, 3, \cdots$, *then there exist* n *points* $C_{14}, \cdots, C_{n4}$, $a<C_{14}<C_{24}<\cdots<C_{n4}$, *and* n *essentially unique solutions,* $y_1(x), \cdots, y_n(x)$ *of* (1.1) *with the following properties*:

(a) $y_i(x)$ *satisfies* $M(y_i(a))=N(y_i(a))=\alpha_1 y_i(C_{i4})+\alpha_2 y_i'(C_{i4})=\alpha_1 y_i'(C_{i4})+\alpha_2 y_i''(C_{i4})=0$;

(b) $\alpha_1 y_i(x)+\alpha_2 y_i'(x)$ *has precisely* $i+1$ *zeros in* $(a, C_{i4}]$ (*where the double zero is counted according to its multiplicity*);

(c) *for any other solution,* $y(x)$, *which satisfies* $M(y(a))=N(y(a))=0$, $\alpha_1 y(x)+\alpha_2 y'(x)$ *has fewer than* $i+1$ *zeros in* $(a, C_{i4}]$.

Proof. Let $A_1, A_2, \cdots, A_{n+1}$ be the $n+1$ zeros of $\alpha_1 y(x)+\alpha_2 y'(x)$ in (a, ∞). Then, by a compactness argument [5], we will determine a minimal A_{n+1} and a minimizing solution V_m. The condition $(\alpha_1/\alpha_2)(\eta_2+\eta_3\beta_3^{-1}\beta_2)+(\eta_1-\eta_3\beta_3^{-1}\beta_1)\neq 0$ is essential in this argument when $\alpha_1>0$, $\alpha_2>0$.

Consider a solution w of (1.1) which satisfies $w(a)=w'(a)=0$, $w''(a)=w'''(a)=1$. Then by [5, Lemma 1.2], $\alpha_1(V_m-Kw)+\alpha_2(V'_m-Kw)$ has a double zero α, in (A_n, A_{n+1}) (we are now considering minimal values). Without loss of generality, we assume $\alpha_1 V_m+\alpha_2 V'_m>0$ in (A_n, A_{n+1}). Then $f=K(\alpha_1 w+\alpha w')$ has at least two points of intersection with every positive arc located in (A_1, A_n). Indeed, since $[\alpha_1(V_m-Kw)+\alpha_2(V'_m-Kw')](\alpha)$ and both $(\alpha_1 V_m+\alpha V'_m)(\alpha)$ and $(\alpha_1 w+\alpha_2 w')(\alpha)$ are positive, the constant K must be positive. If there existed a positive arc of $f=\alpha_1 V_m+\alpha_2 V'_m$ in (A_1, A_n), which is not intersected by $g=K(\alpha_1 w+\alpha_2 w')$, there would then, by virtue of [5, Lemma 1.2], exist a constant K_1 such that $h=K_1(\alpha_1 w+\alpha_2 w')$ had a common tangent with f at some point of this arc. Since $g>f$, in the interval in question, we would necessarily have $K_1<K$ and h and f would intersect in the interval (A_n, A_{n+1}). The function $f-h$ would thus have a double zero somewhere in (A_1, A_n). This situation clearly contradicts Lemmas 1.17 and 1.18. Hence $f-g$ has at least $n+1$ zeros in $(a, \alpha]$, provided n is odd. Now if n is even we see by the unique nature of A_{n+1} that $f-g$ must vanish in (a, A_1). Hence $f-g$ has at least $n+1$ zeros in $(a, \alpha]$. Since A_{n+1} is minimal, we must have $\alpha=A_{n+1}$ and $K=0$. Hence $V_m=y_n$.

The essential uniqueness of $y_i(x)$ is proved as follows: Assume there are two solutions $u(x)$ and $d(x)$ which satisfy (a). Then $(\alpha_1 u''(b)+\alpha_2 u'''(b))d=z_1$ and $(\alpha_1 d''(b)+\alpha_2 d'''(b))u=z_2$, are nontrivial by Lemma 1.19. But $(\alpha_1(z_1-z_2)''+\alpha_2(z_1-z_2)''')(b)=0$. This contradicts Lemma 1.19. Note that b represents any of the points, C_{i4}. It follows that $z_1=z_2$, hence, the essential uniqueness.

As in the above theorem our zeros will be counted according to their multiplicities. We leave to the reader the formulation and proof for the cases (1.1)–(1.6), (1.7). Whenever we refer to such theorems we do so through the previous theorem as "Theorem 2.1". In these cases we do not have to assume $r(x)=1$.

Theorem 2.1 illustrates the type of result we can obtain. Other situations can also be covered by using different β_i's and η_i's.

We designate the ith eigenfunctions of (1.1)–(1.5), (1.6) and (1.7) as $F_{i1}(x)$, $F_{i2}(x)$ and $F_{i3}(x)$, respectively. In the case, (1.1)–(1.8), we assume the existence of a sequence of nonnegative eigenvalues $l_{i4}(b)$. In the following self-adjoint case of (1.1)–(1.8), the existence is assured:

$$\begin{gathered}\beta_2=-\beta_1\neq 0, \quad \beta_3=\beta_4=0, \quad \sum_{i=1}^{4}\eta_i^2\neq 0, \quad \eta_i\geqq 0,\\ \eta_4=\eta_3\neq 0, \quad \eta_2-\eta_1\geqq 0, \quad \alpha_2=0.\end{gathered} \tag{2.2}$$

We note that "Theorem 2.1" applies in this case.

We let $F_{i4}(x)$ be an ith eigenfunction of (1.1)–(1.8). We then have the following theorem.

THEOREM 2.3. *Let* $\beta_4=-\beta_3\neq 0$, $\eta_4=\eta_3\neq 0$. *Let* $\sum_{i=1}^{4}\beta_i^2\neq 0$, $\eta_i\geqq 0$, *and* $\alpha_1^2+\alpha_2^2\neq 0$, $\alpha_1\geqq 0$, $\alpha_2\geqq 0$. *If* $\alpha_1>0$, $\alpha_2>0$, *we assume* $(\alpha_1/\alpha_2)(\eta_2+\eta_3\beta_3^{-1}\beta_2)+(\eta_1-\eta_3\beta_3^{-1}\beta_1)\neq 0$. *If* $C_{i4}(l_{i4}(b))=b$ *and* $l_{i4}(C_{i4}(l))=l$, *then the zeros of* $\alpha_1 F_{i4}+\alpha_2 F'_{i4}$ *separate the zeros of* F_{i4} *on* (a, b). F_{i4} *has at most* i *zeros on* (a, b), $i=1, 2, 3, \cdots$.

Proof. If the hypotheses hold, then Theorem 2.1 applies, and then $\alpha_1 F_{i4}+\alpha_2 F'_{i4}$ has $i-1$ zeros on (a, b). Assume F_{i4} has $i+1$ zeros on (a, b). These zeros are simple, by

[5] and the conditions here. It is clear that $\alpha_1 F_{i4} + \alpha_2 F'_{i4}$ vanishes between consecutive zeros of F_{i4}, hence has i zeros on (a, b), a contradiction. The theorem follows.

In the case (2.2), the conditions $C_{i4}(l_{i4}(b)) = b$ and $l_{i4}(C_{i4}(l)) = l$, are satisfied, by [5] and the calculus of variations. In this case, since $\alpha_2 = 0$, F_{i4} has exactly $i-1$ simple zeros on (a, b). Also $r(x)$ need not be equal to one. We will consider some nonselfadjoint cases in a subsequent paper.

We also note that these theorems can be extended to the before-mentioned situation where $(\alpha_1 y' + \alpha_2 y'')(b) = (\alpha_1 y'' + \alpha_2 y''')(b) = 0$.

The following nonselfadjoint case does not present much irregularity and satisfies the conditions $C_{i4}(l_{i4}(b)) = b$ and $l_{i4}(C_{i4}(l)) = l$. Consider (1.1)–(1.8):

$$K_{34} = \beta_3\eta_4 + \eta_4\beta_3 \neq 0, \quad \beta_1 = -\beta_2 \neq 0, \quad \eta_1 = \eta_2 \neq 0,$$

$$\frac{(\beta_4\eta_2 - \beta_2\eta_4)}{K_{34}} \geqq 0, \quad \frac{(\beta_3\eta_1 - \beta_1\eta_3)}{K_{34}} \leqq 0, \quad \alpha_2 = 0. \tag{2.4}$$

We leave the proof of (2.4) for a subsequent paper.

THEOREM 2.5. *$F_{i3}(x)$ has $i+1$ simple zeros in (a, b). $i = 1, 2, 3, \cdots$.*

Proof. By "Theorem 2.1" and (1.21), $(rF''_{i3})(x)$ has $i+3$ zeros on $[a, b]$. By Rolle's theorem $(rF''_{i3})'(x)$ vanishes at least $i+2$ times on $[a, b]$ and hence $F_{i3}(x)$ vanishes at least $i+1$ times on (a, b).

Assume $F_{i3}(x)$ vanishes more than $i+1$ times on (a, b). Then by Rolle's theorem $(rF''_{i3})(x)$ vanishes more than $i+3$ times on $[a, b]$, a contradiction.

The zeros are simple by [5].

Note that [1, Thm. 5.7] does not cover this particular case.

3. Existence and oscillation theorems.

THEOREM 3.1. *Let $r(x) = 1$ in (1.1). Let $\alpha_1^2 + \alpha_2^2 \neq 0$, $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$, $\eta_i \geqq 0$, $\delta_1 \geqq 0$, $\delta_2 \geqq 0$, $(\alpha_1/\alpha_2)(\eta_2 + \eta_3\beta_3^{-1}\beta_2) + (\eta_1 - \eta_3\beta_3^{-1}\beta_1) \neq 0$ if $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta_4 = -\beta_3 \neq 0$, $\eta_4 = \eta_3 \neq 0$, $\eta_1\beta_3 = \beta_1\eta_3$, $\beta_1\eta_2 + \eta_1\beta_2 \neq 0$, $\beta_1 \neq 0$, $\beta_1^{-1}\beta_3 \geqq 1$. If $b_1 - a > 1$, let $b - a = 1$ and $(b_1 - a)^2/2 - \beta_1^{-1}\beta_3 < 0$. If $C_{i4}(l_{i4}(b)) = b$, $l_{i4}(C_{i4}(l)) = l$, $i = 1, 2, 3, \cdots$, then (1.1)–(1.9) has a denumerably infinite number of nonnegative eigenvalues, λ_{i4}, $i = 1, 2, 3, \cdots$; $\lambda_{i4} < \lambda_{i+1,4}$; $\lambda_{i4} \to \infty$, $i \to \infty$. The eigenfunctions of (1.1)–(1.9) are essentially unique.*

Proof. Let $K(x, l)$ be a solution of (1.1) which satisfies $M(K(a)) = N(K(a)) = \alpha_1 K(b) + \alpha_2 K'(b) = 0$. Clearly $K(x, l_{i4}) = CF_{i4}(x)$, Then Theorem 2.1 can be applied provided we know the sign of $\alpha_1 K(x) + \alpha_2 K'(x)$ in (a, a_1), where a_1 is the first zero of $\alpha_1 K(x, l_{i4}) + \alpha_2 K'(x, l_{i4})$ in (a, ∞). We note that C can be taken as $C = 1$. The sign of $\alpha_1 K(x) + \alpha_2 K'(x)$ can be obtained if we know whether $K_0 < 1$, $K_0 = 1$ or $K_0 > 1$, where $K_0 = (\alpha_1 u_2(b) + \alpha_2 u_2'(b))/(\alpha_1 u_1(b) + \alpha_2 u_1'(b))$, $u_1(b) = y_3(b) - \beta_1^{-1}\beta_3 y_1(b)$, $u_2(b) = y_4(b) + \beta_1^{-1}\beta_3 y_1(b)$. Now, by [4, p. 273], we have $y_3(x) = \sum_{m=0}^{\infty} a_{m3}(x) l^m$ and $y_1(x) = \sum_{m=0}^{\infty} a_{m1}(x) l^m$, where

$$a_{m3} = \int_a^x dt \int_t^x (u - t)(x - u) p(t) a_{m-1,3}(t)\, du,$$

$$a_{m1} = \int_a^x dt \int_t^x (u - t)(x - u) p(t) a_{m-1,1}(t)\, du,$$

$$a_{03}(x) = \frac{(x-a)^2}{2}, \qquad a_{01}(x) = 1.$$

Clearly, for $(x - a) \leqq 1$ we have $K_0 < 0$. Then from Lemmas 1.17 and 1.18 and Theorem 2.1, we have that $(P(K(c, l_{i+1,4})))(P(K(c, l_{i4}))) < 0$. By continuity, eigenvalues exist. The transformation of Theorem 3.3 obtains existence for $b_1 - a > 1$.

Our conditions easily imply the essential uniqueness of the eigenfunctions.

In Theorems 3.3, 3.4, 3.5 and 3.6, if $b_1 - a > 1$, we let $b - a = 1$ and $|(b_1 - a)\alpha_1/\alpha_2| < 1$.

LEMMA 3.2. *$F_{ik}(x), F'_{ik}(x), rF''_{ik}(x)$ and $(rF''_{ik})'(x)$, $k = 1, 2, 3$, are less than zero for $x > b$ and i odd, and greater than zero for $x > b$ and i even.*

Proof. We prove the case $k = 3$. Let i be even. Then $F_{i3}(x)$ has $i+1$ simple zeros in (a, b), by Theorem 2.5. Hence $F_{i3}(b) > 0$ (we assume $F_{i3}(a) < 0$), by [5]. The lemma follows by [5]. For i odd consider $-F_{i3}(x)$.

Let $H_3(b, x, l) = (\alpha_1(ry_1'')(b, l) + \alpha_2(ry_1'')'(b, l))y_2(x, l) - (\alpha_1(ry_2'')(b, l) + \alpha_2(ry_2'')' \times (b, l))y_1(x, l)$.

THEOREM 3.3. *If $\sum_{i=1}^{4} \gamma_i^2 \neq 0$, $\gamma_i \geqq 0$, $i = 1, 2, 3, 4$ and* (i) $\alpha_1^2 + \alpha_2^2 \neq 0$; $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$; *or* (ii) $|\alpha_1/\alpha_2| < 1$; *then the problems* (1.1)–(1.2), (1.3), (1.4) *have a denumerably infinite number of positive eigenvalues λ_{ik}, $k = 1, 2, 3$, $i = 1, 2, 3, \cdots$, respectively and corresponding essentially unique eigenfunctions $H_{ik}(b, x, \lambda_{ik})$, $\lambda_{ik} < \lambda_{i+1,k}$, $\lambda_{ik} \to \infty$, $i \to \infty$.*

Proof. We prove the case $k = 3$. Let n be odd. Then by an obvious extension of [5, Thm. 2.6], we have $H_3(b, x, l_{n3}) = KF_{n3}(x)$, $n = 1, 2, 3, \cdots$. If (i) holds and we assume $F_{n3}(a) < 0$, then clearly we have $K > 0$, $n = 1, 2, 3, \cdots$.

If (ii) holds we proceed as follows: We have $(ry_1'')(x, l) = \sum_{i=0}^{\infty} ra_{i1}''(x)l^i$ and $(ry_1'')'(x, l) = \sum_{i=0}^{\infty} (ra_{i1}'')'(x)l^i$, where $a_{i1}(x) = \int_a^x dt \int_t^x ((u-t)(x-u)/r(u))p(t)a_{i-1,1}(t)\, du$, $a_{01}(x) = 1$. These relations essentially follow from elementary existence theory; note [4, p. 273].

Clearly, $ra_{i1}'' = \int_a^x (x-t)p(t)a_{i-1,1}(t)\, dt$, $(ra_{i1}'')' = \int_a^x p(t)a_{i-1,1}(t)\, dt$. If $(b-a) \leqq 1$, we have $(ra_{i1}'')'(x) \geqq ra_{i1}''(x)$, $i = 1, 2, 3, \cdots$. It follows, then, that $(ry_1'')'(b, l)/(ry_1'')(b, l) \geqq 1$, $l > 0$. Then $(ry_1'')'(b, l)/(ry_1'')(b, l) > |\alpha_1/\alpha_2|$ or $|\alpha_2|(ry_1'')'(b, l) - |\alpha_1|(ry_1'')(b, l) > 0$, $l > 0$. (If $\alpha_1 > 0$, $\alpha_2 < 0$, use $-H_3(x, l)$.)

In any case we have $K > 0$ for $n = 1, 2, 3, \cdots$, Now, by Lemma 3.2, we have $(L(H_3(c, l_{n3})))(L(H(c, l_{n+1,3}))) < 0$. Hence λ_{i3}, $i = 1, 2, 3, \cdots$, exists and $H_3(b, x, \lambda_{i3}) = H_{i3}(b, x, \lambda_{i3})$, is essentially unique by a simple extension of [5, Thm. 2.6].

The case $b_1 - a > 1$ can be obtained from the transformation $x = At + B$, where $A = (b-a)/(b_1 - a)$, $B = a[1 - (b-a)/(b_1 - a)]$. The equation $(r(t)y''(t))'' - lp(t)y(t) = 0$, transforms into $(\hat{r}(x)\hat{y}''(x))'' - l((b_1 - a)^4/(b-a)^4)\hat{p}(x)\hat{y}(x) = 0$, where $a \leqq t \leqq b_1$ $\hat{r}(x) = r((x-B)/A)$, $\hat{p}(x) = p((x-B)/A)$, $\hat{y}(x) = y((x-B)/A)$.

If $b_1 - a > 1$ and $(b-a) = 1$, then the above proves the theorem for variable x and hence the transformation obtains the theorem for variable t. The theorem is proved.

THEOREM 3.4. *If $\sum_{i=1}^{4} \gamma_i^2 \neq 0$, $\gamma_i \geqq 0$ and* (i) $\alpha_1^2 + \alpha_2^2 \neq 0$, $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$; *or* (ii) $|\alpha_1/\alpha_2| < 1$; *then $(rH_3'')(b, x, l)$ has m zeros on $[a, b]$, $m = i+2$ or $i+3$, $i = 0, 1, 2, \cdots$ for $l_{i3}(b) \leqq l < l_{i+1,3}(b)$, $l_{03}(b) = 0$. $H_3(b, x, l)$ has n zeros on (a, b), $i \leqq n \leqq i+3$, $i = 0, 1, 2, \cdots$.*

Proof. Assume there exists an l_0, $l_{i3}(b) \leqq l_0 < l_{i+1,3}(b)$ for which $(rH_3'')(b, x, l_0)$ has more than $i+3$ zeros on $[a, b]$. Then $C_{i+1,3}(l_0) < b$ by "Theorem 2.1". But, by (1.21) and Theorem 1.22, we have $C_{i+1,3}(l_{i+1,3}(b)) = b < C_{i+1,3}(l_0)$, a contradiction. Hence $(rH_3'')(b, x, l_0)$ has $\leqq i+3$ zeros on $[a, b]$.

Next we show that $(rH_3'')(b, x, l)$ has at least $i+2$ zeros on $[a, b]$ for $l_{i3}(b) \leqq l < l_{i+1,3}(b)$. Let $A_1, A_2, \cdots, A_i$ be the i zeros of $(rH_3'')(b, x, l_{i+1,3}(b))$ on (a, b), by "Theorem 2.1" and (1.21). Since $(rH_3'')'(b, A_k, l_{i+1,3}(b)) \neq 0$, by [5], the implicit function theorem gives us i functions $A_k(l)$, $k = 1, \cdots, i$ in C' with respect to l and $(rH_3'')(b, A_k(l), l) = 0$, all in some neighborhood $N(l_{i+1,3}(b))$. Let L_1 be the left boundary point of N.

We consider $\lim_{l \to L_1} \sup A_k(l) = y_k$ and $\lim_{l \to L_1} \inf A_k(l) = w_k$, $k = 1, 2, \cdots, i$. We assert that $a < w_k \leqq y_k < b$. Suppose this were not so. Then a would be a zero of

$H_3(b, x, l)$ for some l, $L_1 \leqq l < l_{i+1,3}(b)$. The zero at a contradicts [5, Lemma 2.1]. The zero of $(rH_3'')(b, x, l)$ at b contradicts the fact that $l_{i3}(b)$ and $l_{i+1,3}(b)$ are consecutive eigenvalues.

By "Theorem 2.1" we have $a < A_1(l) < C_{13}(l)$, $C_{13}(l) < A_2(l) < C_{23}(l), \cdots, C_{i-1,3}(l) < A_i(l) < C_{i3}(l) < b$ in $N(l_{i+1,3}(b))$.

It is clear that $C_{k-1,3}(L_1) < w_k \leqq y_k < C_{k3}(L_1)$, $k = 1, \cdots, i-1$, where we define $C_{03}(L_1) = a$. Also we have $C_{k-1,3}(L_1) < w_i \leqq y_i \leqq C_{i3}(L_1)$. Now if $w_k \neq y_k$, we would have a point C_3 between $C_{k-1,3}(L_1)$ and $C_{k3}(L_1)$, which contradicts the definition of $C_{k-1,3}(L_1)$ and $C_{k3}(L_1)$. If $y_i = C_{i3}(L_1)$ and $w_i \neq y_i$ we contradict Theorem 1.22.

By the argument above we have shown that $A_1(L_1), A_2(L_1), \cdots, A_{i-1}(L_1)$ exist and $(rH_3'')'(b, A_k(L_1), L_1) \neq 0$. Hence $A_k(l)$ exists as a C' function $(k = 1, \cdots, i-1)$ for $l_{i3}(b) < l < l_{i+1,3}(b)$. Also, $a < A_k(l) < b$, $k = 1, 2, \cdots, i$. If $y_i = C_{i3}(L_1)$, $m = i+3$. If $y_i \neq C_{i3}(L_1)$, $m \geqq i+2$.

Now if $m = i+2$, then it is easy to see that $n \geqq i$ by Rolle's theorem. If $m = i+3$, then $n \geqq i+1$. If $m = i+2$, then $n \leqq i+2$. If $m = i+3$, $n \leqq i+3$. Hence $i \leqq n \leqq i+3$. Let $H_1(b, x, l) = (\alpha_1 y_3(b, l) + \alpha_2 y_3'(b, l))y_4(x, l) - (\alpha_1 y_4(b, l) + \alpha_2 y_4'(b, l))y_3(x, l)$ and $H_2(b, x, l) = (\alpha_1 (ry_3')(b, l) + \alpha_2 (ry_3'')'(b, l))y_3(x, l) - (\alpha_1 (ry_4''(b, l) + \alpha_2 (ry_4'')'(b, l))y_3(x, l)$.

We obtain the following corresponding theorems for (1.1)–(1.2), (1.3).

THEOREM 3.5. *If* $\sum_{L=1}^{4} \gamma_i^2 \neq 0$, $\gamma_i \geqq 0$ *and* (i) $\alpha_1^2 + \alpha_2^2 \neq 0$, $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$; *or* (ii) $|\alpha_1/\alpha_2| < 1$, *then* $H_1(b, x, l)$ *has* m *zeros on* $[a, b]$, $m = i+2$ *or* $m = i+3$ *for* $l_{i1}(b) \leqq l < l_{i+1,1}(b)$, $l_{01}(b) = 0$, $i = 0, 1, 2, \cdots$.

THEOREM 3.6. *If* $\sum_{i=1}^{4} \gamma_i^2 \neq 0$, $\gamma_i \geqq 0$ *and* (i) $\alpha_1^2 + \alpha_2^2 \neq 0$, $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$ *or* (ii) $|\alpha_1/\alpha_2| < 1$, *then* $(rH_2'')(b, x, l)$ *has* m *zeros on* $[a, b]$, $m = i$ *or* $m = i+1$ *for* $l_{i2}(b) \leqq l < l_{i+1,2}(b)$. $l_{02}(b) = 0$. $H_2(b, x, l)$ *has* n *zeros on* $[a, b]$, $i-2 \leqq n \leqq i+1$, $i = 0, 1, 2, \cdots$.

COROLLARY 3.7. *If* $\alpha_1 > 0$, $\alpha_2 = 0$ *in Theorem* 3.4, *then* $m = i+3$. *If only* (i) *holds, the zeros of* (rH_3'') (b, x, l) *and* $H_3(b, x, l)$ *in* (a, b) *are simple. If* $\alpha_1 > 0$, $\alpha_2 = 0$, *in Theorem* 3.5, *then* $m = i+3$. *If only* (i) *holds, then the zeros of* $H_1(b, x, l)$ *in* (a, b) *are simple. If* $\alpha_1 > 0$, $\alpha_2 = 0$ *in Theorem* 3.6, *then* $m = i+1$. *If only* (i) *holds, the zeros of* $(rH_2'')(b, x, l)$ *and* $H_2(b, x, l)$ *in* (a, b) *are simple.*

Proof. The proof follows from theorems in [5].

4. Eigenvalue comparison theorems.

THEOREM 4.1. *If* $\gamma_1 > 0$, $\gamma_2 = \gamma_3 = \gamma_4 = 0$ *and* $\alpha_1 > 0$, $\alpha_2 = 0$ *in* (1.1)–(1.2) *and* $\lambda(b, c)$ *is any eigenvalue which lies in the interval* $(l_{i1}(b), l_{i+1,1}(b))$, $i = 0, 1, 2, \cdots$, $l_{01}(b) = 0$, *then* $\lambda(b, c) > l_{i+1,1}(c)$.

Proof. Assume $\lambda(b, c) < l_{i+1,1}(c)$. $H_1(b, x, \lambda(b, c))$ has $i+3$ zeros in $[a, b]$ by Corollary 3.7 and one at c. Therefore $C_{i+1,1}(\lambda(b, c)) \leqq c$, by "Theorem 2.1". But $C_{i+1,1}(\lambda(b, c)) > C_{i+1,1}(l_{i+1,1}(c)) = c$, by Theorem 1.22. Hence $\lambda(b, c) \geqq l_{i+1,1}(c)$.

If $\lambda(b, c) = l_{i+1,1}(c)$, then $H_1(b, x, \lambda(b, c))$ has $i+5$ zeros in $[a, c]$, $i+3$ in $[a, b]$ and two at c. But $H_1(b, x, l_{i+1,1}(c))$ has $i+4$ zeros in $[a, c]$ by relation (1.21) and "Theorem 2.1". Hence $\lambda(b, c) \neq l_{i+1,1}(c)$ and the theorem follows.

Corresponding theorems can be proved for (1.1)–(1.3), (1.4), (1.9).

The following theorem extends a result stated in [2, p. 207].

THEOREM 4.2. $l_{i2}(b) < l_{i1}(b)$, $b > a$, $i = 1, 2, 3, \cdots$.

Proof. Assume $l_{i1}(b) \leqq l_{i2}(b)$. By Rolle's theorem, "Theorem 2.1", (1.21) and Theorem 1.22, we have $C_{i2}(l_{i2}(b)) = b \leqq C_{i2}(l_{i1}(b)) < b$, a contradiction. The theorem follows.

5. Generalizations and discussion.

Let

$$M(y(x)) = A_{11}y(x) + A_{12}y'(x) + A_{13}(ry'')(x) + A_{14}(ry'')'(x),$$

$$N(y(x)) = B_{11}y(x) - B_{12}y'(x) + B_{13}(ry'')(x) - B_{14}(ry'')'(x),$$

$$P(y(x)) = (K_{11}y(x) + K_{12}y'(x)) + (K_{11}y'(x) + K_{12}y''(x)),$$

$$Q(y(x)) = (K_{11}y'(x) + K_{12}y''(x)) + (K_{11}y''(x) + K_{12}y'''(x)),$$

$$S(y(x)) = M_{11}(y(x) + y'(x)) + M_{12}(y'(x) + y''(x)),$$

$$T(y(x)) = M_{11}(y'(x) + y''(x)) + M_{12}(y''(x) + y'''(x)),$$

$$R(y(x)) = P(y(x)) + Q(y(x)) + y(x) + K_{11}y(x) + K_{12}y'(x) + K_{11}y''(x) + K_{12}y'''(x),$$

$$W(y(x)) = y(x) + y(x) + y'(x) + y'(x) + y''(x) + y''(x) + y'''(x).$$

Then we may consider the problems:

$$M(y(a)) = N(y(a)) = P(y(b)) = Q(y(b)) = 0, \tag{5.1}$$

$$M(y(a)) = N(y(a)) = P(y(b)) = R(y(c)) = 0, \tag{5.2}$$

$$M(y(a)) = N(y(a)) = S(y(b)) = T(y(b)) = 0, \tag{5.3}$$

$$M(y(a)) = N(y(a)) = S(y(b)) = W(y(c)) = 0, \qquad 0 < a < b < c < \infty. \tag{5.4}$$

The methods presented in this paper can be applied to the above problems.

We note that Theorem 2.5 shows that the above approach covers many cases not covered by the theory in [1].

Theorem 4.1, although apparently simple, is actually somewhat involved because we are comparing two problems whose eigenfunctions belong to two different sets of functions. See for example [3, Chapt. VI].

The proof of the relation in Theorem 2.3, which corresponds to (1.21), is of interest in nonselfadjoint, as well as in selfadjoint cases. See for example [6], which contains related material.

Lemma 1.17 can be proved for cases where $r(x)$ is not constant.

6. Acknowledgments. I wish to thank the people of SIAM who handled this paper, for their help and understanding. I also wish to thank Antoinette Clement for typing the manuscript.

REFERENCES

[1] D. BANKS AND G. KUROWSKI, *A Prüfer transformation for the equation of the vibrating beam*, Trans. Amer. Math. Soc., 199 (1974) pp. 203–222.

[2] J. BARRETT, *Systems-disconjugacy of a fourth-order differential equation*, Proc. Amer. Math. Soc., 12 (1961), pp. 205–213.

[3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1953.

[4] E. L. INCE, *Ordinary Differential Equations*, 2nd ed., Dover, New York, 1956.

[5] LEIGHTON AND NEHARI, *On the oscillation of solutions of self-adjoint linear differential equations of the fourth order*, Trans. Amer. Math. Soc., 89 (1958), pp. 325–377.

[6] W. T. REID, *Variational aspects of oscillation phenomena for higher order differential equations*, J. Math. Anal. Appl., 40 (1972), pp. 446–470.

[7] J. ROVDER, *Three point value problem for a third order linear differential equation*, Math. Slovaca, 27 (1977), pp. 97–111.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0008 $01.00/0

POSITIVE SOLUTIONS OF NEGATIVE EXPONENT GENERALIZED EMDEN-FOWLER BOUNDARY VALUE PROBLEMS*

C. D. LUNING† AND W. L. PERRY‡

Abstract. A constructive proof of existence of positive solution for negative exponent and sublinear generalized Emden–Fowler boundary value problems is given. The proof utilizes a monotone iterative scheme of Picard type.

1. Introduction.

The class of boundary value problems

$$y''(x)+a(x)y^{\mu}(x)=0, \qquad a<x<b, \quad \mu\in\mathbb{R}, \tag{1.1}$$

$$\alpha y(a)-\beta y'(a)=0, \qquad \gamma y(b)+\delta y'(b)=0 \tag{1.2}$$

has proved to be very important in applied mathematics. Among the equations in this class are:

(A) The Thomas–Fermi equation where $\mu=\frac{3}{2}$ and $a(x)=-x^{1/2}$. This equation was developed in studies of atomic structures initiated in 1927 by L. H. Thomas [1] and E. Fermi [6]. The equation still has significant use in atomic calculations [5]. The boundary conditions (1.2) are obtained from the usual Thomas–Fermi boundary conditions by a change of variable and a normalization [8].

(B) The generalized Emden–Fowler equation where $\mu>0$ and $a(x)>0$. This equation arises in the fields of gas dynamics, Newtonian fluid mechanics, nuclear physics, and chemically reacting systems [16]. Recently the equation has arisen in the study of multipole toroidal plasmas [1]. The original Lane–Emden equation, developed in 1869 and subsequently studied by Fowler, can be put in the form (1.1) whence the name generalized Emden–Fowler. The cases $0<\mu<1$ and $\mu>1$ are called sublinear and superlinear respectively.

(C) The negative exponent sublinear equation. Recently (1.1) has been used in modeling non-Newtonian fluids such as coal slurries [2]. In this case $a(x)>0$ and $\mu<0$.

In most of these applications, the physical interest lies in existence and uniqueness of positive solutions. In previous papers we studied problem (A) [8], [9] and the superlinear case of (B) [10] from the point of view of developing constructive proofs for existence of positive solutions. By means of the change of variable $y=\lambda^{1/\mu}u$ and a change of independent variable, problem (1.1), (1.2) is transformed to the nonlinear eigenvalue problem,

$$u''(x)+\lambda a(x)u^{\mu}(x)=0, \qquad 0<x<1, \quad \mu\in\mathbb{R}, \tag{1.3}$$

$$\alpha u(0)-\beta u'(0)=0, \qquad \gamma u(1)+\delta u'(1)=0. \tag{1.4}$$

We showed that, under certain restrictions on $a(x)$, α, β, γ and δ, the iteration defined by

$$u_n''(x)+\lambda_n a(x)u_{n-1}^{\mu-1}(x)u_n(x)=0, \qquad 0<x<1, \tag{1.5}$$

$$\alpha u_n(0)-\beta u_n'(0)=0, \qquad \gamma u_n(1)+\delta u_n'(1)=0, \tag{1.6}$$

with $u_n(x)$ appropriately normalized and $u_0(x)$ appropriately chosen, generates a sequence $\{u_n, \lambda_n\}$ that converges uniformly to a positive solution (u, λ) of (1.3), (1.4).

* Received by the editors December 4, 1979, and in revised form January 11, 1981.
† Department of Mathematics, Sam Houston State University, Huntsville, Texas 77340.
‡ Department of Mathematics, Texas A & M University, College Station, Texas 77843.

Detailed results of computational implementation of the scheme just described may be found in [11], [12].

In this paper we are primarily interested in the negative exponent sublinear eigenvalue problem, $-1 \leqq \mu < 0$, and the sublinear eigenvalue problem, $0 < \mu < 1$,

$$u''(x) + \lambda a(x) u^{\mu}(x) = 0, \qquad 0 < x < 1, \tag{1.7}$$

$$u(0) = u(1) = 0. \tag{1.8}$$

Other authors have also developed constructive techniques for problems similar to (1.7), (1.8). In [2], [13] Callegari and Nachman develop explicit power series solutions for problem (1.7) with $\mu \leqq -1$ and boundary conditions $u'(0) = 0$, $u(1) = 0$. Callegari and Reiss in [3] and Callegari, Reiss and Keller in [4] apply shooting methods to the problem $(x^3 u')' + a(x) u^{-2} = 0$, $0 < x < 1$ for certain functions $a(x)$ and boundary conditions $u'(0) = 0$, and either $u(1) = \lambda$ or $u'(1) + \alpha u(1) = U$. They consider the various cases $\lambda > 0$, $\lambda = 0$, $\lambda < 0$, $U = 0$, $U < 0$ and prove comprehensive existence and uniqueness results for each case. Earlier constructive results of H. Keller and D. Cohen [7] apply to the problem

$$Lu = \lambda f(x, u), \qquad x \in D \subseteq \mathbb{R}^n,$$

with appropriate boundary conditions, if $f(x, 0) > 0$, $f(x, u)$ is sufficiently smooth, and $f(x, u)$ satisfies a monotonicity condition. In our cases $f(x, 0)$ equals 0 or is undefined, so the results do not apply. Our approach differs from those in [2], [3], [4], [10] and [13] in that in this paper we choose a function $u_0(x)$ and use the Picard type iteration.

$$u_n''(x) + \lambda_n a(x) u_{n-1}^{\mu}(x) = 0, \qquad 0 < x < 1, \tag{1.9}$$

$$u_n(0) = u_n(1) = 0, \tag{1.10}$$

where λ_n is determined by a normalization to obtain a sequence $\{u_n, \lambda_n\}_{n=1}^{\infty}$ which is shown to converge uniformly to a solution of (1.7), (1.8). We remark that the proof for the sublinear case $0 < \mu < 1$ is actually valid for all $\mu > 0$ and thus includes the superlinear case $\mu > 1$. Since the iteration scheme (1.9), (1.10) is significantly easier to implement than the scheme in [10] this iteration should supplant that of [10] in the superlinear case.

We remark that Taliaferro [14] has given necessary and sufficient conditions for the existence and uniqueness of a positive solution of (1.7), (1.8). He also gives necessary and sufficient conditions for $y'(0)$ or $y'(1)$ to be finite. Our results complement those of Taliaferro in that our proof is constructive and thus gives a method for obtaining an approximate solution.

We assume $a(x) \in C(0, 1)$, $a(x) > 0$, $0 < x < 1$ and $\int_0^1 \xi^{\mu} a(\xi)\, d\xi < \infty$. Let

$$K(x, \xi) = \begin{cases} (1-x)\xi, & 0 < \xi < x, \\ (1-\xi)x, & x < \xi < 1. \end{cases} \tag{1.11}$$

The Picard iterates $\{u_n\}_{n=0}^{\infty}$ and the sequence $\{\lambda_n\}_{n=1}^{\infty}$ are defined: $u_0(x) = x$, and for $n \geqq 1$

$$u_n(x) = \lambda_n \int_0^1 K(x, \xi) a(\xi) u_{n-1}^{\mu}(\xi)\, d\xi, \tag{1.12}$$

where $u_n(x)$ is normalized by choosing λ_n so that

$$\lambda_n \int_0^1 (1-\xi) a(\xi) u_{n-1}^{\mu}(\xi)\, d\xi = 1. \tag{1.13}$$

For $n \geqq 1$ it follows from (1.12) that $u_n(x)$ satisfies (1.9), (1.10) and from (1.13) that $u_n'(0) = 1$. Of course $u_n(x) > 0$, $0 < x < 1$ and $\lambda_n > 0$.

In § 2 we consider the case $\mu > 0$ and prove

THEOREM 1. *If $\mu > 0$ and if $a(x)$, $\{u_n\}_{n=0}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ are as in* (1.12), (1.13), *then $0 < u_{n+1}(x) < u_n(x)$, $0 < x < 1$ and $0 < \lambda_n < \lambda_{n+1}$. Moreover, there is a positive solution $\{u, \lambda\}$ of* (1.7), (1, 8) *such that $\lim_{n\to\infty} \lambda_n = \lambda$ and $\lim_{n\to\infty} u_n(x) = u(x)$ uniformly on* [0, 1].

In § 3 we consider the case $-1 \leqq \mu < 0$ and prove

THEOREM 2. *If $-1 \leqq \mu < 0$, and if $a(x)$, $\{u_n\}_{n=0}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ are as in* (1.12), (1.13), *then for $n \geqq 1$ $0 < u_{2n-1}(x) < u_{2n+1}(x) < u_{2n}(x) < u_{2n-2}(x)$, $0 < x < 1$, $\lambda_{2n} < \lambda_{2n+2} < \lambda_{2n+1} < \lambda_{2n-1}$. Moreover, there is a positive solution $\{u, \lambda\}$ of* (1.7), (1.8) *such that $\lim_{n\to\infty} \lambda_n = \lambda$ and $\lim_{n\to\infty} u_n(x) = u(x)$ uniformly on* [0, 1].

2. Convergence of the Picard iterates for $\mu > 0$. In this section we consider the iterations (1.12), (1.13) with $u_0(x) = x$ and $\mu > 0$. We first prove a lemma proving that in this case the iterations (1.12), (1.13) generate monotone sequences.

LEMMA 1. *For $n \geqq 1$, $0 < u_n(x) < u_{n-1}(x)$, $0 < x < 1$, and $0 < \lambda_n < \lambda_{n+1}$. Also, for any $0 < M < 1$ there is at most one value of $x \in (0, 1)$ such that $Mu_{n-1}^{\mu}(x) - u_n^{\mu}(x) = 0$.*

Proof (by induction). For $k \geqq 0$ let $f_k(x) = M^{1/\mu} u_k(x) - u_{k+1}(x)$. Then $Mu_{n-1}^{\mu}(x) - u_n^{\mu}(x) = 0$ for at most one value of $x \in (0, 1)$.

We have $(u_0 - u_1)''(x) = \lambda_1 a(x) u_0^{\mu}(x) > 0$, $0 < x < 1$, $(u_0 - u_1)(0) = (u_0 - u_1)'(0) = 0$. Thus $(u_0 - u_1)(x) > 0$, $0 < x < 1$. Also $f_0''(x) = \lambda_1 a(x) u_0^{\mu}(x) > 0$, $0 < x < 1$, $f_0(0) = 0$, $f_0'(0) = M^{1/\mu} - 1 < 0$. Thus $f_0(x)$ can cross the x-axis at most once interior to the interval (0, 1). From the normalization (1.13) and the result $u_0(x) > u_1(x)$ we conclude $\lambda_2 > \lambda_1$.

Assume $(u_{k-1} - u_k)(x) > 0$, $0 < x < 1$ and that $f_{k-1}(x)$ crosses the x-axis at most once interior to the interval (0, 1). From (1.13) these assumptions imply $\lambda_{k+1} > \lambda_k$. We have

$$(u_k - u_{k+1})''(x) = a(x)[\lambda_{k+1} u_k^{\mu}(x) - \lambda_k u_{k-1}^{\mu}(x)].$$

Thus $(u_k - u_{k+1})''(x) > 0$ for x near zero, and, by the induction hypothesis on $f_{k-1}(x)$, we have that there is at most one $x \in (0, 1)$ such that $(u_k - u_{k+1})''(x) = 0$. Thus $(u_k - u_{k+1})(x)$ is convex for x near zero and has at most one inflection point interior to the interval (0, 1). In order to satisfy the boundary conditions $(u_k - u_{k+1})(0) = (u_k - u_{k+1})'(0) = (u_k - u_{k+1})(1) = 0$ for $k \geqq 1$, we conclude that $(u_k - u_{k+1})(x) > 0$, $0 < x < 1$. Similarly, since $f_k''(x) = a(x)[\lambda_{k+1} u_k^{\mu}(x) - M^{1/\mu} \lambda_k u_{k-1}^{\mu}(x)]$, we have $f_k''(x) > 0$ for x near zero, and by the induction hypothesis on $f_{k-1}(x)$ we conclude $f_k''(x)$ can be zero at most once interior to (0, 1). In order to satisfy the boundary values $f_k(0) = f_k(1) = 0$, $f_k'(0) = M^{1/\mu} - 1 < 0$ we conclude that $f_k(x)$ can cross the x-axis at most once interior to the interval (0, 1). From the normalization (1.13) and $(u_k - u_{k+1})(x) > 0$, $0 < x < 1$, we conclude that $\lambda_{k+1} < \lambda_{k+2}$.

We now show that there is a continuous function $w(x)$ which is positive in some neighborhood of zero and such that $u_n(x) > w(x)$ from whence we will be able to conclude the sequence $\{u_n\}$ does not converge to the zero function.

LEMMA 2. *There exists $w(x) \in C[0, 1]$, $w(x) \geqq 0$, $w(x) > 0$ for x near zero, such that $u_n(x) \geqq w(x)$, $0 \leqq x \leqq 1$.*

Proof. Let

$$T(x) = \begin{cases} x, & 0 \leqq x \leqq \frac{1}{2}, \\ 1 - x, & \frac{1}{2} \leqq x \leqq 1. \end{cases} \tag{2.1}$$

Since for $n \geqq 1$ $u_n(0) = u_n(1) = 0$, $u_n(x) \geqq 0$, $u_n''(x) < 0$, $0 < x < 1$, it follows that $u_n(x) \geqq$

$\|u_n\|_\infty T(x)$, where $\|u_n\|_\infty = \sup_{0\leqq x\leqq 1}|u_n(x)|$. From (1.12),

$$u_n(x) \geqq \lambda_n \|u_{n-1}\|_\infty^\mu \int_0^1 K(x,\xi)a(\xi)T^\mu(\xi)\,d\xi, \tag{2.2}$$

and since $0<u_n(x)<u_0(x)<1$, $0<x<1$, it follows that

$$1 > \lambda_n \|u_{n-1}\|_\infty^\mu \int_0^1 K(x,\xi)a(\xi)T^\mu(\xi)\,d\xi, \tag{2.3}$$

and thus

$$\|u_{n-1}\|_\infty^\mu < \lambda_n^{-1}\left\| \int_0^1 K(x,\xi)a(\xi)T^\mu(\xi)\,d\xi\right\|_\infty^{-1}. \tag{2.4}$$

Hence

$$-u_n''(x) = \lambda_n a(x)u_{n-1}^\mu(x) < a(x)\left\| \int_0^1 K(x,\xi)a(\xi)T^\mu(\xi)\,d\xi\right\|_\infty^{-1}. \tag{2.5}$$

That is, there is a $C>0$ such that $u_n''(x) > -Ca(x)$, $0<x<1$. Using $u_n(0)=0$, $u_n(1)=0$, $u_n'(0)=1$ we conclude the existence of $w(x)$.

To complete the proof of Theorem 1 we have, from (1.13),

$$\lambda_n = \left(\int_0^1 (1-\xi)a(\xi)u_{n-1}^\mu(\xi)\,d\xi\right)^{-1} \leqq \left(\int_0^1 (1-\xi)a(\xi)w^\mu(\xi)\,d\xi\right)^{-1}.$$

Thus the increasing sequence $\{\lambda_n\}_{n=1}^\infty$ is bounded above and there is a $\lambda>0$ such that $\lim_{n\to\infty}\lambda_n = \lambda$. Now, using (1.12), we can conclude there exists $K>0$ such that

$$|u_n(x_2) - u_n(x_1)| = \left| \int_{x_1}^{x_2} u_n'(\xi)\,d\xi\right| \leqq K|x_2 - x_1|. \tag{2.6}$$

Of course $0\leqq u_n(x)\leqq 1$, $0\leqq x\leqq 1$. Thus the sequence of functions $\{u_n(x)\}_{n=0}^\infty$ is equicontinuous and uniformly bounded on $[0,1]$. By Ascoli's lemma there exists $u\in C[0,1]$ such that $\lim_{n\to\infty} u_n = u$ uniformly on $[0,1]$. From Lemma 2 we have $u(x)$ is not identically zero, and from (1.12) and the dominated convergence theorem $u(x) = \lambda\int_0^1 K(x,\xi)a(\xi)u^\mu(\xi)\,d\xi$. Thus $u(x)>0$, $0<x<1$ and $\{u,\lambda\}$ is a positive solution of (1.7), (1.8) with $\mu>0$.

3. Convergence of the Picard iterates for $-1\leqq\mu<0$. In this section we consider the iterations (1.12), (1.13) with $u_0(x)=x$ and $\mu<0$. We first prove a lemma showing the iterations generate alternating monotone sequences.

LEMMA 3. *For $n\geqq 1$, $u_{2n-1}(x)<u_{2n+1}(x)<u_{2n}(x)<u_{2n-2}(x)$, $0<x<1$, $\lambda_{2n}<\lambda_{2n+2}<\lambda_{2n+1}<\lambda_{2n-1}$. Moreover, for any $0<M<1$ there is at most one value of $x\in(0,1)$ such that $Mu_i^{-\mu}(x) - u_j^{-\mu}(x) = 0$, where pair (i,j) can be $(2n-2,2n-1)$, $(2n,2n-1)$, $(2n-2,2n)$ or $(2n+1,2n-1)$.*

Proof. The proof of Lemma 3 is very similar to the proof of Lemma 1. Again we use that $Mu_i^{-\mu}(x) - u_j^{-\mu}(x) = 0$ for at most one $x\in(0,1)$ is equivalent to $M^{-1/\mu}u_i(x) - u_j(x) = 0$ for at most one $x\in(0,1)$. The argument that $(u_0-u_1)(x)>0$ $0<x<1$ and $(M^{-1/\mu}u_0 - u_1)(x) = 0$ for at most one $x\in(0,1)$ are the same as in Lemma 1. Similarly, it can be shown that $(u_0-u_2)(x)>0$, $0<x<1$ and $(M^{-1/\mu}u_0 - u_2)(x) = 0$ for at most one $x\in(0,1)$. From the normalization (1.13) we conclude $\lambda_2<\lambda_1$, $\lambda_3<\lambda_1$. For the induction hypothesis we assume $(u_{2n}-u_{2n+1})(x)>0$, $0<x<1$, $(u_{2n}-u_{2n+2})(x)>0$, $0<x<1$, there is at most one $x\in(0,1)$ such that $(M^{-1/\mu}u_{2n} - u_{2n+1})(x) = 0$, and there

is at most one $x \in (0, 1)$ such that $(M^{-1/\mu}u_{2n} - u_{2n+1})(x) = 0$. By the normalization (1.13) we have $\lambda_{2n+2} < \lambda_{2n+1}$ and $\lambda_{2n+2} < \lambda_{2n+1}$. To show that $(u_{2n+2} - u_{2n+1})(x) > 0$, $0 < x < 1$, we proceed as follows

$$\begin{aligned}(u_{2n+2} - u_{2n+1})''(x) &= \lambda_{2n+1}a(x)u_{2n}^{\mu}(x) - \lambda_{2n+2}a(x)u_{2n+1}^{\mu}(x)\\ &= -\lambda_{2n+1}a(x)u_{2n}^{\mu}(x)u_{2n+1}^{\mu}(x)\\ &\quad \times\left[\frac{\lambda_{2n+2}}{\lambda_{2n+1}}u_{2n}^{-\mu}(x) - u_{2n+1}^{-\mu}(x)\right].\end{aligned}$$

The argument now parallels that of Lemma 1, and we conclude $(u_{2n+2} - u_{2n+1})(x) > 0$, $0 < x < 1$. Similarly it is shown that there is at most one $x \in (0, 1)$ such that $(M^{-1/\mu}u_{2n+2} - u_{2n+1})(x) = 0$. From the normalization (1.13) we conclude $\lambda_{2n+3} > \lambda_{2n+2}$. One then proceeds in the same manner to consider $(u_{2n+3} - u_{2n+1})(x)$, $(M^{-1/\mu}u_{2n+3} - u_{2n+1})(x)$, $(u_{2n+2} - u_{2n+3})(x)$, $(M^{-1/\mu}u_{2n+2} - u_{2n+3})(x)$, and $(u_{2n+2} - u_{2n+4})(x)$, $(M^{-1/\mu}u_{2n+2} - u_{2n+4})(x)$ along with the normalization (1.13) to order the λ's. This then completes the induction proof of the lemma.

Since $0 < \lambda_{2n} < \lambda_{2n+2} < \lambda_{2n+1} < \lambda_{2n-1}$, there exists $0 < \check{\lambda} \leqq \hat{\lambda}$ such that $\lim_{n\to\infty} \lambda_{2n} = \check{\lambda}$, $\lim_{n\to\infty} \lambda_{2n+1} = \hat{\lambda}$. Since $0 < u_{2n+1}(x) < u_{2n}(x) < u_{2n-2}(x) \leqq u_0(x)$ $0 < x < 1$, the sequences $\{u_{2n-1}\}$ and $\{u_{2n}\}$ are uniformly bounded monotone sequences. Repeating the argument of (2.6), we also conclude that the sequences are equicontinuous. Thus by Ascoli's lemma there exist functions $\check{u}, \hat{u} \in C[0, 1]$ such that $0 < \check{u}(x) \leqq \hat{u}(x)$, $0 < x < 1$ and $\lim_{n\to\infty} u_{2n-1} = \check{u}$ and $\lim_{n\to\infty} u_{2n} = \hat{u}$ uniformly on $[0, 1]$. Using the dominated convergence theorem and (1.12), we have

$$\begin{aligned}\hat{u}(x) &= \check{\lambda}\int_0^1 K(x, \xi)a(\xi)\check{u}^{\mu}(\xi)\,d\xi,\\ \check{u}(x) &= \hat{\lambda}\int_0^1 K(x, \xi)a(\xi)\hat{u}^{\mu}(\xi)\,d\xi\end{aligned} \tag{3.1}$$

or, equivalently,

$$\begin{aligned}&\hat{u}''(x) + \check{\lambda}a(x)\check{u}^{\mu}(x) = 0, \qquad 0 < x < 1,\\ &\check{u}''(x) + \hat{\lambda}a(x)\hat{u}^{\mu}(x) = 0, \qquad 0 < x < 1,\\ &\hat{u}(0) = \check{u}(0) = \hat{u}(1) = \check{u}(1) = 0.\end{aligned} \tag{3.2}$$

Since $u_n'(0) = 1$ and the convergence is uniform, we also have $\hat{u}'(0) = \check{u}'(0) = 1$. To show $\hat{\lambda} = \check{\lambda}$ and $\hat{u} = \check{u}$, we consider $f(x) = \hat{u}(x)\check{u}'(x) - \check{u}(x)\hat{u}'(x)$. Then $f(0) = f(1) = 0$ and $f'(x) = \check{\lambda}a(x)\hat{u}^{1+\mu}(x) - \hat{\lambda}a(x)\hat{u}^{1+\mu}(x)$. Using $0 < \check{\lambda} < \hat{\lambda}$, $0 < \check{u}(x) \leqq \hat{u}(x)$, $0 < x < 1$ and $1 + \mu \geqq 0$, we have $f'(x) \geqq 0$, $0 < x < 1$. In order to satisfy the boundary conditions $f(0) = f(1) = 0$, we must have $f'(x) = 0$, $0 < x < 1$, from which we conclude $\hat{\lambda} = \check{\lambda}$, $\hat{u}(x) = \check{u}(x)$, $0 \leqq x \leqq 1$ which completes the proof of theorem 2.

REFERENCES

[1] J. G. BERRYMAN, *Evolution of a stable profile for a class of nonlinear diffusion equations with fixed boundaries*, J. Math. Phys., 18 (1977), pp. 2108–2112.

[2] A CALLEGARI AND A. NACHMAN, *Some singular, nonlinear differential equations arising in boundary layer theory*, J. Math. Anal. Appl., 64 (1978), pp. 96–105.

[3] A. CALLEGARI AND E. REISS, *Nonlinear boundary value problems for circular membranes*, Arch. Rat. Mech. Anal., 31 (1968), pp. 390–400.

[4] A. CALLEGARI,, E. REISS AND H. KELLER, *Membrane buckling: A study of solution multiplicity*, Comm. Pure Appl. Math., 24 (1971), 499–527.
[5] P. CSAVINSZKY, *Universal approximate solution of the Thomas–Fermi equation for ions*, Phys. Rev. A, 8 (1973), pp. 1688–1701.
[6] E. FERMI, *Un metado statistico par la determinazione di alcune proprietá dell'atome*, Rend. Accad. Naz. del Lincei. Cl. Sci. Fis., Mat. e Nat. 6 (1927), pp. 602–607.
[7] H. KELLER AND D. COHEN, *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), 1361–1376.
[8] C. D. LUNING, *An iterative technique for obtaining solutions of a Thomas-Fermi equation*, this journal, 9 (1978), pp. 515–522.
[9] C. D. LUNING AND W. L. PERRY, *An iterative technique for solution of the Thomas-Fermi equation utilizing a nonlinear eigenvalue problem*, Quart. Appl. Math., 35 (1977), pp. 257–268.
[10] ———, *Positive solutions of superlinear eigenvalue problems via a monotone iterative technique*, J. Differential Equations, 33 (1979), pp. 359–367.
[11] C. D. LUNING, W. L. PERRY AND R. FLAGG, *Implementation of an iterative technique for the solution of generalized Emden–Fowler eigenproblems*, Proc. of Conference on Codes for Boundary-Value Problems in Ordinary Differential Equations, Lecture Notes in Computer Science 76, Springer–Verlag, New York, 1979.
[12] ———, *Implementation of new iterative techniques for solutions of Thomas–Fermi and Emden–Fowler equations*, J. Comp. Phys., to appear.
[13] A. NACHMAN AND A. CALLEGARI, *A nonlinear singular boundary value problem in the theory of pseudoplastic fluids*, SIAM J. Appl. Math., 38 (1980), pp. 275–281.
[14] S. TALIAFERRO, *A nonlinear singular boundary value problem*, Nonlinear Analysis, Th., Meth. and Appl., 3 (1979), pp. 897–904.
[15] L. H. THOMAS, *The calculation of atomic fields*, Proc. Camb. Phil. Soc., 23 (1927), pp. 542–548.
[16] J. S. W. WONG, *On the generalized Emden–Fowler equation*, SIAM Rev., 17 (1975), pp. 339–360.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0009 $01.00/0

TRAVELING WAVE SOLUTIONS FOR SOME NONLINEAR DIFFUSION EQUATIONS*

C. ATKINSON,† G. E. H. REUTER‡ AND C. J. RIDLER-ROWE‡

Abstract. Traveling wave solutions are discussed for nonlinear diffusion equations where the nonlinearity occurs in the diffusion flux as well as in a source term. For a variety of nonlinear diffusion fluxes it is shown that wave solutions exist if and only if the wave speed is greater than some critical value. This critical value is determined explicitly in some special cases, and inequalities are derived for the general case.

1. Introduction. Since the classical paper by Kolmogorov, Petrovsky and Piscounov [11], there has been much work on wave solutions to nonlinear reaction-diffusion equations of the type

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(D\frac{\partial u}{\partial x}\right)+F(u). \tag{1}$$

In most of this work—for recent surveys, see Aronson and Weinberger [3], Fife [5]—the diffusion term in (1) is linear in the sense that D is a constant, and the nonlinearity occurs only in $F(u)$. The reader may recall that in the special case $D\equiv 1$, $F(u)=u(1-u)$ considered by Kolmogorov et al., they look for a wave solution $u=u(X)=u(x+at)$ with $0\leqq u\leqq 1$, $u(-\infty)=0$, $u(+\infty)=1$, $u(X)$ increasing in X and find that such solutions exist if and only if $a\geqq 2$. Thus there is a critical velocity, 2, below which waves do not exist. The major part of their work is then concerned with whether solutions of (1) with given initial conditions approach travelling waves as $t\to\infty$.

We shall concern ourselves with the existence of travelling waves when there is nonlinearity in the diffusion term of (1) because D is a function of u or of $|\partial u/\partial x|$, and shall find in the cases we consider that there is again a critical velocity a^* such that wave solutions of (1) exist above but not below a^*. We note that when D is a constant and $F(u)=Mu(1-u)$, we can always rescale t and x to obtain $D\equiv 1$ and $M=1$, and will always assume that analogous normalizations have been performed, in particular that $F(u)=u(1-u)$. We shall not tie ourselves down to particular applications, but note that (1) with $D=D(u)=u$ occurs in models of population growth considered by Gurney and Nisbet [7] and with $D(u)=u^n$, $n>0$, in Gurtin and MacCamy [8].

We shall consider wave solutions of (1), that is solutions of the form $u=u(X)=u(x+at)$ with $0\leqq u(X)\leqq 1$. We shall only consider the case $a>0$, (for it is easy to see that if $u=u(x+at)$ is a wave solution, then since D is isotropic, $v(x,t)=u(-x+at)$ gives another wave solution travelling in the opposite direction). Furthermore $du/dX>0$ when $0<u<1$ in all the cases which we consider; this is proved for one case in § 3, but the very similar proofs for the other cases are omitted. We shall concentrate mainly on solutions which are strictly increasing (that is throughout $-\infty<X<\infty$, so that $0<u(X)<1$) with $u(-\infty)=0$, $u(+\infty)=1$. Then $u=u(X)$ satisfies the differential equation

$$a\frac{du}{dX}=\frac{d}{dX}\left(D\frac{du}{dX}\right)+u(1-u),\qquad -\infty<X<\infty. \tag{2}$$

* Received by the editors May 7, 1980, and in revised form October 31, 1980.

† Department of Mechanical Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15261.

‡ Department of Mathematics, Imperial College of Science and Technology, London SW7 2BZ, England.

In §§ 2 and 3 we shall look at the case $D = u^n (n > 0)$ and show that there exists a critical velocity a^* such that strictly increasing wave solutions, with $u(-\infty) = 0$, $u(+\infty) = 1$, exist if and only if $a > a^*$, we obtain the explicit evaluation $a^* = \sqrt{\frac{1}{2}}$ when $n = 1$, and estimates for a^* when $n \neq 1$. In § 3 we also show that in the form

$$a\frac{du}{dX} = \frac{1}{n+1}\frac{d}{dX}\left(\frac{d}{dX}u^{n+1}\right) + (1-u)u,$$

(2) admits a weak solution in the sense introduced by Oleinik, Kalashnikov and Chzhou Yui-Lin [12], and employed, for example, by Aronson [1]; for this, one requires absolute continuity of u and $(d/dX)u^{n+1}$ for $-\infty < X < \infty$. This weak solution is strictly increasing from 0 to 1 on a semi-infinite interval $[X_0, \infty)$.

We shall also look at one case in which D is a function of $|\partial u/\partial x|$, namely $D = |\partial u/\partial x|^{N-1}$, $N > 1$. Weak solutions of diffusion equations with nonlinear fluxes of this kind have been considered by Atkinson and Bouillet [4]. Note that properties of the equation considered by Atkinson and Bouillet have similar characteristics to those considered in [7] and [8] and hence may have the same potential for use in population growth models. Since it will be shown that $du/dx \geqq 0$, we can ignore the modulus sign and look at the ordinary differential equation

$$a\frac{du}{dX} = \frac{d}{dX}\left[\left(\frac{du}{dX}\right)^N\right] + u(1-u), \qquad -\infty < X < \infty. \tag{3}$$

We again prove the existence of the critical velocity a^* such that if $a > a^*$ there exist strictly increasing wave solutions $u(X)$ with $u(-\infty) = 0$ and $u(+\infty) = 1$, but not if $0 < a \leqq a^*$. In this case when $a = a^*$ (3) has a solution which is strictly increasing from 0 to 1 on a semi-infinite interval $[X_0, \infty)$. Finally we shall look at $D = (1-u)^{-1}$ and more generally at $D = (1-u)^{-\gamma}$, $0 \leqq \gamma \leqq 1$, finding in all these cases that wave solutions exist if and only if $a \geqq 2$. The motives for looking at $D = (1-u)^{-1}$ are that even when the source term $F(u)$ in (1) is absent, there are then wave solutions for all $a > 0$. This is easily shown by integrating (2), with $u(1-u)$ absent, explicitly; a solution, with $X = x + at$, is

$$u = (1 + e^{-aX})^{-1}$$

if we standardize the X-origin to make $u(0) = \frac{1}{2}$.

We shall treat the ordinary differential equations (2) or (3) by converting them into a system of first order equations involving u and a second unknown related to du/dX.

2. $D(u) = u^n$ with $n = 1$. We now consider the special case $D(u) = u$, $F(u) = u(1-u)$ and look for a solution of

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right) + F(u) \tag{4}$$

of the form $u = u(x + at) = u(X)$, $X = x + at$. The function $u(X)$ is to be strictly increasing in X, with $u(-\infty) = 0$, $u(+\infty) = 1$. Inserting $u = u(X)$ into (4), and defining v by

$$\frac{du}{dX} = av, \tag{5}$$

where we assume $du/dX > 0$ (proved in § 3), we then get

$$a\frac{d(uv)}{dX} = a^2 v - F(u),$$

$$\frac{d(uv)}{du} = \frac{v - Au(1-u)}{v}, \quad \text{where } A = a^{-2}.$$

Hence

$$uv\frac{dv}{du} = v(1-v) - Au(1-u). \tag{6}$$

We now look at (6) over the range $0 < u < 1$, and require $v > 0$. Since u is the running variable, we shall have to check that X runs over the full range $(-\infty, \infty)$. From $dX = du/av$, this amounts to requiring that

$$\int_{0+}^{1/2} \frac{du}{v} = +\infty, \qquad \int_{1/2}^{1-} \frac{du}{v} = +\infty. \tag{7}$$

We now convert (6) into the plane system

$$\frac{du}{ds} = uv, \qquad \frac{dv}{ds} = v(1-v) - Au(1-u) \tag{8}$$

and look for a solution that stays in $0 \leqq u \leqq 1$, $v \geqq 0$, starts from (0, 0) (as $s \to -\infty$) and runs into (1, 0) (as $s \to +\infty$). Note that the system has three critical points at (0, 0), (0, 1), (1, 0). Of these, (1, 0) is a saddle point, and there is a unique solution path S running into it from the region $0 < u < 1$, $v > 0$, with slope tending to a negative limit as S approaches (1, 0). On the other hand, (0, 0) is a degenerate critical point in the sense that its linear approximation has one zero eigenvalue. We therefore investigate what happens to S when it is followed backwards (for decreasing u) from (1, 0), and try to discover for what range of A it approaches (0, 0) as $u \to 0+$, equivalently as $s \to -\infty$.

First, a simple calculation shows that the slope of S at (1, 0) is $-\alpha$, where

$$\alpha = \sqrt{(A + \tfrac{1}{4})} - \tfrac{1}{2}, \tag{9}$$

so that $\alpha < 1$, $= 1$, > 1 according as $A < 2$, $= 2$, > 2. Next, note that when $A = 2$ so that $\alpha = 1$, $v = 1 - u$ is an exact solution to (6) and is then the equation of S; thus $v \to 1$ as $u \to 0$ on S and the first of the conditions in (7) is broken. Now suppose that $A < 2$, so that $\alpha < 1$; choose β with $\alpha < \beta < 1$ and compare S with the line L: $v = \beta(1-u)$. Since S has slope $-\alpha$ at (1, 0), S lies below L for u near 1. We assert that S stays below L for $0 < u < 1$, because otherwise let the last crossing before $u = 1$ occur at $u_0 < 1$. Then $(d/du)[v - \beta(1-u)]$ should be $\leqq 0$ at $u = u_0$. But from (6) this is

$$\begin{aligned}\frac{1-v}{u_0} - A\frac{1-u_0}{v} + \beta &= \frac{1-\beta(1-u_0)}{u_0} - \frac{A}{\beta} + \beta \\ &= \frac{1-\beta}{u_0} + 2\beta - \frac{A}{\beta} > (1-\beta) + 2\beta - \frac{A}{\beta} \\ &= 1 + \beta - \frac{A}{\beta} > 1 + \alpha - \frac{A}{\alpha} = 0.\end{aligned}$$

So on S, $v<\beta(1-u)$ in the whole range $0<u<1$. The last inequality is true for all β satisfying $\alpha<\beta<1$. Hence when $A<2$,

$$v \leqq \alpha(1-u) \quad \text{for } 0<u<1. \tag{10}$$

Similarly one can show that when $A>2$,

$$v \leqq \alpha(1-u) \quad \text{for } 0<u<1. \tag{11}$$

In the case $A>2$ (that is $\alpha>1$), (11) shows that

$$\frac{dv}{du}=\frac{1-v}{u}-A\frac{1-u}{v}\leqq -\frac{\alpha-1}{u}+O(1) \quad \text{as } u\to 0+$$

and hence $v\to+\infty$ as $u\to 0+$. Thus the first of conditions (7) is broken when $A>2$, as happened also when $A=2$.

Next we asume that $A<2$ (that is, $\alpha<1$) and show that there is now a suitable solution leading from (1, 0) to (0, 0). It is easy to see that the region $(0<u<1, v>0)$ is divided into parts, in each of which the slope $dv/du=(1-v)/u-A(1-u)/v$ has constant sign, by two curves on which the slope is zero (see Fig. 1); let C denote the zero slope curve which enters (0, 0). If one follows S backwards (for decreasing u) from (1, 0), S starts in a region of negative slope and must cross C at some point with u-coordinate between 0 and $\frac{1}{2}$, on appealing to (10) in the case $1<A<2$; S then enters a region of positive slope and stays there for all smaller $u>0$. Hence the solution has a nonnegative limit, v_0 say, as $u\to 0+$. But if $v_0>0$ then S satisfies

$$\frac{dv}{du}\sim\frac{1-v_0}{u} \quad \text{as } u\to 0+,$$

giving a contradiction, namely $v\to-\infty$ as $u\to 0+$. Hence $v\to 0$ as $u\to 0+$, which establishes the existence of a unique solution S leading from (1, 0) to (0, 0).

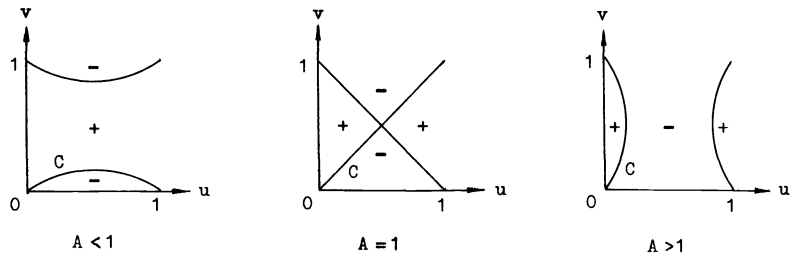


FIG. 1. *Signs of* $\dfrac{dv}{du}-\dfrac{1-v}{u}-A\dfrac{1-u}{v}$.

For $A<2$ it remains to check the first of conditions (7), (the second being dealt with by the approximate linearity of S near the saddle point (1, 0)). Since S lies above the zero slope curve C near (0, 0) and C has slope A at (0, 0), $\liminf_{u\to 0+} v/u \geqq A$. We will now show that $\limsup_{u\to 0+} v/u \leqq A$, so that $v/u\to A$ as $u\to 0+$ and the first of conditions (7) is satisfied. Choose any $k>A$ and consider the line $T: v=ku$. If at any point, S lies on or above T, then S has slope

$$\begin{aligned}\frac{1-v}{u}-A\frac{1-u}{v}&\geqq\frac{1-v}{u}-\frac{A}{ku}+\frac{A}{k}\\ &\sim\frac{k-A}{ku} \quad \text{as } u\to 0+.\end{aligned} \tag{12}$$

Hence there exists some $\delta > 0$ such that if S and T cross, then

$$\text{slope } S > \text{slope } T \quad \text{when } 0 < u < \delta.$$

Consequently there are three possibilities: either (i) S lies below T whenever $0 < u < \delta$, or (ii) S crosses T for some u_1 between 0 and δ, and then S stays below T for $0 < u < u_1$, or else (iii) S lies above T whenever $0 < u < \delta$. However if (iii) is true, we obtain a contradiction from (12), namely $v \to -\infty$ as $u \to 0+$. Hence (i) or (ii) is true so that $\limsup_{u \to 0+} v/u \leqq k$. The choice of k implies that $\limsup_{u \to 0+} v/u \leqq A$, as required.

By extending the "crossing argument" just given we will show that the solution curve S has an asymptotic expansion $\sum_1^\infty b_n u^n$ as $u \to 0+$. The existence of the second term of this expansion is enough to ensure that the wave solution $u(X)$ of (4) is exponentially small as $X \to -\infty$, and in fact $u(X) \exp(-X/a)$ has a positive limit. Likewise we should mention that the saddle point behavior near (1, 0) in the u-v plane ensures that $[1 - u(X)] \exp(\alpha a X)$ has a positive limit as $X \to +\infty$. Temporarily assuming that the expansion does exist, the differential equation (6) implies that the coefficients are

$$b_1 = A, \qquad b_2 = 2A^2 - A, \tag{13}$$
$$b_n = 2b_1 b_{n-1} + 3b_2 b_{n-2} + \cdots + n b_{n-1} b_1, \qquad n \geqq 3.$$

If $A > \frac{1}{2}$ the expansion in fact diverges since all its coefficients are positive and $b_n \geqq (n+2) b_1 b_{n-1}$ for $n \geqq 3$.

Our first step in proving that the asymptotic expansion exists is to show that, given any sufficiently large $\lambda > 0$, S is bounded above by $Au(1 + \lambda u)$ for sufficiently small $u > 0$. To do this one may replace T in the previous crossing argument by $Au(1 + \lambda u)$ and use the fact that, given any $k > A$, S is bounded above by ku for sufficiently small $u > 0$. The argument then is like that in the next step in the proof, and we go directly to this next step since it indicates better the general induction argument which completes the proof.

We now show that a closer upper bound for S near $u = 0$ is obtained by taking T to be the curve $Au(1 + \rho u + \mu u^2)$ in the crossing argument. Here we intend to show that ρ may be chosen to make the slopes of S and T agree in powers of u up to the constant term at any crossing, and choose $\mu > 0$ suitably large. From the previous step we may assume that for some $\lambda > 0$, S satisfies

$$v < Au(1 + \lambda u) \quad \text{for sufficiently small } u > 0. \tag{14}$$

If at any point S lies on or above T, then S has slope

$$\frac{1-v}{u} - A\frac{1-u}{v} \geqq \frac{1}{u} - A(1 + \lambda u) - \left(\frac{1}{u} - 1\right)\left\{1 + \sum_{n=1}^{\infty} (-\rho u - \mu u^2)^n\right\} \tag{15}$$

$$= -A + 1 + \rho + \mu u + O(u) + O(u^2) \quad \text{as } u \to 0+ \tag{16}$$

in the sense that the term $O(u)$ does not depend on μ while $O(u^2)$ does. Hence on defining ρ by $\rho = 2A - 1$, and making any suitably large choice of μ, we have slope $S >$ slope T at any crossing of S and T for small $u > 0$. If possibility (iii) of the crossing argument were true, then integrating the lower bound (16) for slope S from 0 to u would (since $\rho = 2A - 1$) contradict the upper bound (14) provided that μ were initially chosen large enough. Then only possibilities (i) and (ii) of the crossing argument remain. Hence given any sufficiently large $\mu > 0$, S is bounded above by $Au(1 + \rho u + \mu u^2)$ for sufficiently small $u > 0$.

Similarly, further upper bounds are obtained inductively in the form $Au(1+\rho_1 u+\cdots+\rho_j u^j+\mu_j u^{j+1})$ for u near 0, $(\rho_1=\rho,\ \mu_1=\mu)$. The coefficients μ_j are merely chosen sufficiently large, but the crucial step is that the coefficients ρ_j are successively determined by the condition that the slopes of S and T agree at any intersection up to terms of order u^{j-1}, each new coefficient ρ_j appearing via the first term of a geometric expansion in the same way as in (15). Lower bounds of the form $Au(1+\rho_1 u+\cdots+\rho_j u^j+\nu_j u^{j+1})$ may be obtained similarly, and hence the asymptotic expansion follows (with, in the notation of (13), $b_n=A\rho_{n-1}$ for $n\geqq 2$).

It is interesting to note that when $A=\frac{1}{2}$, the coefficients b_n for $n\geqq 2$ are all zero, so that the expansion reduces to the single term Au. It is easily checked that this is an exact solution to (6), but we also know that there is another solution v which $\to 0$ as $u\to 0+$ and as $u\to 1-$. In fact from the previous arguments one can see that for each $A>0$, all solutions starting near (0, 0) in the positive quadrant of the u-v plane must enter (0, 0) with slope A and the same asymptotic expansion; furthermore any two such solutions differ by an exponentially small amount as $u\to 0+$. For if v_1 and v_2 are two such solutions and $w=v_1-v_2$ then

$$\frac{dw}{du}=-\frac{w}{u}-A(1-u)\left(\frac{1}{v_1}-\frac{1}{v_2}\right)=w\left\{-\frac{1}{u}+\frac{A(1-u)}{v_1 v_2}\right\}.$$

Hence

$$\frac{1}{w}\frac{dw}{du}\sim\frac{1}{Au^2}\quad\text{as } u\to 0+,$$

so that on integrating from u to u_0 (both small with u_0 fixed) one obtains

$$w(u)=w(u_0)\,e^{-f(u)}\quad\text{where } f(u)\sim 1/Au\quad\text{as } u\to 0+.$$

We can now summarize the main results of this section as follows.

THEOREM. *If $a>\sqrt{\frac{1}{2}}$, equation* (4) *has a unique wave solution $u(x+at)=u(X)$, $X=x+at$, such that $u(X)$ is strictly increasing in X for $-\infty<X<\infty$, and $X\to-\infty,+\infty$ respectively as $u\to 0+,\ 1-$; if $0<a\leqq\sqrt{\frac{1}{2}}$ no such solution exists. When $a>\sqrt{\frac{1}{2}}$ the wave solution has asympotically exponential tails as $X\to\pm\infty$. $du/dX=av$ has an asymptotic expansion $a\sum_1^\infty b_n u^n$ as $u\to 0+$, whose coefficients b_n are given in* (13) *(with $A=a^{-2}$).*

3. $D(u)=u^n$, $n>0$. We now consider the general case $D(u)=u^n$, $n>0$, with $F(u)=u(1-u)$. In looking for solutions $u=u(x+at)$ it will again turn out that there is a critical value a^* such that strictly increasing wave solutions exist if and only if $a>a^*$; when $a=a^*$ it will be seen that a weak solution exists in the sense mentioned in § 1. Though we can no longer give an explicit formula for a^*, (recall that $a^*=\sqrt{\frac{1}{2}}$ when $n=1$), we will give estimates for a^*.

We will first show that $du/dX>0$ when $0<u<1$ by an argument like that of Fife and McLeod [6], and then prove the existence of a^*. The equation to be considered is

$$a\frac{du}{dX}=\frac{d}{dX}\left(u^n\frac{du}{dX}\right)+u(1-u),\qquad -\infty<X<\infty. \tag{17}$$

This may conveniently be transformed into a first order equation on defining the variable Y by

$$aY=u^n\frac{du}{dX}, \tag{18}$$

which leads to the equation

$$\frac{dY}{du} = 1 - \frac{Au^{n+1}(1-u)}{Y} \tag{19}$$

where $A = a^{-2}$. Now if $du/dX < 0$ for any value of X we have $Y < 0$ in (19) so that $dY/du > 1$; if we then follow the solution backwards (i.e., for u decreasing) we still have $Y < 0$, $dY/du > 1$ and hence $Y \to c < 0$ as $u \to 0+$. This leads to $u \sim [a|c|(X_0 - X)]^{1/n+1}$ as $X \to$ some $X_0 -$, and does not give an admissible solution (even as a weak solution). It follows that we should require $Y > 0$ (and so $du/dX > 0$) over the range $0 < u < 1$, and we shall seek solutions such that $X = x + at$ runs over the full range $(-\infty, \infty)$, with $u(-\infty) = 0$, $u(+\infty) = 1$. Thus we shall look for a solution Y running from (0, 0) to (1, 0), both of which are critical points of equation (19).

At (1, 0), (19) has a saddle point and there is a unique solution, S say, running into (1, 0) from the region $0 < u < 1$, $Y > 0$. S has slope $-\alpha$ at (1, 0) where $\alpha = \sqrt{(A + \frac{1}{4})} - \frac{1}{2}$. From (19), dY/du is positive, zero or negative in the region $0 < u < 1$, $Y > 0$ according as Y lies above, on or below the curve C defined by

$$Y = Au^{n+1}(1-u), \qquad 0 < u < 1. \tag{20}$$

Hence, on following S backwards from (1, 0) (with u decreasing), S initially has negative slope and lies below C, but must eventually cross C and then remain above C in a region of positive slope for all smaller $u > 0$. Thus $Y(u) > 0$ on S for $0 < u < 1$, and $Y(0) \geqq 0$ is well-defined by continuity.

We shall show that at each u in [0, 1), S descends as A decreases and there is some positive critical value A^* of A such that for $A > A^*$, $Y(0) > 0$ on S, while for $A = A^*$, S enters (0, 0) with behavior $Y \sim u$ as $u \to 0+$. Using (18) it is then easily seen that no admissible wave solution can exist for $A > A^*$, that is for $0 < a < a^*$ where $a^* = (A^*)^{-1/2}$. For $0 < A < A^*$ we shall show that S enters (0, 0), but now with the behavior $Y \sim Au^{n+1}$ as $u \to 0+$. From this, the approximate linearity of S near the saddle point (1, 0) and (18), it follows that a unique strictly increasing solution of (17) exists for $0 < A < A^*$, that is for $a > a^*$. However, for $a = a^*$ it can be seen that the curve S actually gives a weak solution in the sense mentioned in § 1. For $n = 1$, so that $a^* = \sqrt{\frac{1}{2}}$, this can be shown by explicit calculation. If we let $Y = uv$, then as remarked after (9) in § 2 we can take $v = 1 - u$, i.e., $du/dX = a(1-u)$. It then follows easily that for $a = \sqrt{\frac{1}{2}}$ we have a wave solution $u = 0$ for $X \leqq X_0$, $= 1 - e^{-a(X - X_0)}$ for $X > X_0$, with the property that u and $d(u^2)/dX$ are absolutely continuous. Its only defect is that du/dX fails to exist at $X = X_0$.

The fact that S descends as A decreases follows from the behavior of its slope at (1, 0) and a simple comparison argument using equation (19). It can also be shown by arguments similar to those of Johnson [9, pp. 48–49] that, given any u_0 with $0 < u_0 < 1$, $Y(u_0)$ on S depends continuously on $A \geqq 0$. (Here we define S by $Y \equiv 0$ when $A = 0$.) Furthermore, $Y(0) > 0$ on S for all sufficiently large A. For suppose $Y(0) = 0$, so that $Y(u) < u$ for u in (0, 1] from (19). Then write (19) as

$$YY' = Y - Au^{n+1}(1-u),$$

and integrate from 0 to 1 to obtain

$$A \int_0^1 u^{n+1}(1-u)\, du = \int_0^1 Y(u)\, du < \int_0^1 u\, du = \frac{1}{2},$$

which is impossible for A sufficiently large.

We next consider the differential equation (19) near (0, 0). A contraction mapping argument shows that for each $A \geqq 0$, there is a solution T emerging from (0, 0) with the behavior $Y \sim u$ as $u \to 0+$, and given any c satisfying $0 < c < 1$, T is the unique solution lying in the wedge $cu \leqq Y \leqq u$ for all small enough $u > 0$. To construct T one may rewrite (19), with the requirement $Y(0) = 0$, as $Y = \theta Y$ where

$$\theta Y(u) = \int_0^u \left\{1 - As^{n+1} \frac{1-s}{Y(s)}\right\} ds.$$

Now let any c be given with $0 < c < 1$. One can show that there is some $\delta > 0$ such that if $Y(u)$ lies between cu and u (inclusively) for $0 < u < \delta$, then so does $\theta Y(u)$. Next put $Y_0 \equiv u$ and $Y_{m+1} = \theta Y_m$, $m \geqq 0$. Then one can deduce that for some positive $\delta_1 (< \delta)$ and $K < 1$, $\Delta_{m+1} \leqq K \Delta_m$, $m \geqq 1$, where $\Delta_m = \sup \{|Y_m(u) - Y_{m-1}(u)|: 0 < u \leqq \delta_1\}$. By standard arguments Y_m converges for $0 < u \leqq \delta_1$ to a solution path T whose above-mentioned properties are easily checked. As with S it is easy to see that T rises as A decreases.

We now compare S and T near $u = 0$ as A varies. We know that $Y(0) > 0$ on S for A large enough, and S rises as A increases. Thus the values of A for which $Y(0) > 0$ on S form an infinite interval whose infimum we denote by $A^* \geqq 0$. For $A > A^*$, T lies below S (otherwise one would have $Y(0) = 0$ on S). Hence, since S descends and T rises as A decreases, it can be seen that T_1, the solution T corresponding to any fixed $A_1 > A^*$, provides a lower bound on some interval $0 < u < \delta_1$ for *all* solutions S with $A > A^*$; by the continuity of S with respect to $A \geqq 0$, T_1 does likewise for S^*, the solution S corresponding to A^*. Hence

$$\liminf_{u \to 0+} \frac{Y(u)}{u} \geqq 1 \quad \text{on } S^*, \tag{21}$$

which implies that $A^* > 0$ (because $Y \equiv 0$ when $A = 0$). It now follows that $Y(0) = 0$ on S^*, for otherwise $Y(0) > 0$ on S^* and then, by continuity, also on S for some $A < A^*$. From (21) and the fact that $dY/du < 1$ we obtain $Y \sim u$ as $u \to 0+$ on S^*. Hence we may conclude, using the uniqueness of T already mentioned, that S and T must merge when $A = A^*$.

Now suppose that $0 < A < A^*$. Since S descends and T rises as A decreases, it follows immediately that S now lies below T (for S and T merge when $A = A^*$). Also S lies above the zero slope curve (20) near (0, 0), and hence S enters (0, 0). We now show that

$$Y \sim Au^{n+1} \quad \text{on } S \text{ as } u \to 0+, \tag{22}$$

which will complete our main argument on the existence of A^*. For any $\lambda > A$ a crossing argument similar to those in § 2 shows that, near $u = 0$, S lies below the curve Γ: $Y = \lambda u^{n+1}(1-u)$, (which is a multiple of the zero slope curve (20)). If S stays above Γ near $u = 0$, $dY/du > 1 - A/\lambda$ so that $Y > (1 - A/\lambda)u$ on S near $u = 0$, and then, by the uniqueness property of T, S and T would merge—a contradiction; further details of this crossing argument are omitted. Since $\lambda (> A)$ is arbitrary, and S lies above the zero slope curve (20) near (0, 0), (22) follows. Our arguments also show that for each $A > 0$ any solution lying between T and the u-axis satisfies $Y \sim Au^{n+1}$ as $u \to 0+$; by an argument like that ending § 2, any two such solutions differ by an exponentially small amount as $u \to 0+$.

Having established the existence of a critical value A^* of A, it remains to estimate A^* (in terms of n) by suitable inequalities. We could do this by quoting results from

Johnson and Nachbar [10] for a closely related problem, but will give a slightly simpler argument for our special choices of $D(u)=u^n$, $F(u)=u(1-u)$. It is convenient here to define the variable v by $u^n\,du/dX=auv$ and rewrite (17) as

$$2vv'=\frac{2v(1-v)}{u}-2Au^{n-1}(1-u),$$

writing $'$ for d/du. (This variable v was used in our early investigations of the existence of A^*.) Note that $Y=uv$. Taking $A=A^*$, there is a solution v with $v(0+)=1$, $v(1-)=0$, and $0<v<1$ for $0<u<1$ (since (19) gives $dY/du<1$, so that $0<Y<u$). Hence

$$2vv'>-2A^*u^{n-1}(1-u)$$

and, on integrating from $u=0$ to $u=1$,

$$-1>-2A^*\int_0^1 u^{n-1}(1-u)\,du=-2A^*\left(\frac{1}{n}-\frac{1}{n+1}\right)=-\frac{2A^*}{n(n+1)}$$

whence

$$A^*>\frac{n(n+1)}{2}.$$

On the other hand, for $A<A^*$ we have a solution v with $v(0+)=v(1-)=0$. Then

$$\begin{aligned}(uv^2)'&=2uvv'+v^2=2v-2v^2+v^2-2Au^n(1-u)\\&=1-(1-v)^2-2Au^n(1-u)\\&\leqq 1-2Au^n(1-u).\end{aligned}$$

Integrate from 0 to 1 to obtain

$$0\leqq 1-2A\int_0^1 u^n(1-u)\,du=1-\frac{2A}{(n+1)(n+2)},$$

$$A\leqq\frac{(n+1)(n+2)}{2}.$$

Letting $A\uparrow A^*$, we get

$$A^*\leqq\frac{(n+1)(n+2)}{2}.$$

Thus

$$\frac{n(n+1)}{2}\leqq A^*\leqq\frac{(n+1)(n+2)}{2}.$$

Note that for $n=1$ this gives $1<A^*<3$ compared with the known result $A^*=2$ (from § 2).

4. $D(u)=|\partial u/\partial x|^{N-1}$, $N>1$. In this section we deal with the case $D(u)=|\partial u/\partial x|^{N-1}$, $N>1$ and $F(u)=u(1-u)$. $N=1$ gives $D(u)\equiv 1$, which is the classical KPP equation to be covered by § 5. As mentioned in § 1 we can rule out the possibility $du/dX<0$ by an argument similar to that in § 3 (here using the substitution $Y=-|du/dX|^{N-1}\,du/dX$) and so seek a wave solution with $u=u(x+at)=u(X)$ with $u(X)$

strictly increasing from 0 to 1 as X runs from $-\infty$ to $+\infty$. Equation (2) becomes

$$a\frac{du}{dX}=\frac{d}{dX}\left(\left(\frac{du}{dX}\right)^N\right)+u(1-u), \qquad -\infty<X<\infty.$$

Putting $aY=(du/dX)^N$, we obtain

$$\frac{dY}{du}=1-\frac{Au(1-u)}{Y^{1/N}}, \qquad A=a^{-(1+1/N)}, \tag{23}$$

and look for a solution running from (0, 0) to (1, 0) in the region $0<u<1$, $Y>0$. We also need X to run from $-\infty$ to $+\infty$, and therefore require that

$$\int_{0+} Y^{-1/N}\,du=+\infty, \qquad \int^{1-} Y^{-1/N}\,du=+\infty. \tag{24}$$

Indicating proofs in outline only, we shall show that, as in the preceding case, there is a positive critical value A^* of A such that this problem has a unique strictly increasing solution if and only if $A<A^*$, that is if and only if $a>a^*$, where $a^*=(A^*)^{-N/(N+1)}$. When $A=A^*$ there is a solution $u(X)$ strictly increasing from 0 to 1 on a semi-infinite interval $[X_0,\infty)$. Again we have no explicit general formula for A^*, but shall give estimates.

Existence and uniqueness for solutions of (23) entering (1, 0) from the region $0<u<1$, $Y>0$ can be proved by imitating the methods of Johnson [9]. Firstly rewrite (23) as

$$\frac{d}{du}Y^{(N+1)/N}=\frac{N+1}{N}(Y^{1/N}-Au(1-u))$$

or, with $Z=Y^{(N+1)/N}$,

$$\frac{dZ}{du}=\frac{N+1}{N}(Z^{1/(N+1)}-Au(1-u)). \tag{25}$$

It then follows from simple arguments based on the Peano existence theorem that (25) has a solution with $Z=0$ at $u=1$, Z existing for $0\leqq u\leqq 1$, and $Z>0$ for $0<u<1$. Further arguments in the manner of Johnson [9, pp. 48–49] show that Z is unique and depends continuously on $A\geqq 0$ for each fixed u in [0, 1). The same facts hold for $Y=Z^{N/(N+1)}$, which gives a solution path S for (23). Note also that when $A=0$, $Z=0$ is the unique solution of (25) in $0\leqq u\leqq 1$ with $Z(1)=0$.

To find the behavior of the solution path S as $u\to 1-$, a simple comparison of S and its slope with the curve $[\lambda u(1-u)]^N$ and its slope (for $\lambda<A$ and $\lambda\geqq A$) shows that $Y\sim[A(1-u)]^N$ on S as $u\to 1-$. It then follows that S satisfies the second of conditions (24).

One may now proceed using arguments quite similar to those for the case $D(u)=u^n (n>0)$—so details are omitted— and conclude that the critical value A^* of A exists, as required. When $A=A^*$, it can be seen that it is again the curve S which gives a solution strictly increasing from 0 to 1 on a semi-infinite interval $[X_0,\infty)$; the corresponding solution $u=u(X)$ can be shown to be valid in the classical sense because du/dX and (d/dX) $((du/dX)^N)$ exist even at $X=X_0$.

It remains to obtain estimates for A^*. A convenient variable for this is v, defined by $v=Y/u$ (as in the case $D=u^n$), and then (23) becomes

$$uv'=1-v-\frac{Au(1-u)}{(uv)^{1/N}},$$

or

$$\frac{N}{N+1}(v^{1+1/N})' = \frac{v^{1/N}(1-v)}{u} - Au^{-1/N}(1-u), \tag{26}$$

where $'$ denotes d/du. Now assume that $A = A^*$, so that S runs from $(1, 0)$ to $(0, 0)$ with behavior $Y \sim u$ as $u \to 0+$, and $v = Y/u$ runs from $(1, 0)$ to $(0, 1)$. Note that $0 \leqq v \leqq 1$, since $0 \leqq Y \leqq u$ on S as in § 3. Then from (26)

$$\frac{N}{N+1}(v^{1+1/N})' \geqq -A^* u^{-1/N}(1-u)$$

so that on integrating from 0 to 1,

$$-\frac{N}{N+1} \geqq -A^* \int_0^1 u^{-1/N}(1-u)\, du,$$

giving $A^* \geqq (N-1)(2N-1)/N(N+1)$.

Now rewrite (26) as follows:

$$\begin{aligned}\frac{N}{N+1}(uv^{1+1/N})' &= \frac{N}{N+1}\{u(v^{1+1/N})' + v^{1+1/N}\} \\ &= v^{1/N}(1-v) - Au^{1-1/N}(1-u) + \frac{N}{N+1}v^{1+1/N} \\ &= v^{1/N} - \frac{1}{N+1}v^{1+1/N} - Au^{1-1/N}(1-u).\end{aligned}$$

But $v^{1/N} - (1/(N+1))v^{1+1/N}$ is increasing in v for $0 \leqq v \leqq 1$, and so is bounded above by its value at 1, namely $N/(N+1)$. Hence

$$\frac{1}{N+1}(uv^{1+1/N})' \leqq \frac{N}{N+1} - Au^{1-1/N}(1-u).$$

Then taking $A = A^*$ again, $uv^{1+1/N} = 0$ at $u = 0, 1$, so that integrating the last inequality gives

$$0 \leqq \frac{1}{N+1} - A^* \int_0^1 u^{1-1/N}(1-u)\, du.$$

Hence $A^* \leqq (2N-1)(3N-1)/N(N+1)$.

5. $D(u) = (1-u)^{-\gamma}$, $0 \leqq \gamma \leqq 1$. We finally turn to the case $D(u) = (1-u)^{-\gamma}$, $0 \leqq \gamma \leqq 1$, and $F(u) = u(1-u)$, where $\gamma = 0$ gives the classical KPP equation. We seek a wave solution $u = u(X) = u(x + at)$ of

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(D(u)\frac{\partial u}{\partial x}\right) + F(u)$$

with the same conditions as before. Defining Y by $D(u)\, du/dX = aY$, we obtain

$$\frac{dY}{du} = 1 - \frac{Au(1-u)^{1-\gamma}}{Y}, \qquad A = a^{-2}, \tag{27}$$

and look for a solution with $Y > 0$ for $0 < u < 1$ running from $(0, 0)$ to $(1, 0)$ in the uY-plane, checking that X runs over the full range $(-\infty, \infty)$.

Assume first that $0 \leqq \gamma < 1$. Then (27) has critical points at (0, 0), (1, 0). For $A > \frac{1}{4}$, (0, 0) is a spiral singularity so that any solution approaching (0, 0) will leave the region $0 < u < 1$, $Y > 0$, and we may therefore asume that $0 < A \leqq \frac{1}{4}$. It is known from the work of Johnson [9] that there is a unique solution S approaching (1, 0) from the region $0 < u < 1$, $Y > 0$, and we now show that S followed backwards (for decreasing u) approaches (0, 0) as $u \to 0+$.

To see this, let C be the curve $Y = Au(1-u)^{1-\gamma}$. The solution curve S has positive, zero or negative slope according as S lies above, on or below C. Following S backwards from (1, 0), it lies below S for u near 1 but ultimately crosses C and lies above C for all smaller values of u. Let C' be the curve $Y = 2Au(1-u)^{1-\gamma}$. If S were to cross C' it would then have slope $\frac{1}{2}$ whereas C' has slope

$$2A\{(1-u)^{1-\gamma} - u(1-\gamma)(1-u)^{-\gamma}\} = 2A(1-u)^{-\gamma}\{1-(2-\gamma)u\}$$
$$< 2A(1-u)^{1-\gamma} < 2A \leqq \tfrac{1}{2},$$

so that we should have slope $S >$ slope C'. But S lies below C, hence below C', near $u = 1$ which implies that S must lie below C' for all u in $0 < u < 1$. Thus $Y < 2u(1-u)^{1-\gamma}$ for $0 < u < 1$, which shows that S approaches (0, 0) as $u \to 0+$.

For $\gamma = 1$, still assuming that $0 < A \leqq \frac{1}{4}$, explicit calculation (left to the reader) shows that there is a solution S leading from (0, 0) to (1, 0).

In the range $0 \leqq \gamma \leqq 1$, we have $Y < u$ for all u in $0 < u < 1$, since $dY/du < 1$. Since $dX = (D(u)/aY)\,du$, we have $dX > (\text{constant}/u)\,du$ for small u, so that $X \to -\infty$ as $u \to 0+$. For the behavior of S near (1, 0), hence of X as $u \to 1-$, integrate (27) from u to 1 to obtain

$$Y(u)^2 = 2\int_u^1 [As(1-s)^{1-\gamma} - Y(s)]\,ds$$
$$\leqq 2A\int_u^1 s(1-s)^{1-\gamma}\,ds = O((1-s)^{2-\gamma}).$$

Hence $Y(u) = O((1-s)^{1-1/2\gamma})$ and

$$Y(u)^2 = 2A\int_u^1 (1-s)^{1-\gamma}\,ds + \int_u^1 f(s)\,ds,$$

where

$$f(s) = -2[A(1-s)^{2-\gamma} + Y(s)] = O((1-s)^{1-1/2\gamma}).$$

It follows that

$$Y(u)^2 \sim \frac{2A}{2-\gamma}(1-u)^{2-\gamma} \quad \text{as } u \to 1-,$$

assuming now that $\gamma > 0$. But $dX = (D(u)/aY)\,du$ then shows that

$$dX \sim \frac{\text{constant}}{(1-u)^{1+1/2\gamma}},$$

which ensures that $X \to +\infty$ as $u \to 1-$. ($\gamma = 0$ can be treated by using the fact that (1, 0) is then a saddle point of standard type.)

Acknowledgment. We wish to thank the referee for helpful suggestions and for drawing our attention to the work of Aronson [2].

REFERENCES

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

[2] D. G. ARONSON, *Density dependent interaction diffusion systems*, (in) Proceedings of the Advanced Seminar on Dynamics and Modeling of Reactive Systems, Academic Press, 1980.

[3] D. G. ARONSON AND H. F. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, (in) Partial Differential Equations and Related Topics, Lecture Notes in Mathematics 446, Springer-Verlag, Berlin, 1975.

[4] C. ATKINSON AND J. E. BOUILLET, *Some qualitative properties of solutions of a generalized diffusion equation*, Math. Proc. Cambridge Philos. Soc., 86 (1979), pp. 495–510.

[5] P. C. FIFE, *Asymptotic states for equations of reaction and diffusion*, Bull. Amer. Math. Soc., 84 (1978), pp. 693–726.

[6] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal. 65 (1977), pp. 333–361.

[7] W. S. C. GURNEY AND R. M. NISBET, *The regulation of inhomogeneous populations*, J. Theoret. Biol., 52 (1975), pp. 441–457.

[8] M. E. GURTIN AND R. C. MACCAMY, *On the diffusion of biological populations*, Math. Biosci. 33 (1977), pp. 35–49.

[9] W. E. JOHNSON, *On a first-order boundary value problem from laminar flame theory*, Arch. Rational Mech. Anal., 13 (1963), pp. 46–54.

[10] W. E. JOHNSON AND W. NACHBAR, *Laminar flame theory and the steady burning of a monopropellant*, Arch. Rational Mech. Anal., 12 (1963), pp. 58–92.

[11] A. N. KOLMOGOROV, I. G. PETROVSKY AND N. S. PISCOUNOV, *Étude de l'equation de diffusion avec croissance de la quantité de matière et son application à une problème biologique*, Bull. Univ. État Moscuo (ser. intern.) A1(6) (1937), pp. 1–25.

[12] O. A. OLEINIK, A. S. KALASHNIKOV AND CHZHOU YUI-LIN, *The Cauchy problem and boundary problems for equations of the type of nonstationary filtration*, Izv. Akad. Nauk SSSR Ser. Mat. 22 (1958), pp. 667–704.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0010 $01.00/0

THE QUENCHING OF SOLUTIONS OF SEMILINEAR HYPERBOLIC EQUATIONS*

PETER H. CHANG† AND HOWARD A. LEVINE‡

Abstract. We consider the problem $u_{tt} = u_{xx} + \phi(u(x, t))$, $0 < x < L$, $t > 0$; $u(0, t) = u(L, t) = 0$; $u(x, 0) = u_t(x, 0) = 0$. Assume that $\phi : (-\infty, A) \to (0, \infty)$ is continuously differentiable, monotone increasing, convex, and satisfies $\lim_{u \to A^-} \phi(u) = +\infty$. We prove that there exist numbers L_1 and L_2, $0 < L_1 \leqq L_2$ such that if $L > L_2$, then a weak solution u (to be defined) quenches in the sense that u reaches A in finite time; if $L < L_1$, then u does not quench. We also investigate the behavior of the weak solution for small L and establish the local (in time) existence of u.

1. Introduction. In [3], Kawarada investigated the following nonlinear initial boundary value problem:

$$\text{(P)} \qquad \begin{aligned} &u_t = u_{xx} + \frac{1}{1-u}, && 0 < x < L, \quad t > 0, \\ &u(0, t) = u(L, t) = 0, && t > 0, \\ &u(x, 0) = 0, && 0 \leqq x \leqq L. \end{aligned}$$

There, he established the following interesting results:

(A) If $L > 2\sqrt{2}$, then $u(L/2, t)$ reaches one in finite time.

(B) If $u(L/2, t)$ reaches one in finite time, then $u_t(L/2, t)$ is unbounded in finite time.

Whenever (B) occurs, Kawarada says that *u quenches in finite time.* We shall say that u quenches if (A) occurs. This is a weaker definition than Kawarada's.

In [1] and independently in [6] it was established that there is a number L_0 such that if $L > L_0$, u quenches in finite time while if $L < L_0$, u tends monotonically to the smaller of the two solutions of the stationary problem

$$f''(x) + \frac{1}{1-f(x)} = 0, \qquad 0 < x < L,$$
$$f(0) = f(L) = 0.$$

In [6] it was also shown that this latter situation also obtains at $L = L_0$, where the two stationary solutions coalesce into a single stationary solution. The number L_0 can be found exactly, in fact $L_0 \cong 1.5307 \cdots$. These papers also included extensions to more general nonlinear parabolic problems where the nonlinear term has the same qualitative properties as $1/(1-u)$ (convex, positive, monotone increasing and singular at the right endpoint of $(-\infty, a)$.) The principal tools employed there were the maximum principle and the various comparison theorems derived from it.

* Received by the editors April 2, 1980; and in final form December 12, 1980. Funds for numerical computation provided by the Science and Humanities Research Institute of Iowa State University.

† Department of Mathematics, University of Nebraska at Omaha, Omaha, Nebraska 68182.

‡ Department of Mathematics, Iowa State University, Ames, Iowa 50011. The work of this author was supported in part by the National Science Foundation under Grant MCS 78-02729.

Motivated by the preceding remarks, we were led to examine the analogous problem for the wave equation. That is, we studied the problem

$$
\begin{aligned}
& u_{tt} = u_{xx} + \frac{1}{1-u}, && 0 < x < L, \quad t > 0, \\
\text{(W)} \qquad & u(0, t) = u(1, t) = 0, && t > 0, \\
& u(x, 0) = u_t(x, 0) = 0, && 0 \leqq x \leqq L.
\end{aligned}
$$

Although we do not have any physical application in mind, we believe the study of problem (W) to be of theoretical interest. Since parabolic equations are in some sense on the borderline between elliptic and hyperbolic equations, it is of interest to know which of the properties of their solutions are possessed by solutions of the other two types of equations and what form the properties take in these cases. For example, the maximum principle for parabolic equations has a stronger version for elliptic equations and a much weaker version for hyperbolic equations. See [9] and references therein.

The first result we obtained on this problem is contained in Theorem 3.2. For this problem it says that if $L > L_1 \cong 1.418 \cdots$, then u quenches (reaches one) in finite time. Since $L_1 < L_0$, we conjectured that for any $L > 0$, u must quench in finite time. However, when (W) was solved numerically for small L, the results obtained seemed to contradict this conjecture.

Guided by the numerical results, we were able to show that if $L < 1.238$, then $u \leqq 0.7732$ for all time. That is, if L is small, u cannot quench, even in infinite time. This result is contained in Theorem 4.1.

Because for problem (W) we do not have as useful a maximum principle available, the arguments we use are much different than those used for the parabolic problem.

Rather than studying problem (W), we treat the somewhat more general problem (W').

$$
\begin{aligned}
& u_{tt} = u_{xx} + \varepsilon\varphi(u(x, t)), && 0 < x < 1, \quad t > 0, \quad \varepsilon > 0. \\
\text{(W')} \qquad & u(0, t) = u(1, t) = 0, && t > 0, \\
& u(x, 0) = u_t(x, 0) = 0, && 0 \leqq x \leqq 1,
\end{aligned}
$$

which reduces to (W) when $\varepsilon = L^2$ and $\varphi(u) = 1/(1-u)$, after a change of variables. Here $\varphi : (-\infty, A) \to (0, \infty)$ is continuously differentiable, monotone increasing, convex and satisfies

$$\lim_{u \to A^-} \varphi(u) = +\infty.$$

The solution $u(x, t; \varepsilon)$ for fixed $\varepsilon > 0$, is shown to exist in the weak sense (defined precisely later) on the largest set $[0, 1] \times [0, T)$, where $|u| < A$. If $T = +\infty$, we say that u is a global solution. If $T < \infty$, then $\sup \{u(x, t) : (x, t) \in [0, 1] \times [0, T)\} = A$ and we say u quenches (reaches A) in finite time. If $T = +\infty$, and this supremum is A, u quenches in infinite time. Thus, if u does not quench at all, $u \leqq A(1-\delta)$ for some $\delta \in (0, 1)$, on the half strip.

We then summarize our results for (W') as follows: There exist two numbers ε_1, ε_2, $0 < \varepsilon_1 \leqq \varepsilon_2 < +\infty$ such that if $\varepsilon < \varepsilon_1$, $u(x, t; \varepsilon)$ (the solution of (W')) cannot quench. If $\varepsilon > \varepsilon_2$ then $u(x, t; \varepsilon)$ quenches in finite time. We do *not* prove $\varepsilon_1 = \varepsilon_2$, although we believe this to be the case. The numerical results indicate that this is so for (W) and that $L_1 = \sqrt{\varepsilon_1} = \sqrt{\varepsilon_2} \cong 1.365 \cdots$. Also, we believe that if $\varepsilon = \varepsilon_1 = \varepsilon_2$, then u quenches in infinite time.

The plan of the paper is as follows: In § 2 we define the notion of a weak solution which we shall use in the sequel. We establish local existence there also. In § 3 we show

that if ε is "large" u quenches in finite time whereas in § 4 we show that if ε is "small", u cannot quench at all, even in infinite time. In § 5 we discuss the behavior of u as $\varepsilon \to 0^+$. We conclude with some remarks in the final section.

2. The definition of a weak solution. We say u is a weak solution of (W′) on $D_T \equiv (0, 1)\times(0, T)$ if:

(i) u is continuous in $\bar{D}_T$ and satisfies the initial and boundary conditions there.

(ii) $|u| \leqq A(1-\delta)$ on $\bar{D}_T$.

(iii) u has weak derivatives u_x, u_t on D_T and for all $t \in (0, T)$, u_x, $u_t \in L^2(0, 1)$.

(iv) For any function $\psi(x, t) \in C^2(\bar{D}_T)$ satisfying the boundary conditions and $0 \leqq t \leqq T$,

$$\int_0^1 \psi(x, t)u_t(x, t)\,dx = \int_0^t \int_0^1 [\psi_\tau(x, \tau)u_\tau(x, \tau) - \psi_x(x, \tau)u_x(x, \tau)]\,dx\,d\tau + \varepsilon \int_0^t \int_0^1 \psi(x, \tau)\varphi(u(x, \tau))\,dx\,d\tau. \tag{2.1}$$

(v) The total energy associated with (W′) is conserved, i.e.,

$$E_T(t) = \frac{1}{2}\int_0^1 (u_x^2 + u_t^2)\,dx - \varepsilon \int_0^1 \int_0^{u(x,t)} \varphi(\eta)\,d\eta\,dx = E_T(0) = 0. \tag{2.2}$$

We next examine the question of the local (in time) existence of the weak solution defined above. The singular value of the nonlinearity and the consequent restriction $|u| \leqq A(1-\delta)$ on D_T prevent straightforward application of Reed [10, Thm. 1, p. 5], because the nonlinearity is now not defined on the domain of d^2/dx^2. Nevertheless, a local existence theorem of the desired kind can be obtained from the contraction mapping principle used for hyperbolic systems, to be found in Garabedian [11, p. 110]. One still has to deal with the strange nonlinearity and the boundary conditions however.

We proceed as follows: Let $\delta \in (0, 1)$ be fixed. Consider the problem (W′) with nonzero initial data

$$\begin{aligned} &u_{tt} = u_{xx} + \varepsilon\varphi(u(x, t)), && 0 < x < 1, \quad T \geqq t > 0, \quad \varepsilon > 0,\\ &u(0, t) = u(1, t) = 0, && T \geqq t > 0,\\ &u(x, 0) = u_0(x),\\ &u_t(x, 0) = v_0(x), \end{aligned} \tag{W″}$$

where u_0, $v_0 \in C^1(0, 1)$ and $u_0(0) = u_0(1) = 0$. Letting $\|\cdot\|_\infty$ denote the sup norm of a function of x, we assume that

$$\|u_0\|_\infty + T\|v_0\|_\infty < A(1-2\delta). \tag{*}$$

Define u, u_0, v_0 by odd periodic (with period two) reflection (in x) on $R^1 \times [0, T]$. Define the following function

$$F : R^1 \times [0, \infty) \times (-A, A) \to R^1,$$

by

$$F(x, t, u) = \begin{cases} \varphi(u), & x \in [2n, 2n+1),\\ \varphi(-u), & x \in [2n-1, 2n), \end{cases} \quad n = 0, \pm 1, \pm 2, \cdots.$$

Then by standard arguments, u solves (W'') if and only if u solves, on $R^1 \times [0, T)$, the integral equation

$$u(x, t) = u_1(x, t) + \frac{\varepsilon}{2} \int_0^t \int_{x-t+\eta}^{x+t-\eta} F(\xi, \eta, u(\xi, \eta))\, d\xi\, d\eta, \tag{2.3}$$

where

$$u_1(x, t) = \frac{1}{2}(u_0(x+t) + u_0(x-t)) + \frac{1}{2} \int_{x-t}^{x+t} v_0(\sigma)\, d\sigma.$$

Clearly, if $(*)$ holds,

$$|||u_1|||_\infty \equiv \sup_{0 \leqq t \leqq T} \|u_1(t)\|_\infty < A(1-2\delta).$$

Let B_τ be the Banach space of odd (in x) continuous functions on $R^1 \times [0, \tau]$, which vanish on the lines $x = n$, n an integer, and are of period two in x. Let $\bar{B}(u_1, A\delta)$ denote the closed ball of radius $A\delta$ in this Banach space. (Note that $u_1 \in B_\tau$.) Define

$$\tilde{T} : \bar{B}(u_1, A\delta) \to B_\tau,$$

by

$$(\tilde{T}u)(x, t) = u_1(x, t) + \frac{\varepsilon}{2} \int_0^t \int_{x-t+\eta}^{x+t-\eta} F(\xi, \eta, u(\xi, \eta))\, d\xi\, d\eta.$$

In view of the definition of F, this map is well defined. It is then easy to check that

$$|||\tilde{T}u - u_1|||_\infty < A\delta, \tag{2.4}$$

$$|||\tilde{T}u - \tilde{T}v|||_\infty < \lambda\, |||u - v|||_\infty, \qquad 0 < \lambda < 1 \tag{2.5}$$

for $u, v \in B(u_1, A\delta)$ provided

$$\tau < \min\left\{ T, \left(\frac{2}{\varepsilon}\right)^{1/2} [\delta A/\varphi((1-\delta)A)]^{1/2}, \left(\frac{2}{\varepsilon}\right)^{1/2} [\varphi'((1-\delta)A)]^{-1/2} \right\},$$

so that $\tilde{T} : B(u_1, A\delta) \to B(u_1, A\delta)$ and is a contraction. Thus $\tilde{T}$ has a unique fixed point. This establishes the following theorem.

THEOREM 2.1. *A weak (C^1) solution of* (W') *exists on D_T if T is sufficiently small, for any $\varepsilon > 0$. The solution is piecewise C^2 in D_T and* (2.1) *and* (2.2) *hold there. Furthermore, if u exists on D_T and $|u| \leqq A(1-\delta)$ on $\bar{D}_T$, then u may be continued to $D_{T+\tau}$ for τ sufficiently small (and positive).*

It can be shown that the solution of (2.3) is regular enough that (2.1) and (2.2) hold when $u_0 \equiv v_0 \equiv 0$. In this case, from (2.3), one easily calculates

$$u_x(x, t) = \frac{\varepsilon}{2}\left(\int_x^{x+t} F(\sigma, x+t-\sigma, u(\sigma, x+t-\sigma))\, d\sigma \right.$$
$$\left. - \int_{x-t}^{x} F(\sigma, \sigma-x+t, u(\sigma, \sigma-x+t))\, d\sigma \right),$$

$$u_t(x, \sigma) = \frac{\varepsilon}{2}\left(\int_x^{x+t} F(\sigma, x+t-\sigma, u(\sigma, x+t-\sigma))\, d\sigma \right.$$
$$\left. + \int_{x-t}^{x} F(\sigma, \sigma-x+t, u(\sigma, \sigma-x+t))\, d\sigma \right),$$

so that, because $F(x, t, u)$ is piecewise continuous, u_x and u_t are continuous everywhere and differentiable in x and t except on the lines $x = n$, $x + t = n$ or $x - t = n$, where n is an integer. In fact, except on this point set,

$$u_{xx}(x, t) = \varepsilon[F(x + t, 0, 0) + F(x - t, 0, 0)] - \varepsilon F(x, t, u(x, t)) + I,$$
$$u_{tt}(x, t) = [F(x + t, 0, 0) + F(x - t, 0, 0)] + I,$$
$$u_{xt}(x, t) = u_{tx}(x, t) = \frac{\varepsilon}{2}[F(x + t, 0, 0) + F(x - t, 0, 0)] + J$$

and piecewise continuous, where

$$I \equiv \frac{\varepsilon}{2}\int_x^{x+t} F_3(\sigma, x + t - \sigma, u(x + t - \sigma))u_2(\sigma, x + t - \sigma)\, d\sigma$$
$$+ \frac{\varepsilon}{2}\int_{x-t}^{x} F_3(\sigma, \sigma - x + t, u(\sigma - x + t))u_2(\sigma, \sigma - x + t)\, d\sigma$$
$$\equiv I_1 + I_2,$$
$$J \equiv I_1 - I_2.$$

Thus the solution is classical except on

$$\{(x, t) \in R^1 \times [0, \tau] \mid x, x - t, x + t \text{ are integers}\}.$$

It is then an easy matter to show, using care when integrating across jumps in u_{xx}, u_{xt}, u_{tt}, that (2.1) and (2.2) hold.

3. Nonexistence or quenching for large ε. The results here are analogous to those in [5]. We repeat them here for completeness and because the class of nonlinearities here is different from that considered in [5].

Throughout this section, $\varphi : (-\infty, A) \to (0, \infty)$ satisfies the following conditions:

(a) $\varphi > 0$, $\varphi' \geqq 0$, φ is convex;

(b) $\lim_{u \to A^-} \varphi(u) = +\infty$.

Let

$$\Phi(x) = \int_0^x \varphi(s)\, ds \tag{3.1}$$

and

$$H(x) = -\pi^2\frac{x^2}{2} + \varepsilon\Phi(x), \qquad -\infty < x < A. \tag{3.2}$$

For H, we suppose that $H(x) > 0$ on $(0, A)$ and $\lim_{x \to A^-} H(x) > 0$.

LEMMA 3.1. *Under the above hypothesis*

$$\infty > \int_0^A [H(\sigma)]^{-1/2}\, d\sigma.$$

This is clear since $H(\sigma) = \varepsilon\varphi(0)\sigma + O(\sigma^2)$ for σ small and positive and bounded away from zero near $\sigma = A$.

THEOREM 3.2. *If $\varepsilon > 0$ is such that the above holds for $H(x)$ and φ satisfies* (a), (b) *above, then a weak solution of* (W′) *must quench in finite time.*

Proof. Assume the contrary: that $|u|<A$ for $(x, t)\in[0, 1]\times[0, \infty)$. Define

$$F(t)\equiv\frac{\pi}{2}\int_0^1 \sin(\pi x)u(x, t)\,dx; \tag{3.3}$$

the choice $\psi(x, t)=t\sin\pi x$ in (2.1) yields

$$\begin{aligned} tF'(t) &= \frac{t\pi}{2}\int_0^\pi \sin(\pi x)u_t(x, t)\,dx \\ &= \frac{\pi}{2}\int_0^t\int_0^1 [\sin(\pi x)u_\eta(x, \eta)-\pi\eta\cos(\pi x)u_x(x, \eta)]\,dx\,d\eta \\ &\quad +\frac{\pi\varepsilon}{2}\int_0^t \eta\int_0^1 \sin(\pi x)\varphi(u(x, \eta))\,dx\,d\eta. \end{aligned} \tag{3.4}$$

Thus $tF'(t)$ is differentiable and hence so is $F'(t)$. Therefore

$$tF''(t)+F'(t)=F'(t)-\frac{t\pi^2}{2}\int_0^1 \cos(\pi x)u_x(x, t)\,dx+t\frac{\pi\varepsilon}{2}\int_0^1 \sin(\pi x)\varphi(u(x, t))\,dx,$$

so that, after integration by parts,

$$F''(t)=-\pi^2F(t)+\frac{\varepsilon\pi}{2}\int_0^1 \sin(\pi x)\varphi(u(x, t))\,dx.$$

The use of Jensen's inequality yields

$$F''(t)\geqq-\pi^2F(t)+\varepsilon\varphi(F(t))\equiv H'(F(t)).$$

Since $F''(0)\geqq\varepsilon\varphi(0)>0$ and $F(0)=F'(0)=0$, we have $F'(t)>0$ and $F(t)>0$ on some interval $(0, \eta)$. Therefore, on this interval

$$\tfrac{1}{2}(F'(t))^2\geqq H(F(t)).$$

From this it follows that F' cannot change sign so that $F(t)\in(0, A)$ for all $t\in[0, \infty)$ and thus

$$\int_0^A [H(\sigma)]^{-1/2}\,d\sigma\geqq\sqrt{2}t$$

for all t which is a contradiction.

Notice that we are invoking part of Theorem 2.1 here, to the effect that if $|u|\leqq A(1-\delta)$ on $[0, 1]\times[0, T]$, then u can be continued as a weak solution on $[0, 1]\times[0, T+t']$ with $|u|\leqq A(1-\delta')(\delta'<\delta)$ and $t'>0$.

Example. For the above problem, we take $\varphi(u)=(1-u)^{-\beta}$, $\beta>0$, $|u|<1$. It is easy to verify that $H(x)>0$ on $(0, 1)$ if either

(i) $\varepsilon\geqq\dfrac{\pi\beta^\beta}{(\beta+1)^{\beta+1}}$

or

(ii) if $\varepsilon<\pi^2\beta^\beta/(\beta+1)^{\beta+1}$ and $(1-\beta)^{-1}\{1-(\pi^2/\varepsilon)x_0+(\pi^2/\varepsilon)((1+\beta)/2)x_0^2\}>0$, where x_0 is the larger root of $x_0(1-x_0)^\beta=\varepsilon/\pi^2$ if $\beta\neq1$ or $\ln(1/(1-x_0))-\pi^2x_0^2/2\varepsilon>0$, where x_0 satisfies $x_0/1-x_0)=\varepsilon/\pi^2$ and is the larger root if $\beta=1$. (This amounts to showing H is positive at a local minimum.)

Since, for $\beta=1$, $x_0(1-x_0)^\beta=\varepsilon/\pi^2$ is readily solved for x_0, we have the following corollary.

COROLLARY 3.3. *If* $\varphi(u)=(1-u)^{-1}$, *then* u *reaches one in finite time if*

$$\varepsilon > \frac{2\pi^2\theta_0}{(1+2\theta_0)^2} = \varepsilon_2,$$

where $e^{\theta_0}=1+2\theta_0$ *and* $x_0=1-e^{-\theta_0}(\theta_0\cong 1.25643$ *and* $L_2=\sqrt{\varepsilon_2}\cong 1.41766)$.

It is of interest to note that the larger β is, the wider is the range of ε's for which quenching in finite time must occur.

4. Global existence. The energy equation (2.2) can be written as

$$\frac{1}{2}\int_0^1 u_x^2\,dx+\frac{1}{2}\int_0^1 u_t^2\,dx-\varepsilon\int_0^1 \Phi(u(x,t))\,dx=0, \tag{4.1}$$

where Φ is given by (3.1). We write, for φ as in § 3,

$$\Phi(u)=\int_0^u \varphi(\eta)\,d\eta=\varphi(0)u+\frac{u^2}{2}\psi(u). \tag{4.2}$$

Here ψ will be singular at $u=A$ if and only if $\varphi\notin L^1(0,A)$.

THEOREM 4.1. *For a weak solution* u *of* (W′) *over* D_T, *if there is* $\delta\in(0,1]$ *such that*

$$\varepsilon < A(1-\delta)\frac{\pi^2}{\pi\varphi(0)+A(1-\delta)M_\delta}\equiv \mathcal{F}(\delta), \tag{4.3}$$

where $M_\delta=\sup_{|u|\leqq A(1-\delta)}\psi(u)$, *where* $0<\delta\leqq 1$, *then*

$$|u(x,t;\varepsilon)|<A(1-\delta) \tag{4.4}$$

for all $(x,t)\in[0,1]\times[0,\infty)$.

Proof. Assume that T is the first number such that

$$\operatorname{Max}\{u(x,t;\varepsilon)\,|\,(x,t)\in[0,1]\times[0,T]\}=A(1-\delta). \tag{4.5}$$

From (4.1), (4.2), Schwarz's and Poincaré's inequalities for $t\in[0,T]$,

$$\begin{aligned}\pi^2\int_0^1 u^2\,dx &\leqq \int_0^1 u_x^2\,dx\\ &\leqq 2\varepsilon\left(\int_0^1 u^2\,dx\right)^{1/2}\left\{\varphi(0)+M_\delta\left(\int_0^1 u^2\,dx\right)^{1/2}\right\}.\end{aligned}$$

From this we obtain the bound

$$\left(\int_0^1 u^2\,dx\right)^{1/2}\leqq 2\varepsilon\varphi(0)(\pi^2-\varepsilon M_\delta)^{-1},$$

the implied denominator on the right-hand side being positive in view of (4.3). If this bound is used in the right hand side of the preceding inequality and if the (sharp) inequality

$$4u^2(x,t)\leqq\int_0^1 u_x^2(x,t)\,dx,$$

is also employed, we obtain the pointwise bound

$$\begin{aligned}u^2(x,t)&\leqq\varepsilon^2\pi^2\varphi^2(0)[\pi^2-\varepsilon M_\delta]^{-2}\\ &\leqq A^2(1-\delta)^2-\delta'\end{aligned} \tag{4.5a}$$

for some $\delta'>0$ (by (4.3)) since the latter inequality is equivalent to (4.3) this contradicts the choice of T in (4.5).

Actually, we have proved a little more. Since ψ is continuous on $(-\infty, A)$, M_δ is monotone decreasing in δ and continuous on $[0, 1]$ so that $\lim_{\delta \to 0^+} \mathcal{F}(\delta)$ exists. Defining $\mathcal{F}(0)$ to be this limit, we see that $\mathcal{F}$ has a (unique) maximum in $[0, 1)(\mathcal{F}(1) = 0)$, and that if the maximum occurs at $\delta_0 \in (0, 1)$, then $\varepsilon < \mathcal{F}(\delta_0)$ implies $|u(x, t; \varepsilon)| < A(1 - \delta_0)$ on the half strip while if $\delta_0 = 0$ and $\varepsilon < \mathcal{F}(0)$, then $\varepsilon \leqq \mathcal{F}(\delta_1)$ for all δ_1 sufficiently close to zero so that $|u(x, t; \varepsilon)| < A(1 - \delta_1)$, again on the half strip. It is also clear that if $\varphi \in L^1(0, A)$, the maximum must occur in $(0, 1)$.

COROLLARY 4.2. *If* $\varepsilon < \max\{\mathcal{F}(\delta): 0 \leqq \delta \leqq 1\}$, *then* $u(x, t; \varepsilon)$ *can never quench.*

Example. If $\varphi(u) = (1 - u)^{-\beta}$, $\beta > 0$, then

$$\psi(u) = 2u^{-2}(1 - \beta)^{-1}[1 - (1 - \beta)u - (1 - u)^{1-\beta}].$$

Since, in $(-\infty, 0)$, $u^2\psi(u)$ is concave and vanishes with its first derivative at $u = 0$, we have $\psi(u) \leqq 0$ in $(-\infty, 0]$. Furthermore, expanding $\psi(u)$ in a Taylor series about $u = 0$, we see that on $[0, 1)$,

$$\psi(u) = 2 \sum_{i=0}^{\infty} \frac{1}{(i+2)!} (\pi_{j=0}^{i+1}(\beta + j))u^i$$

and therefore $M_\delta = \psi(1 - \delta)$.

For the case $\beta = 1$, $\psi(u) = 2u^{-2}[\ln(1/(1-u)) - u]$. We find, by direct computation, that $\delta_0 = 0.22684$ and that $u \leqq 1 - \delta_0$ if $\varepsilon < \mathcal{F}(\delta_0) = (1.2379)^2$. Thus, combining this with the results of the example from the last section, we see that, with reference to (W): If $L > 1.41766 \equiv L_2$, then u quenches in finite time, while if $L < 1.2379 \equiv L_1$, then $u \leqq 0.7732$ for all time.

It is fair to ask whether one could improve the arguments involved in Theorem 4.1 by writing

$$\Phi(u) = \sum_{0}^{N-1} \varphi^{(k-1)}(0) \frac{u^k}{k!} + \left(\frac{1}{N!}\right) u^N \psi_N(u),$$

and employing Holder's inequality on the first $N - 1$ terms and the (Sobolev) inequality

$$C^2(N)\left(\int_0^1 |u|^N \, dx\right)^{2/N} \leqq \int_0^1 u_x^2 \, dx, \tag{4.6}$$

in place of Poincaré's inequality to obtain a bound on the L^N-norm of u. The constant $C(N)$ is known[1] and the inequality (4.6) is best possible. We did this for $\varphi(u) = 1/(1 - u)$ and for various N. However the case $N = 2$ seems to yield the best value of L_2.

The numerical results indicate that if $L > 1.365 \cdots$, then u quenches in finite time while if $L < 1.365 \cdots$, u does not quench, even in infinite time. More precisely, what is observed numerically is the following. The solution, for small ε, has a discrete sequence of local maxima located along the line $x = \frac{1}{2}$, $t > 0$. The first of these local maxima appears to be an absolute maximum, which, as L increases to $1.365 \cdots$, from below, approaches one from below. However, the time to reach this maximum value increases without bound as L increases to $1.365 \cdots$. Moreover, if $L > 1.365 \cdots$, this maximum value is one and as L decreases to $1.365 \cdots$, the time taken to reach one increases without bound. This, if $L = 1.365 \cdots$, then u quenches in infinite time.

[1] $C(p) = (2\pi p)^{1/2}(2/(2+p))^{(p-2)/2p}\Gamma(1 + 1/p)/\Gamma(2 + 1/p)$, $1 \leqq p < \infty$. ($C(2) = \pi$, $C(\infty) = 2$.)

5. Perturbation analysis of (W'). In this section we investigate the behavior of solutions of (W') as $\varepsilon \to 0$. The arguments involved are standard and will only be sketched. We write

$$\varphi(u) = \varphi(0) + \varphi'(0)u + u^2\psi_1(u)$$

and assume that $u \leqq 1-\delta$ implies $|\psi_1(u)| \leqq M(\delta)$. We take $A = 1$ here for convenience. The linear problem

$$\text{(L)} \qquad \begin{aligned} v_{tt} - v_{xx} &= \varepsilon\varphi(0) + \varepsilon\varphi'(0)v, \\ v(0,t) &= v(1,t) = 0, \\ v(x,0) &= v_t(x,0) = 0, \end{aligned}$$

is easily solved by elementary methods and found to have the following properties if $\varepsilon < \pi^2/\varphi'(0) \equiv \varepsilon_1$,

(L-1) $$v(x,t;\varepsilon) = \varepsilon v_1(x,t;\varepsilon);$$

(L-2) $|v_1(x,t;\varepsilon)| \leqq M_1$ where M_1 is an absolute constant independent of x, t, ε for $\varepsilon \leqq \varepsilon_1(1-\sigma)$, $\sigma \in [0,1]$;

(L-3) $\lim_{\varepsilon\to 0^+}(x,t;\varepsilon) = v_1(x,t;0) (\not\equiv 0)$, which solves

$$v_{1tt} - v_{1xx} = \varphi(0).$$

The following result then holds.

THEOREM 5.1. *Let u solve* (W") *on* $[0,1]\times[0,\infty)$ *and suppose* $|u(x,t;\varepsilon)| \leqq 1-\delta$ *for all* $\varepsilon < \varepsilon'$ *say, on the half strip. Then*

$$u(x,t;\varepsilon) = v(x,t;\varepsilon) + w(x,t;\varepsilon),$$

where

$$\lim_{\varepsilon\to 0^+} \varepsilon^{-\sigma} w(x,t;\varepsilon) = 0$$

for every σ, $0 \leqq \sigma < 2$, *convergence being uniform on compact subsets of* $[0,1]\times[0,\infty)$.

Proof. The difference $w = u - v$ satisfies (weakly) the equation

$$w_{tt} - w_{xx} = \varepsilon\varphi'(0)w + \varepsilon u^2\psi_1(u),$$

with the same initial and boundary data as u and v. The following energy principle then holds for w

$$\begin{aligned} E(t) &\equiv \frac{1}{2}\int_0^1 w_t^2\,dx + \frac{1}{2}\int_0^1 w_x^2\,dx \\ &= \tfrac{1}{2}\varepsilon\varphi'(0)\int_0^1 w^2\,dx + \varepsilon\int_0^t\int_0^1 u^2\psi_1(u)w_\eta\,dx\,d\eta. \end{aligned}$$

If we write

$$u^2\psi_1(u)w_\eta = \varepsilon v_1 u\psi(u)w_\eta + u\psi_1(u)ww_\eta,$$

choosing ε so small that $\frac{1}{2}\varepsilon\varphi'(0) < \pi^{-2}/2$, we find that

$$E(t) \leqq \varepsilon^2 A\int_0^t\int_0^1 |w_\eta|\,dx\,d\eta + B\varepsilon\int_0^t\int_0^1 |ww_\eta|\,dx\,d\eta, \tag{5.1}$$

where A, B are computable constants depending only on δ, $M(\delta)$, M_1.[2] For (not necessarily the same) A, B, we have further that

$$E(t) \leqq \varepsilon A \int_0^t E(\eta)\, d\eta + \varepsilon^3 Bt, \tag{5.2}$$

where we have used the assumption that $|\psi(u)| \leqq M(\delta)$ if $u \leqq 1-\delta$ and

$$\varepsilon^3 \int_0^1 |w_\eta|\, dx \leqq \varepsilon^3 \left(\int_0^1 |w_\eta|^2\, dx \right)^{1/2} \leqq \tfrac{1}{2}\varepsilon^3 + \tfrac{1}{2}\varepsilon \int_0^1 |w_\eta|^2\, dx$$

and where the arithmetic-geometric mean inequality and Poincaré's inequality have been used in the second term in (5.1).

Gronwall's inequality applied to (5.2) yields

$$E(t) \leqq A_1(t)\varepsilon^2 + A_2(t)\varepsilon^3,$$

where $A_1(t)$ and $A_2(t)$ are uniformly bounded on $[0, T]$ for $\varepsilon \in [0, \varepsilon_0]$ say. Putting this back into (5.2) yields (for different A_1, A_2 with the aforementioned properties)

$$E(t) \leqq A_1(t)\varepsilon^3 + A_2(t)\varepsilon^4 = O(\varepsilon^3). \tag{5.3}$$

This can be improved, at the expense of worse order constants, by using the following consequence of (5.1),

$$E(t) \leqq A\varepsilon^2 \int_0^t \sqrt{E(\eta)}\, d\eta + B\varepsilon \int_0^t E(\eta)\, d\eta. \tag{5.4}$$

Use of (5.3) in (5.4) yields

$$E(t) = O(\varepsilon^{7/2}).$$

Using this in (5.4) once again, we find $E(t) = O(\varepsilon^{15/4})$ etc. Thus, on every compact subset of $[0, 1] \times [0, \infty)$ and for every $\delta' > 0$,

$$E(t) = O(\varepsilon^{4-\delta'}) \quad \text{as } \varepsilon \to 0.$$

From the inequality

$$w^2(x, t) \leqq \frac{1}{4} \int_0^1 w_x^2\, dx \leqq \tfrac{1}{2} E(t),$$

we see that on every compact subset of $[0, 1] \times [0, \infty)$,

$$\lim_{\varepsilon \to 0^+} \varepsilon^{2-\delta'} w(x, t, \varepsilon) = 0,$$

uniformly.

In particular, $w/\varepsilon \to 0$ uniformly on compact subsets of the half strip. Thus, for small ε, $v_1 = v/\varepsilon$ will make the dominant contribution to u/ε. This function is, for $\varphi(u) = 1/(1-u)$,

$$v_1(x, t, \varepsilon) = \frac{4}{\pi} \sum_1^\infty \frac{[1 - \cos((2n+1)^2\pi^2 - \varepsilon)^{1/2} t]}{(2n+1)[(2n+1)^2\pi^2 - \varepsilon]} \sin((2n+1)\pi x).$$

[2] Here $M(\delta) = \sup\{|\psi_1(u)|, -\infty < u \leqq 1-\delta\}$. Actually this supremum need only be taken over $[-(1-\delta), 1-\delta]$ in view of (4.5a).

This fact has been observed numerically. That is, we observed numerically that for small ε, u was not only bounded away from one but also was oscillatory. This seemed surprising in view of our (mistaken) belief that since u quenched in finite time for L's less than L_0, u would quench in finite time for all $L > 0$.

Finally, we make the following observations: We can extend the results of the paper to the case of nonzero, appropriately restricted initial data. Furthermore it is possible to obtain analogous results in higher dimensions, at least in so far as §§ 3 and 4 are concerned since only Poincaré's inequality and the positivity of the first eigenfunction for the membrane problem are used. Preliminary calculations indicate that results along the lines of this paper and those in [1], [6] may be possible if the nonlinearity appears in the boundary condition. The second author is currently investigating this possibility.

Acknowledgment. The authors thank the referees for several helpful comments and suggestions.

REFERENCES

[1] A. ACKER AND W. WALTER, *On the global existence of solutions of parabolic differential equations with a singular nonlinear term*, Nonlinear Anal., 2 (1978), pp. 499–505.

[2] A. DOUGLIS, *Existence theorems for hyperbolic systems*, Comm. Pure Appl. Math., 5 (1952), pp. 119–154.

[3] H. KAWARADA, *On the solutions of initial boundary value problem for* $u_t = u_{xx} + 1/(1-u)$, Pub. RIMS Kyoto Univ., 10 (1975), pp. 729–736.

[4] P. D. LAX, *Nonlinear hyperbolic equations*, Comm. Pure Appl. Math., 6 (1953), pp. 231–258.

[5] H. A. LEVINE, *On the nonexistence of global weak solutions of some properly and improperly posed problems of mathematical physics: the method of unbounded Fourier coefficients*, Math. Ann., 214 (1975), pp. 205–220.

[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] J. RAUCH, *Singularities of solutions to semilinear wave equations*, J. Math. Pures Appl., 58 (1979), pp. 299–308.

[8] 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.

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

[10] M. REED, *Abstract Nonlinear Wave Equations*, Lecture Notes on Mathematics 507, Springer Verlag, New York, 1976.

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

[12] J. RAUCH AND M. REED, *Propagation of singularities for semilinear hyperbolic equations in one space variable*, Ann. Math., 111 (1980), pp. 531–552.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0011 $01.00/0

ASYMPTOTIC BEHAVIOR IN AGE-DEPENDENT POPULATION DYNAMICS WITH HEREDITARY RENEWAL LAW*

PIERANGELO MARCATI†

Abstract. This paper is concerned with the age dependent population dynamics where the renewal law involves hereditary effects. The first section is devoted to a model of Von Foerster type (i.e., a single species linear model). The second section is devoted to the same model with spatial spread. Both models are studied by means of the Laplace transform applied to a convolution integral equation (renewal equation). Actually in the model with diffusion we convert the equation to an abstract differential equation on a Banach space. The semigroup approach combined with a spectral mapping argument allow us to use the Laplace transform also in this case. An asymptotic expansion as in the classical Von Foerster model is obtained.

Introduction. In this paper we shall study a model for the age dependent population dynamics of a single species (e.g., a cell population, a bird population or the female human population). These models are studied using integral equations theory. Indeed the model's equation can be reduced to an equivalent integral equation of convolution type, the so called "renewal equation". For this equation we refer to the papers of Feller [7], Paley–Wiener [14], Coale [2] and the books of Hoppensteadt [11] and Bellmann–Cooke [1]. In the classical theory (see [11]), if we denote by $u(a, t)$ the density of population at age a and time t, the "birth law" is assumed to verify the following equation:

$$u(0, t) = \int_0^A b(a)u(a, t)\, da,$$

where $b(a) \geqq 0$ is the age specific birth rate. If we consider, for instance, a bird population, this law does not hold because we have to take into account the maturation period of the eggs.

In this case we consider a function $g(s) \geqq 0$ which denotes the proportion of the eggs that will yield living birds s units of time after the eggs are laid. Therefore, the "birth law" can be assumed to be of the following "hereditary" type:

$$u(0, t) = \int_{t-r}^t g(t-s) \int_0^A b(a)u(a, s)\, da\, ds,$$

where $r > 0$ is the maximum maturation period for the eggs. This condition has been considered by Cushing [3], when $r = +\infty$.

Let us denote by $m(a) > 0$ the age specific death rate; then in a way similar to [11], we get

$$\text{(P)} \qquad \begin{aligned} &\frac{\partial u}{\partial a} + \frac{\partial u}{\partial t} = -m(a)u(a, t), \qquad (a, t) \in [0, A] \times \mathbb{R}_+, \\ &u(a, \theta) = \phi(a, \theta), \qquad (a, \theta) \in [0, A] \times [-r, 0] \\ &u(0, t) = \int_{t-r}^t g(t-s) \int_0^A b(a)u(a, s)\, da\, ds, \qquad t \in \mathbb{R}_+. \end{aligned}$$

In order to consider spatial spread, let $u(a, t, x)$ denote the density of population per unit age, unit time and unit surface, where x belongs to an open bounded subset

* Received by the editors November 7, 1980.

† Dipartimento di Matematica, Libera Università degli Studi di Trento 38050 POVO (TN)-Italy. This work was supported by CNR Grant No. 79-00696-01.

Ω of $\mathbb{R}^2$. Thus, the birth law is given

$$u(0, t, x) = \int_{t-r}^{t} g(t-s) \int_0^A b(a)u(a, s, x)\, da\, ds.$$

In a way similar to the case without delay (see Marcati–Serafini [13]), we get

$$\frac{\partial u}{\partial a} + \frac{\partial u}{\partial t} = -m(a)u(a, t, x) + \operatorname{div}_x [G(x) \operatorname{grad} u(\partial, t, x)],$$
$$(a, t, x) \in [0, A] \times \mathbb{R}_+ \times \Omega,$$

(P′)
$$u(a, \theta, x) = \phi(a, \theta, x), \qquad (a, \theta, x) \in [0, A] \times [-r, 0] \times \bar{\Omega},$$
$$u(0, t, x) = \int_{t-r}^{t} g(t-s) \int_0^A b(a)u(a, s, x)\, da\, ds, \qquad (t, x) \in \mathbb{R}_+ \times \bar{\Omega},$$
$$\frac{\partial u}{\partial \nu} = 0 \quad \text{or} \quad u = 0 \quad \text{on } \partial\Omega, \qquad C(x) \geqq c_0 > 0.$$

(ν is the outward unit normal).

The problem of spatial spread in age dependent population dynamics was proposed by Gurtin [9] and investigated by Gopalsamy [8], Di Blasio–Lamberti [4] and Marcati–Serafini [13]. Gurtin [9] proposed a more general diffusion operator of the type

$$-\left[\operatorname{div}_x \int_0^A K(a, a', x) \operatorname{grad}_x u(a', t, x)\, da'\right]$$

that has been studied in [4]. Some other diffusion models can be found in Webb [17] for epidemics.

Our purpose here, is to study the asymptotic behavior of (P) and (P′) in order to obtain an asymptotic expansion for the solution (as in the classical case). The methods employed here are Laplace transform theory, for the problem (P), and the Laplace transform combined with semigroup theory and spectral analysis for (P′).

Remark. This model can be applied also to a female population. In this case $g(s)$ is concentrated in a narrow interval near r. Let us denote, by $b(a)u(a, s)\, da$, the number of females with ages between a and $a + da$ who conceived at time s. Then

$$\exp\left[-\int_a^{a+(t-s)} m(\sigma)\, d\sigma\right] \cdot b(a)u(a, s)\, da$$

is the number of females who survive till the time t. The number of offsprings will be given by

$$u(0, t) = \int_{t-r}^{t} g_0(t-s) \int_0^A \exp\left[-\int_a^{a+(t-s)} m\, d\sigma\right] b(a)u(a, s)\, da\, ds.$$

Actually, the function m in the age interval where the females can be conceived is nearly a constant value $\underline{m}$. Then we can assume

$$u(0, t) = \int_{t-r}^{t} g_0(t-s) \exp[-\underline{m}(t-s)] \int_0^A b(a)u(a, s)\, da\, ds$$

holds also for the female population; that is,

$$g(s) = g_0(s) \exp[-\underline{m}s].$$

1. Asymptotic behavior for (P). Assume that the following hypotheses hold;

$$
\begin{aligned}
&b \in C[0, A], \qquad b(0) = b(A) = 0, \quad b \geqq 0,\\
&m \in C[0, A], \qquad m > 0, \qquad \int_0^A m(a)\, da = +\infty,\\
&\phi \in C([0, A]x[-r, 0]), \qquad \phi \geqq 0,\\
&\phi(0, 0) = \int_{-r}^{0} g(-s) \int_0^A b(a)\phi(a, s)\, da\, ds,\\
&g \in C[0, r], \qquad g \geqq 0, \qquad \|g\|_{L^1} = 1.
\end{aligned}
\tag{1.1}
$$

Let

$$
p(t) = \exp\left[-\int_0^t m(a)\, da\right],
$$

$$
V(t) = b(t)p(t) \quad \text{if } t \in [0, A], \qquad V(t) = 0 \quad \text{outside } [0, A], \tag{1.2}
$$

$$
K(t) = \int_{t-r}^{t} g(t-s)V(s)\, ds \quad \text{if } t \in [0, A+r], \qquad K(t) = 0 \text{ outside } [0, A+r] \tag{1.3}
$$

and

$$
\begin{aligned}
f(t) = &\int_{(t-r)\vee 0}^{t} g(t-s) \int_s^A \{p(a)/p(a-s)\}\phi(a-s, 0)\, da\, ds\\
&+ \int_{(t-r)\wedge 0}^{0} g(t-s) \int_0^A b(a)\phi(a, s)\, da\, ds,
\end{aligned}
\tag{1.4}
$$

(where $\vee$ and $\wedge$ are the sup and inf symbols).

Solving (P), along the characteristics $t - a = \text{const.}$, we obtain

$$
u(a, t) = \begin{cases} p(a)B(t-a) & \text{if } t > a,\\ [p(a)/p(a-t)]\phi(a-t, 0) & \text{if } t \leqq a, \end{cases} \tag{1.5}
$$

where $B(t) = u(0, t)$.

If we insert the above expression for u in the "renewal law", we get

$$
B(t) = f(t) + \int_0^t K(t-s)B(s)\, ds. \tag{1.6}
$$

The equation (1.6) is called the "renewal equation".

In order to have a solution $u \in C^1$ to (P), we need a solution $B \in C^1$. So we have to make stronger assumptions on b, g, m. But the considerations that we shall examine about the existence and the asymptotic behavior of a solution to (1.6), do not require any differentiability of the data.

THEOREM 1.1. *The following conclusions hold*:

(i) *Assuming only condition* (1.1), *we have the existence and the uniqueness of a positive solution* $B \in C(\mathbb{R}_+)$ *to* (1.6). *Then, by* (1.5), *the problem* (P) *has a unique continuous positive solution in the sense that for all* $(a, t) \in [0, A] \times \mathbb{R}_+$ *one has the existence of the directional derivative*

$$
D_{1,1}(a, t) = \lim_{h \to 0} \frac{u(a+h, t+h) - u(a, t)}{h}
$$

and

$$D_{1,1}(a, t) = -m(a)u(a, t).$$

(ii) *Moreover, if* g, b, ϕ *are* C^1, *then* u *is* C^1 *except the characteristic line* $t = a$.

Proof. The first part of (i) is a classical result concerning linear Volterra integral equations (see, for instance, [1]). The second part follows by solving (P) along the characteristics $t - a = \text{const}$ and calculating $D_{1,1}$ by (1.5). To prove (ii), we remark that p is differentiable in $[0, A]$, and $p'(a) = -m(a)p(a)$.

Then p' is in $L^1[0, A] \cap C[0, A)$, so f and B are differentiable in $(0, +\infty)$. In addition, for all $t > 0$,

$$B'(t) = f'(t) + \int_0^t K'(t-s)B(s)\, ds.$$

Since f, in general, is not differentiable in $t = 0$, B is not differentiable along $t - a = 0$. The same holds for u. □

The following result is concerned with the asymptotic expansion for the solution to (1.6). The proof strictly follows the classical theory (see, for instance, [11]).

We shall denote by the superscript $\hat{}$ the Laplace transform of a given function.

THEOREM 1.2. *If the hypotheses in* (1.1) *are fulfilled, we have the following asymptotic expansion for* $B(t)$:

$$B(t) = B_0 \exp(p_r^* t) + o(\exp(p_r^* t)) \quad \text{as } t \to \infty$$

where $B_0 \geqq 0$ *and*

$$\hat{K}(p_r) = 1, \qquad p_r^* = \max\{\operatorname{Re} p : \hat{K}(p) = 1\}.$$

Proof. Since f, K have compact support their Laplace transforms are analytic entire functions. By Gronwall's lemma applied to equation (1.6) there exists $\sigma \in \mathbb{R}$ such that

$$|B(t)| \leqq \text{const. } e^{\sigma t}.$$

So, $\hat{B}(p)$ exists for $\operatorname{Re} p > \sigma$. By (1.6), one has

$$\hat{B}(p) = (1 - \hat{K}(p))^{-1} \hat{f}(p), \qquad \operatorname{Re} p > \sigma.$$

The function $\hat{B}(p)$ has a meromorphic continuation in the half-plane $\operatorname{Re} p \leqq \sigma$, and its poles are given by the roots of the characteristic equation $\hat{K}(p) = 1$. This equation has a unique real simple root p_r^*, such that for all $p \neq p_r^*$ and $\hat{K}(p) = 1$, it follows that $\operatorname{Re} p < p_r$. In addition, the set of the roots of $\hat{K}(p) = 1$ is a countable subset of $\mathbb{C}$. In each vertical strip $|\operatorname{Re} p| \leqq \text{const.}$, there are a finite number of roots (see [11]). Now if we choose $\gamma \in \mathbb{R}$ such that $p_r > \gamma$, and $\operatorname{Re} p < \gamma$ if $p \neq p_r$, and $\hat{K}(p) = 1$, then

$$B(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} (1 - \hat{K}(p))^{-1} f(p)\, e^{pt}\, dp + \operatorname*{Res}_{p = p_r^*} [e^{pt} \hat{B}(p)]$$

(the convergence of the integrals follows by the Riemann–Lebesgue lemma (see [6]), since f, K have compact support). Therefore one has

$$B(t) = B_0\, e^{p_r^* t} + o(e^{p_r^* t}) \quad \text{as } t \to \infty. \quad \square$$

In the next result we shall make a comparison of the solution to (1.6) with the qualitative behavior of the problem without delay, where $r = 0$ and g is equal to the Dirac distribution concentrated in 0.

THEOREM 1.3. *Denote by p^* the real root of $\hat{V}(p)=1$, that is the characteristic root of the problem without delay; then the following conclusions hold. For all $r>0$ it follows that:*

i) *If $p^*<0$ then $p^*<p_r^*<0$.*

ii) *If $p^*=0$ then $p_r^*=0$.*

iii) *If $p^*>0$ then $0<p_r^*<p^*$.*

Then $\int_0^A V(t)\,dt<1$ is necessary and sufficient for $p_r^*<0$, while $\int_0^A V(t)\,dt>1$ is necessary and sufficient for $p_r^*>0$.

Proof. It is sufficient to point out that the functions

$$p\in\mathbb{R}\to\hat{V}(p)\in\mathbb{R},\qquad p\in\mathbb{R}\to\hat{K}(p)\in\mathbb{R}$$

are monotonic decreasing. Then, if $p^*<0$, it follows that

$$\hat{K}(p_r^*)=1=\hat{V}(p^*)>\hat{g}(p^*)\hat{V}(p^*)=\hat{K}(p^*)$$

so $\hat{K}(p_r^*)>\hat{K}(p^*)$ and hence $p^*<p_r^*$.

Moreover, $1>\hat{V}(0)>\hat{K}(0)$ and $\hat{K}(p_r^*)>\hat{K}(0)$, so $p_r^*<0$, and (i) is proven. The proof of (ii) is obvious. In the same way as for (i) one can get (iii). □

Remark 1.4. The above results are also true if $g(s)\,ds=d\eta$, where $\eta:[0,r]\to\mathbb{R}$ is a nondecreasing function such that $\eta(r)=1+\eta(0)$.

Remark 1.5. An important consequence of the above theorem is that the behavior at infinity of the solution to equation (1.6) is qualitatively determined by the behavior at infinity of the solution to the equation of the model without delay,

$$D(t)=F(t)+\int_0^t V(t-s)D(s)\,ds,$$

where

$$F(t)=\int_t^A\{p(a)/p(a-t)\}\phi(a-t,0)\,da.$$

In particular, the gestation delay does not yield oscillations or bifurcations, while these phenomena can be found in many nonlinear models (see [3]).

2. Investigation of the model (P′).

A. Existence and uniqueness. We convert the model (P′) to an abstract differential equation on a suitable Banach space.

Let L be the infinitesimal generator of a strongly continuous semigroup $\{T(t): t\geqq 0\}$ on the Banach space X. Let us denote by $D(L)$ the domain of L (that is densely embedded in X) and, by $|\cdot|$, the norm on X. Let $K\subset X$ represent a closed convex cone with zero vertex. Assume K is invariant under the action of the semigroup, that is,

$$T(t)K\subseteq K,\qquad t>0. \tag{2.1}$$

The equation (P′) can be put into the following abstract setting:

$$\text{(A)}\qquad\begin{aligned}&\frac{\partial u}{\partial a}+\frac{\partial u}{\partial t}=-m(a)u+Lu,\\&u(0,t)=\int_{t-r}^t g(t-s)\int_0^A b(a)u(a,s)\,da\,ds,\\&u(a,\theta)=\phi(a,\theta),\\&a\in[0,A],\qquad t\in\mathbb{R}_+,\qquad \theta\in[-r,0],\end{aligned}$$

where the following hypotheses are fulfilled:

(2.2) g, m, b verify (1.1),

(2.3) $\phi \in C([0, A]\times[-r, 0]; K)$, $\int_{-r}^{0} g(-s) \int_{0}^{A} b(a)\phi(a, s)\, da\, ds = \phi(0, 0)$.

To obtain (P′) from (A), we assume

(2.4) $Lu = \operatorname{div}_x [C(x) \operatorname{grad}_x u]$, $C \in C^{1,\alpha}(\bar{\Omega})$, $C(x) \geqq C_0 > 0$, $\alpha \in [0, 1)$.

Then let us consider $\partial\Omega$ sufficiently smooth (e.g., $\partial\Omega \in C^2$),

$$X = C_0(\bar{\Omega}) = \{u \in C(\bar{\Omega}): u = 0 \text{ on } \partial\Omega\},$$

$$D(L) = \{u \in C^1(\bar{\Omega}) \cap C_0(\bar{\Omega}): Lu \in C_0(\bar{\Omega})\},$$

if we require Dirichlet boundary conditions; otherwise,

$$X = C(\bar{\Omega}), \qquad D(L) = \left\{u \in C^1(\bar{\Omega}): Lu \in X, \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega\right\}$$

when Neumann boundary conditions occur. L^p spaces can also be used. In all these cases, we assume $K = \{u \in X: u \geqq 0\}$. For the semigroup properties we refer to Kato [12] and Stewart [15], [16].

We introduce here the abstract form of the "renewal equation"

(2.5) $$B(t) = f(t) + \int_0^t K(t-s)B(s)\, ds, \qquad B(t) = u(0, t),$$

where[1]

$$K(t) = \int_{(t-r)\vee 0}^{t} g(t-s)V(s)T(s)\, ds \in \mathcal{L}(X),$$

$$f(t) = \int_{(t-r)\vee 0}^{t} g(t-s) \int_{s}^{A} b(a)\{p(a)/p(a-s)\}T(s)\phi(a-s, 0)\, da\, ds + \int_{t-r}^{0} g(t-s) \int_{0}^{A} b(a)\phi(a, s)\, da\, ds.$$

In a way similar to [13], it is possible to prove the following existence and uniqueness theorem.

THEOREM 2.1. *The following propositions hold*:

i) *Assume that the hypotheses* (2.2), (2.3) *are fulfilled*; *then there exists a unique solution* $B \in C(\mathbb{R}_+, K)$ *to the "renewal equation"* (2.5).

ii) *Moreover, if*

$$\phi(a, \theta) \in D(L) \cap K \quad \textit{for all } (a, \theta) \in [0, A]\times[-r, 0],$$

$$(a, \theta) \in [0, A]\times[-r, 0] \to L\phi(a, \theta) \in X$$

is continuous, then

(2.6) $$B(t) \in D(L) \cap K \quad \textit{for all } t \geqq 0$$

[1] $\mathcal{L}(X)$ denotes the Banach algebra of bounded linear operators on X, endowed with the norm $\|U\| = \sup\{|Ux|: |x| = 1, x \in X\}$.

and $t \to LB(t) \in X$ *is continuous. So* (A) *has a unique solution given by the expression*

$$u(a,t) = \begin{cases} (p(a)/p(a-t))T(t)\phi(a-t,0) & \text{if } t \leqq a, \\ p(a)T(a)B(t-a) & \text{if } t > a, \end{cases} \tag{2.7}$$

in the sense that the directional derivative $D_{1,1}$ *(defined as in Theorem* (1.1)*) exists, is continuous, and* $D_{1,1}(a,t) = -m(a)u(a,t) + Lu(a,t)$.

Proof. By standard contraction mapping arguments, one has a unique solution $B \in C(\mathbb{R}_+, X)$. Moreover, this solution can be approximated by the Picard iteration scheme

$$B_0(t) = f(t), \qquad B_{n+1}(t) = f(t) + \int_0^t K(t-s)B_n(s)\,ds.$$

Since $f(t)$ is in K for all $t \geqq 0$, then by (2.1) one has $B_1(t) \in K$, for all $t \geqq 0$. For the same reasons, if $B_n(t) \in K$ then $B_{n+1}(t) \in K$. Therefore, $B \in C(\mathbb{R}_+, K)$. Now we go on to prove (ii). Now, let us define the so-called Yosida approximations $L_n = nL(n-L)^{-1}$, for sufficiently large n. Then one has

$$L_nB(t) = L_nf(t) + \int_0^t K(t-s)L_nB(s)\,ds \tag{2.8}$$

since L_n commutes with $T(t)$. By Gronwall's lemma, it follows that

$$|L_nB(t) - L_mB(t)| \leqq \text{const. } e^{\delta t}|L_nf(t) - L_mf(t)|,$$

for a suitable $\delta > 0$. By the above assumptions (ii), we have that for all $t \geqq 0$, $f(t)$ is in $D(L)$. Then, by well-known properties of Yosida approximations, $L_nf(t)$ converges to $Lf(t)$. In this way, $L_nB(t)$ is a Cauchy sequence in X and $B(t) \in D(L)$. Finally, passing through the limit for $n \to \infty$ in (2.8), one has $LB(t)$ is a solution of

$$LB(t) = Lf(t) + \int_0^t K(t-s)LB(s)\,ds.$$

Solving (A) along the characteristics $t - a = \text{const}$, we get (2.7). Since $Lf(t)$ is continuous in t, $LB(t)$ is also continuous in t. In this way, it follows from (2.7) that

$$u(a,t) \in D(L) \quad \text{for all } (a,t) \in [0,A] \times \mathbb{R}_+$$

and $(a,t) \to Lu(a,t)$ is continuous. Then (2.7) is a solution to (A). It is unique, since if there are two different solutions u, v to equation (A), then $u(0,t)$ and $v(0,t)$ would be two different solutions to (2.5), which is not true. □

B. Asymptotic behavior. The aim of this section is to find an asymptotic expression for the solution to the "renewal equation" (2.5), and then by (2.7) to the solution of (A).

The main result of this section is seen in the following theorem.

THEOREM 2.2. *Assume that L verifies the following assumptions*:

(H_1) $T(t)$ *is an analytic semigroup of bounded linear operator on X.*

(H_2) *There exists* λ *in the resolvent set of L such that* $(\lambda - L)^{-1}$ *is completely continuous.*

(H_3) *There exists a unique eigenvalue* λ_0 *of L such that*[2]

$$\lambda_0 = \max\{\operatorname{Re}\lambda : \lambda \in \sigma(L)\}.$$

[2] $\sigma(L)$ is the spectrum of L and $\sigma_p(L)$ is the point spectrum of L.

Let us denote by

$$G_0(t) = \int_{t-r}^{t} g(t-s)V(s)\,e^{\lambda_0 s}\,ds,$$

$$p_r^{(0)} = \max\{\operatorname{Re} p\colon \hat{G}_0(p) = 1\}.$$

Then the solution B to (2.5) *has the following asymptotic expression*:

$$B(t) = B_0(t)\exp[p_r^{(0)}t] + o(\exp[p_r^{(0)}t]) \quad \text{as } t \to \infty,$$

where $B_0(t)$ is a polynomial in t with coefficients in K.

COROLLARY 2.3 (Marcati–Serafini [13]). *If $g = \delta_0$ and $r = 0$ and if*

$$z_0 = \max\{\operatorname{Re} z\colon \hat{V}(z) = 1\},$$

then $p_r^{(0)} = z_0 + \lambda_0$.

The proof of this proposition will be given at the end of the paper. Our first remark is concerned with some consequences of the hypotheses (H_1), (H_2), (H_3) given above.

Remark 2.4. With the above hypotheses (H_1), (H_2), $T(t)$ is a completely continuous semigroup, and $\sigma(L) = \sigma_p(L) = \{\lambda_n : n \in \mathbb{N}\}$.

By (H_1), one has the existence of $\delta > 0$ such that $0 < \delta < \pi/2$ and

$$\delta + \pi/2 < \operatorname{Arg} \lambda_n < (3/2)\pi - \delta$$

(see, for instance, Kato [12]).

In order to study the asymptotic behavior of (2.5), our goal is to apply here the Laplace transform approach as made in the first part of the paper. Since f, g, K have compact support, their Laplace transforms $\hat{f}$, $\hat{g}$, $\hat{K}$ are analytic entire functions (actually the analyticity of $\hat{K}(p)$ is proven in the strong operator topology, but by the uniform boundedness theorem, one has that $\hat{K}(p)$ is analytic in the uniform operator topology).

By applying Gronwall's lemma to (2.5), there exists $\beta > 0$ such that $|B(t)| \leqq$ const. $e^{\beta t}$ as $t \geqq 0$. Actually, β can be chosen so large that $\|\hat{K}(p)\| < 1$. Then, by (2.5), one has

$$\hat{B}(p) = (1 - \hat{K}(p))^{-1}\hat{f}(p), \qquad \operatorname{Re} p > \beta.$$

We wish to extend $\hat{B}$ to the entire complex plane except a set of "singular points". More precisely:

DEFINITION 2.1. We say $p \in \mathbb{C}$ is a singular point for $(1 - \hat{K}(p))^{-1}$ if $1 \in \sigma(\hat{K}(p))$. Let us denote by $\mathcal{S}$ the set of all singular points.

LEMMA 2.5. *The operator $\hat{K}(p)$ is a completely continuous linear map on X. Therefore $p \in \mathcal{S}$ if and only if $1 \in \sigma_p(\hat{K}(p))$.*

Proof. Let

$$\hat{H}(p) = \int_0^A e^{-pt}V(t)T(t)\,dt;$$

then $\hat{K}(p) = \hat{g}(p)\hat{H}(p)$. Therefore, $\hat{K}(p)$ is completely continuous if $\hat{H}(p)$ is too. First, assume that V is C^1 and p is in the resolvent set of L. We have

$$\hat{H}(p) = (p - L)^{-1}\int_0^A e^{-pt}T(t)V'(t)\,dt.$$

Since $(p - L)^{-1}$ is compact, then for all p in the resolvent it follows that $\hat{K}(p)$ is completely continuous. However, the resolvent set is dense in $\mathbb{C}$, so we can approximate, in the uniform operator topology, a generic $\hat{H}(p)$ by a sequence $\{\hat{H}(p_n)\}$, where

p_n is in the resolvent set, so that $\hat{H}(p_n)$ is completely continuous; then $\hat{H}(p)$ is completely continuous, too. If V is not C^1, we can find a sequence $\{V_n\} \subset C^1$ such that V_n converges uniformly towards V. We obtain a corresponding sequence of completely continuous maps $\hat{H}_n(p)$ converging to $\hat{H}(p)$ in the uniform operator topology. This guarantees that $\hat{H}(p)$ is completely continuous, for all $p \in \mathbb{C}$. □

Let us recall now some results on the "functional calculus for unbounded linear maps" given in Hille–Philips [10].

THEOREM 2.6. *Suppose Λ is the infinitesimal generator of a C^0 semigroup on the Banach space Y and μ is a real-valued absolutely continuous measure with compact support in $\mathbb{R}_+$. If for all $z \in \mathbb{C}$,*

$$\psi(z) = \int_0^\infty e^{zt}\, d\mu(t)$$

and, for all $y \in Y$,

$$\psi(\Lambda)u = \int_0^\infty e^{t\Lambda} y\, d\mu(t),$$

then the following spectral mapping result holds:

$$\sigma(\psi(\Lambda)) = \psi(\sigma(\Lambda)) \cup \{0\}.$$

Our next result is concerned with a characterization of $\mathcal{S}$ in terms of an infinite system of scalar equations labeled by the eigenvalues of L.

THEOREM 2.7. *Let*

$$G_h(a) = \int_{a-r}^{a} g(a-s)V(s) \exp(\lambda_h s)\, ds, \qquad \lambda_h \in \sigma_p(L).$$

Then for all $p \in \mathcal{S}$ there exists $h \in N$ such that p is a solution to

$$\hat{G}_h(p) = 1.$$

Proof. Let us consider $p \in \mathcal{S}$; then $1 \in \sigma_p(\hat{K}(p))$, that is $1 \in \sigma_p(\hat{g}(p)\hat{H}(p))$. By Theorem 2.6, applied to

$$\Lambda = L - p, \qquad d\mu(t) = V(t)\, dt, \quad \psi(p) = \hat{V}(-p),$$

one has

$$\hat{H}(p) = \hat{V}(p - L).$$

Therefore, $1 \in \sigma_p(\hat{g}(p)\hat{H}(p))$ if and only if $1/\hat{g}(p) \in \sigma_p(\hat{V}(p-L))$. Indeed, $\hat{g}(p) = 0$ implies $\hat{K}(p) = 0$ and then $p \notin \mathcal{S}$. Therefore, $1 \in \sigma_p(\hat{K}(p))$ if and only if there exists $\lambda_h \in \sigma_p(L)$ such that $1/\hat{g}(p) = \hat{V}(p - \lambda_h)$, that is, $\hat{G}_h(p) = \hat{g}(p)\hat{V}(p-\lambda_h) = 1$. □

In order to apply residual calculus to the inverse Laplace transform of $\hat{B}(p)$, it is important to know the distribution of the singular points in the complex plane.

THEOREM 2.8. *Let $p^{(h)} = \max\{\operatorname{Re} p : \hat{G}_h(p) = 1\}$; then one has:*

i) *$\hat{G}_h(p^{(h)}) = 1$ and for all other p such that $\hat{G}_h(p) = 1$ it follows that $\operatorname{Re} p < p^{(h)}$.*

ii) $p^{(0)} > p^{(1)} \geqq \cdots p^{(h)} \geqq p^{(h+1)} \cdots$.

iii) *On each vertical strip $|\operatorname{Re} p| \leqq \text{const}$ there is a finite number of elements of $\mathcal{S}$. Then $\mathcal{S}$ is closed.*

Proof. Statement (i) follows by the classical theory on the Lotka–Von Foerster characteristic equation applied to $\hat{G}_h(p) = 1$ (see [11]). In order to prove (ii), we observe that by (H_3), one has $G_0(a) > G_1(a)$; then $p^{(0)} > p^{(1)}$ since $\hat{G}_0(p) = 1 = \hat{G}_1(p^{(1)})$. Without loss of generality, we assume $\operatorname{Re} \lambda_{n+1} \leqq \operatorname{Re} \lambda_n$; then, as above, $p^{(h+1)} \leqq p^{(h)}$.

Let us now prove (iii). Let $p = \xi + i\eta$. Then the Riemann–Lebesgue lemma implies that $\|\hat{K}(\xi + i\eta)\|$ tends to 0 as $|\eta|$ tends to ∞ uniformly in ξ on compact intervals. □

COROLLARY 2.8. *The above theorem implies that $p^{(0)}$ is an isolated singular point and, for each $p \in \mathcal{S}$, $p \neq p^{(0)}$, it follows that* $\operatorname{Re} p < p^{(0)}$.

THEOREM 2.9. *The singular point $p^{(0)}$ is a pole of finite order for the map $p \to (1 - \hat{K}(p))^{-1}$. That is, it is possible to find a positive integer s, a positive real number ν such that*

$$(1 - \hat{K}(p))^{-1} = M(p)(1 - \hat{G}_0(p))^{-s}, \tag{2.9}$$

for $0 < |p - p^{(0)}| \leqq \nu$, where $M(p)$ is analytic in $|p - p^{(0)}| \leqq \nu$.

Proof. With the same argument used in the proof of Theorem 2.7, we can see that $\{\hat{G}_h(p): h \in N\}$ is the set of the eigenvalues of $\hat{K}(p)$. Choose a neighborhood $|p - p^{(0)}| \leqq \nu$ of $p^{(0)}$, such that $\hat{G}_0(p)$ is the unique eigenvalue of $\hat{K}(p)$ in the open disk $|\lambda - 1| < \sigma$, while the other eigenvalues $\hat{G}_h(p)$, $h \geqq 1$, are outside $|\lambda - 1| = \rho$. Where $\nu, \sigma > 0$, $\rho > \sigma$. We observe that $\hat{G}_0(p)$ is a pole for the map $\lambda \to (\lambda - \hat{K}(p))^{-1}$ having the same order of λ_0 as a pole of $\zeta \to (\zeta - L)^{-1}$. Indeed (see Yosida [18, Chap. VIII, Theorem 4], this order depends on the dimension of the "range" of the operator

$$A_{-1}(p) = \frac{1}{2\pi i} \oint (\lambda - \hat{K}(p))^{-1} \, d\lambda.$$

By induction, it is possible to prove

$$[\hat{K}(p) - \hat{G}_0(p)]^{m-j}[L - \lambda_0]^j x = 0,$$

provided that $[\hat{K}(p) - \hat{G}_0(p)]^m x = 0$. Indeed, $\hat{G}_0(p)x = \hat{K}(p)x$ implies that $e^{\lambda_0 t}x = e^{tL}x$, by inverting the Laplace transforms. Then, by (H_3), it follows that $\lambda_0 x = Lx$. Moreover,

$$\hat{K}(p)[\hat{K}(p) - \hat{G}_0(p)]^{n-j-1}[L - \lambda_0]^j x = \hat{G}_0(p)[\hat{K}(p) - \hat{G}_0(p)]^{n-j-1}[L - \lambda_0]^j x;$$

then

$$L[\hat{K}(p) - \hat{G}_0(p)]^{n-j-1}[L - \lambda_0]^j x = \lambda_0[\hat{K}(p) - \hat{G}_0(p)]^{n-j-1}[L - \lambda_0]^j x.$$

So it follows that

$$[\hat{K}(p) - G_0(p)]^{n-j-1}[L - \lambda_0]^{j+1} x = 0.$$

In this way, we have obtained

$$\operatorname{range}(A_{-1}(p)) = \operatorname{range}\left(\frac{1}{2\pi i} \oint (z - L)^{-1} \, dz\right).$$

This argument shows that $\hat{G}_0(p)$ is a pole of fixed order, independent of p (when p is near $p^{(0)}$, of say, order s.
Then $(\lambda - \hat{K}(p))^{-1}$ can be expanded into a Laurent series in the set $0 < |\lambda - \hat{G}_0(p)| < \operatorname{dist}(\hat{G}_0(p), \partial B_\sigma)$, so that

$$(\lambda - \hat{K}(p))^{-1} = \sum_{j=-s}^{\infty} A_j(p)(\lambda - \hat{G}_0(p))^j,$$

where

$$A_j(p) = \frac{1}{2\pi i} \oint_{\partial B_\sigma} (\lambda - \hat{K}(p))^{-1}(\lambda - \hat{G}_0(p))^{-(n+1)} \, d\lambda, \qquad j = -s, \cdots, 0.$$

If we choose $\lambda = 1$, we get

$$(1-\hat{K}(p))^{-1} = (1-\hat{G}_0(p))^{-s} M(p), \qquad 0<|p-p^{(0)}|\leqq \nu,$$

where $M(p)$ is analytic in $|p-p^{(0)}|\leqq \nu$. □

Remark 2.10. Since $p^{(0)}$ is a simple root of $\hat{G}(p)=1$, then it is a pole of order s for the map $(1-G_0(p))^{-s}$; by (2.9), it is a pole of order s for the map $(1-\hat{K}(p))^{-1}$.

Now we go on to prove Theorem 2.2 stated above.

Proof. Let us consider $c\in\mathbb{R}$ such that $\operatorname{Re} p<c<p^{(0)}(=p_r^{(0)})$, for all $p\in\mathcal{S}\{p^{(0)}\}$. Since $B\in C^1$ by the inversion formula, it follows that

$$B(t)=\frac{1}{2\pi i}\int_{\beta-i\infty}^{\beta+i\infty}\hat{B}(p)\,e^{pt}\,dp.$$

We have that

$$\lim_{\theta\to+\infty}\int_{c+i\theta}^{\beta+i\theta}\hat{B}(p)\,e^{pt}\,dp=\lim_{\theta\to+\infty}\int_{c-i\theta}^{\beta-i\theta}\hat{B}(p)\,e^{pt}\,dp=0. \tag{2.10}$$

Indeed, if $\xi\in[c,\beta]$ it follows that

$$\left|\int_c^\beta \hat{B}(\xi+i\theta)\,e^{(\xi+i\theta)t}\,d\xi\right|\leqq\int_c^\beta \|(1-\hat{K}(\xi+i\theta))^{-1}\|\,|\hat{f}(\xi+i\theta)|\,e^{\xi t}\,d\xi.$$

By the Riemann–Lebesgue lemma, we get

$$\|\hat{K}(\xi+i\theta)\|\to 0, \qquad |\hat{f}(\xi+i\theta)|\to 0 \quad \text{as } \theta \text{ tends to } \infty \text{ uniformly in } \xi\in[c,\beta].$$

Then the above limits are proven. Therefore we get

$$B(t)=\operatorname*{Res}_{p=p^{(0)}}[\hat{B}(p)\,e^{pt}]+\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\hat{B}(p)\,e^{pt}\,dp. \tag{2.11}$$

Since the principal part of the Laurent expansion for $\hat{B}(p)\,e^{pt}$ is given by (near $p^{(0)}$),

$$\frac{a_{-1}}{(p-p^{(0)})}+\cdots+\frac{a_{-s}}{(p-p^{(0)})^s}, \qquad a_{-j}\in K, \quad j=1\cdots s, \tag{2.12}$$

then by residual calculus, we get

$$\operatorname*{Res}_{p=p^{(0)}}[\hat{B}(p)\,e^{pt}]=(a_{-1}+\cdots+a_{-s}t^{s-1})\,e^{p^{(0)}t}=B_0(t)\,e^{p^{(0)}t}.$$

Then it follows that

$$B(t)=B_0(t)\,e^{p^{(0)}t}+\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\hat{B}(p)\,e^{pt}\,dp. \tag{2.13}$$

To complete the proof of Theorem 2.2, we shall obtain the following estimates

$$\left|\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\hat{B}(p)\,e^{pt}\,dp\right|=O(e^{ct}) \quad \text{as } t \text{ tends to } \infty. \tag{2.14}$$

Indeed, one has

$$\hat{B}(p)=\hat{f}(p)+\hat{K}(p)(1-\hat{K}(p))^{-1}\hat{f}(p);$$

then

$$(2.15)\qquad \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\hat{B}(p)\,e^{pt}\,dp = f(t)+\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\hat{K}(p)(1-\hat{K}(p))^{-1}\hat{f}(p)\,dp.$$

We observe that $\hat{K}(c+i\dot{y})$ and $\hat{f}(c+iy)$ and the Fourier transforms of $e^{-ct}K(t)$ and $e^{-ct}f(t)$. Since K, f are continuous maps with compact supports then they belong to L^2. Consequently, by the Plancherel theorem (see [17]), the above transforms are in L^2 in the y variable. But $(1-\hat{K}(c+iy))^{-1}$ is bounded in y; then the product $\hat{K}(c+iy)$ $(1-\hat{K}(c+iy))^{-1}\hat{f}(c+iy)$ is L^1 in y. Thus, by (2.15), it follows that

$$\left|\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\hat{B}(c+iy)\,e^{(c+iy)t}\,dy\right| \leqq |f(t)|+\frac{e^{ct}}{2\pi}\int_{-\infty}^{+\infty}|\hat{K}(c+iy)(1-\hat{K}(c+iy))^{-1}\hat{f}(c+iy)|\,dy.$$

Since f has compact support, it follows that $|f(t)| = O(e^{ct})$. Therefore (2.14) is obtained. In this way, since $O(e^{ct}) = o(e^{p^{(0)}t})$, we get Theorem 2.2. □

Proof of Corollary 2.3. It is sufficient to point out that, in this case, $G_0(t) = V(t)\exp(\lambda_0 t)$, then $\hat{G}_0(p) = \hat{V}(p-\lambda_0)$. □

Remark 2.11. The same considerations made in Theorem 1.3 and Remarks 1.4, 1.5 are true also for the model (P'). Indeed, it is sufficient to repeat the above arguments replacing $\hat{K}$ with $\hat{G}_0$ and $\hat{V}(p)$ with $\hat{V}(p-\lambda_0)$.

COROLLARY 2.12. *The solution to the problem* (A) *has the following asymptotic expansion*:

$$u(a,t) = \exp\left(-\int_0^a m(a)\,da\right)T(a)(B_0(t-a)\,e^{p^{(0)}(t-a)}+o(e^{p^{(0)}(t-a)}) \quad \text{as } t \text{ tends to } \infty.$$

REFERENCES

[1] R. BELLMAN AND K. L. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] A. COALE, *The Growth and Structure of Human Populations*, Princeton University Press, Princeton, NJ, 1972.

[3] J. CUSHING, *Volterra integrodifferential equations in population dynamics*, in C.I.M.E. Mathematics of Biology, Cortona, 1979, to appear.

[4] G. DI BLASIO AND L. LAMBERTI, *An initial boundary value problem for age dependent population diffusion*, SIAM J. Appl. Math., 35 (1978), pp. 592–615.

[5] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[6] J. DOETSCH, *Theory and Applications of Laplace Transforms*, Springer-Verlag, Berlin-Heidelberg-New York, 1971.

[7] W. FELLER, *On the integral equation of renewal theory*, Ann. Math. Statist., 12 (1941), pp. 243–267.

[8] K. GOPALSAMY, *On the asymptotic age distribution in dispersive population*, Math. Biosci., 31 (1976), pp. 191–205.

[9] M. GURTIN, *A system of equations in age dependent population diffusion*, J. Theoret. Biol., 40 (1973), pp. 389–392.

[10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, 31, American Mathematical Society, Providence, RI, 1957.

[11] F. HOPPENSTEADT, *Mathematical Theory and Population Demographics, Genetics and Epidemics*, CBMS-NSF Regional Conference Series in Applied Mathematics 20, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[13] P. MARCATI AND R. SERAFINI, *Asymptotic behaviour in age dependent population dynamics with spatial spread*, Bull. Un. Mat. Ital., 16-B (1979), pp. 734–753.
[14] R. PALEY AND N. WIENER, *Theory and applications of Fourier transforms*, Trans. Amer. Math. Soc., 35 (1933).
[15] B. STEWART, *Generation of analytic semigroup by strongly elliptic linear operators*, Trans. Amer. Math. Soc., 199 (1974) pp. 141–162.
[16] ———, *Generation of a analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299–310.
[17] G. WEBB, *A recovery-relapse epidemic model with spatial spread*, preprint 1980.
[18] K. YOSIDA, *Functional Analysis*, 5th ed., Springer-Verlag, Berlin-Heidelberg-New York, 1978.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0012 $01.00/0

ANALYSIS OF GALERKIN APPROXIMATIONS OF A CLASS OF PSEUDOMONOTONE DIFFUSION PROBLEMS*

G. ALDUNCIN† AND J. T. ODEN†

Abstract. A class of nonlinear parabolic problems characterized by convective terms which depend nonlinearly on the solution and its gradient, are considered. Specifically, the operators characterizing the problems are shown to be pseudomonotone and to satisfy Gårding inequalities. The existence, uniqueness, and Galerkin and Faedo–Galerkin approximations of the general class of nonlinear diffusion problems are investigated.

1. Introduction. In this paper we are concerned with the existence, uniqueness, and Galerkin and Faedo–Galerkin approximations of the following general class of nonlinear diffusion problems: Let Ω be a bounded domain in $\mathbb{R}^n$ with boundary $\partial\Omega$, and $0<T<\infty$. Given data f in $\Omega\times(0,T)$ and initial data u_0 on Ω, find $u=u(x,t)$, $(x,t)\in\Omega\times(0,T)$, such that

$$\frac{\partial u}{\partial t}-\nabla\cdot\mathbf{a}(\nabla u)+b(u,\nabla u)=f \quad \text{in } Q=\Omega\times(0,T),$$

$$u=0 \quad \text{on } \Sigma=\partial\Omega\times(0,T), \tag{1.1}$$

$$u(\cdot,0)=u_0 \quad \text{on } \Omega,$$

where

$$\mathbf{a}(\nabla u)=a\nabla u+k|\nabla u|^{p-2}\nabla u,$$

$$a,k\in L^\infty(\Omega), \qquad 2\leqq p<\infty, \tag{1.2}$$

$$a(x)\geqq a_0\geqq 0 \quad \text{and} \quad k(x)\geqq k_0>0, \quad \text{a.e. } x\in\Omega,$$

and $b(u,\nabla u)=b(x,u(x,t),\nabla u(x,t))$ is subject to the conditions:

CI. $$|b(\zeta,\boldsymbol{\zeta})|\leqq c|\zeta|^q|\boldsymbol{\zeta}|^r \quad \forall(\zeta,\boldsymbol{\zeta})\in\mathbb{R}\times\mathbb{R}^n,$$

$$q=0 \quad \text{or} \quad q\geqq 1, \qquad r=0 \quad \text{or} \quad r\geqq 1, \tag{1.3}$$

$$1\leqq q+r<p-1.$$

CII. $b(\zeta,\boldsymbol{\zeta})$ is totally Fréchet differentiable in $\mathbb{R}\times\mathbb{R}^n$ and its partial derivatives $\partial_\zeta b:\mathbb{R}\times\mathbb{R}^n\to\mathscr{L}(\mathbb{R},\mathbb{R})$ and $\partial_{\boldsymbol{\zeta}} b:\mathbb{R}\times\mathbb{R}^n\to\mathscr{L}(\mathbb{R}^n,\mathbb{R})$ are such that, for (q,r) satisfying (1.3) and $\forall(\zeta,\boldsymbol{\zeta})\in\mathbb{R}\times\mathbb{R}^n$,

$$|\partial_\zeta b(\zeta,\boldsymbol{\zeta})|\leqq c_q|\zeta|^{q-1}|\boldsymbol{\zeta}|^r \quad \text{if } q\neq 0,$$

$$|\partial_{\boldsymbol{\zeta}} b(\zeta,\boldsymbol{\zeta})|\leqq c_r|\zeta|^q|\boldsymbol{\zeta}|^{r-1} \quad \text{if } r\neq 0.$$

The case $r=0$ will be understood as $b=b(u)$ (not a function of ∇u), and the case $q=0$ as $b=b(\nabla u)$ (not a function of u).

It is well known that nonlinear diffusion terms such as those represented by the term $a(\nabla u)$ in (1.1), are useful in modeling nonlinear heat conduction. They also occur in models of the flow of non-Newtonian fluids, particularly in the study of molten metals

* Received by the editors June 21, 1978, and in final revised form November 10, 1980. The work reported here was completed during the course of a project supported by the U.S. Army Research Office–Durham under grant DAAG 29-77-G-0087.

† Texas Institute for Computational Mechanics, University of Texas at Austin, Texas 78712.

and in certain problems of flow through porous media. However, it is now widely recognized that convection and advection play an important role in many of these physical processes, and that adequate mathematical models of such processes should frequently include the effects of nonlinear convective terms, such as $b(u, \nabla u)$. The presence of such convective terms, however, leads to solutions which differ considerably from those obtained in pure diffusion problems; solutions can exhibit shock-like fronts; uniqueness, stability and regularity of solutions become more important issues, and the analysis of the behavior of approximate solutions is significantly more complicated. It is standard practice in studies of Galerkin approximations of such nonlinear convection-diffusion problems to restrict the classes of problems under study in such a way that the methods of monotone operators can be used to obtain error estimates and theorems on convergence. While such studies are not without some value, they sidestep the major difficulties mentioned above and may not be applicable to many problems of physical interest.

Our objective in this paper is to analyze two types of approximations of classes of nonmonotone parabolic problems of the type (1.1) using the theory of pseudomonotone operators. These include fully-discrete Galerkin methods and, under some additional restrictions, semi-discrete Faedo–Galerkin methods of approximation. We note that, in general, this type of semi-discrete approximation is not necessarily well-defined for coercive pseudomonotone parabolic problems.

Following this introduction, in § 2, we show that the spatial operator in (1.1) is coercive and pseudomonotone on a dense continuously embedded subspace of the Banach space $L^p(0, T; W_0^{1,p}(\Omega))$ and that this implies that solutions to (1.1) do exist in $L^\infty(0, T; L^2(\Omega)) \cap L^p(0, T; W_0^{1,p}(\Omega))$ [6]. In general, multiple solutions will exist to (1.1) and there cannot exist a continuous dependence on the data. However, regularity conditions on the solutions can be given which will guarantee their uniqueness, and these are discussed in § 3.

In § 4 we introduce an elliptic regularization of problem (1.1) of the type used by Lions [5], and describe properties corresponding to Galerkin approximations and we give an approximation theorem which establishes their strong convergence. We also derive error estimates for such approximations. Finally in § 5 we describe Faedo–Galerkin (semi-discrete) approximations and show that, in the case of our model problem, sufficient conditions are satisfied which guarantee the existence and also *uniqueness* of these approximations. We also prove sufficient conditions for weak and strong convergence of such approximations and establish corresponding approximation error estimates.

Notation. $(V, \|\cdot\|)$ is a real, separable, reflexive Banach space with topological dual $(V', \|\cdot\|_*)$, $\langle \cdot, \cdot \rangle$ denotes the duality pairing on $V' \times V$. $(H, (\cdot, \cdot), |\cdot|)$ is a real Hilbert space identified with its dual, in which V is densely and continuously embedded: $V \hookrightarrow H \hookrightarrow V'$.

$(\mathcal{V}, |||\cdot|||)$ denotes the usual space $L^p(0, T; V)$, $2 \leqq p < \infty$, which is a separable, reflexive Banach space, whose dual space can be identified as $(\mathcal{V}', |||\cdot|||_*) = L^{p'}(0, T; V')$, $p' = p/(p-1)$, $[\cdot, \cdot]$ denoting the corresponding duality pairing. $(\mathcal{H}, (\cdot, \cdot)_{\mathcal{H}}, |\cdot|_{\mathcal{H}})$ denotes the Hilbert space $L^2(0, T; H)$ which, being identified with its dual, is such that $\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}'$.

$(\mathcal{U}, |||\cdot|||_{\mathcal{U}})$ and $(\mathcal{W}, |||\cdot|||_{\mathcal{W}})$ denote the separable, reflexive Banach spaces

$$\mathcal{U} = \{v : v \in \mathcal{V}, \dot{v} \in \mathcal{H}\}, \qquad |||v|||_{\mathcal{U}} = |||v||| + |||\dot{v}|||_{\mathcal{H}},$$

$$\mathcal{W} = \{v : v \in \mathcal{V}, \dot{v} \in \mathcal{V}'\}, \qquad |||v|||_{\mathcal{W}} = |||v||| + |||\dot{v}|||_*.$$

Here $\dot{v} = \partial v/\partial t$ is the distributional time derivative of v which belongs to $\mathscr{D}'((0, T); V) = \mathscr{L}(\mathscr{D}((0, T)), V)$. Hence (cf. [4] and [5]), $\mathscr{U} \hookrightarrow \mathscr{W} \hookrightarrow \mathscr{V}$, $\mathscr{W}$ is continuously embedded in $C([0, T]; H)$ and, if $u, v \in \mathscr{W}$, u, v satisfy the Green's formula

$$[\dot{u}, v] = (u(T), v(T)) - (u(0), v(0)) - [\dot{v}, u]. \tag{1.4}$$

Moreover, the trace mappings $v \mapsto v(0)$ and $v \mapsto v(T)$ are such that $\{v(0): v \in \mathscr{W}\} = H = \{v(T): v \in \mathscr{W}\}$, $\{v(0): v \in \mathscr{U}\}$ and $\{v(T): v \in \mathscr{U}\}$ are dense in H.

2. Existence analysis. For the model problem (1.1), we take as spaces V and H, the usual Sobolev spaces

$$V = W_0^{1,p}(\Omega), \qquad 2 \leqq p \leqq \infty, \quad H = L^2(\Omega). \tag{2.1}$$

Then, problem (1.1) assumes the following abstract form:

Find $u \in \mathscr{W}$ such that

$$\begin{aligned} &\frac{\partial u}{\partial t} + A(u) = f, \qquad f \text{ given in } \mathscr{V}', \\ &u(0) = u_0, \qquad u_0 \text{ given in } L^2(\Omega), \end{aligned} \tag{2.2}$$

where $A: \mathscr{V} \to \mathscr{V}'$ is defined by

$$\begin{aligned} &[A(u), v] = [A_1(u), v] + [A_2(u), v], \\ &[A_1(u), v] = \int_Q \mathbf{a}(x, \nabla u(x, t)) \cdot \nabla v(x, t)\, dx\, dt, \\ &[A_2(u), v] = \int_Q b(x, u(x, t), \nabla u(x, t)) v(x, t)\, dx\, dt, \end{aligned} \tag{2.3}$$

in which $\mathbf{a}(\nabla u)$ is as defined in (1.2) and $b(u, \nabla u)$ is subject to conditions CI and CII.

We now proceed to establish the existence of solutions to problem (2.2). The following two theorems determine basic properties of the operator A.

THEOREM 2.1. *Let $A: \mathscr{V} \to \mathscr{V}'$ be the operator defined in* (2.3). *Then* i) *A is bounded,* ii) *A is coercive, and* iii) *A is locally Lipschitz continuous in the sense that $\forall u, v \in B_\mu(0) = \{v \in \mathscr{V}: |||v||| < \mu, \mu > 0\}$, $w \in \mathscr{V}$, there is a positive constant $C(\mu)$ such that*

$$|[A(u) - A(v), w]| \leqq C(\mu) |||u - v||| \cdot |||w|||. \tag{2.4}$$

Proof. We shall use the notation

$$a_\infty = \|a\|_{L^\infty(\Omega)}, \quad k_\infty = \|k\|_{L^\infty(\Omega)} \quad \text{and} \quad \|\cdot\|_{s,Q} = \|\cdot\|_{L^s(Q)}.$$

i) Applying Hölder's inequality, we easily obtain that

$$\begin{aligned} |||A(v)|||_* \leqq\ & a_\infty \operatorname{mes}(Q)^{(p-2)/p} |||v||| + k_\infty |||v|||^{p-1} \\ & + c \operatorname{mes}(Q)^{(p-1-q-r)/p} |||v|||^{q+r} \quad \forall v \in \mathscr{V}. \end{aligned} \tag{2.5}$$

ii) From Friedrichs' inequality, it follows that, $\forall v \in L^s(0, T; W_0^{1,s}(\Omega))$, $1 \leqq s < \infty$,

$$\|v\|^s_{L^s(0,T;W_0^{1,s}(\Omega))} \leqq (c^s(s, n) \operatorname{mes}(\Omega)^{s/n} + 1) \|\nabla v\|^s_{s,Q}. \tag{2.6}$$

Thus,

$$\begin{aligned} [A(v), v] \geqq\ & a_0 \|\nabla v\|^2_{2,Q} + k_0 (1 + c^p(p, n) \operatorname{mes}(\Omega)^{p/n})^{-1} |||v|||^p \\ & - c \operatorname{mes}(Q)^{(p-1-q-r)/p} |||v|||^{1+q+r} \quad \forall v \in \mathscr{V}. \end{aligned} \tag{2.7}$$

But, using Young's inequality in the last term in (2.7), leads to

$$[A(v), v] \geqq a_0 \|\nabla v\|_{2,Q}^2 + \gamma_1 |||v|||^p - \gamma_2 T \quad \forall v \in \mathcal{V}, \tag{2.8}$$

where γ_1 and γ_2 are >0. Therefore, $[A(v), v]/|||v||| \to +\infty$ as $|||v||| \to \infty$.

iii) By the inequality in $\mathbb{R}^n$ [10],

$$\begin{aligned} &| |x|^{r-2}x - |y|^{r-2}y| < c\{|x|+|y|\}^{r-2}|x-y|, \\ &c = \sqrt{r-1} \quad \text{if } 2 \leqq r \leqq 3, \qquad c = r-1 \quad \text{if } 3 \leqq r < \infty \end{aligned} \tag{2.9}$$

and Hölder's inequality, we obtain, $\forall u, v \in B_\mu(0) \subset \mathcal{V}$, $w \in \mathcal{V}$,

$$|[A_1(u) - A_1(v), w]| \leqq \{a_\infty \operatorname{mes}(Q)^{(p-2)/p} + k_\infty c(p)(2\mu)^{p-2}\} |||u-v||| \, |||w|||. \tag{2.10}$$

We now use hypothesis CII. First observe that

$$\begin{aligned} [A_2(u) - A_2(v), w] &= \int_Q \int_0^1 \frac{db(\xi, \nabla\xi)}{d\theta} w \, d\theta \, dQ \\ &= \int_Q \int_0^1 \{\partial_\xi b(\xi, \nabla\xi)\eta + \partial_{\nabla\xi} b(\xi, \nabla\xi) \cdot \nabla\eta\} w \, d\theta \, dQ, \end{aligned} \tag{2.11}$$

where $\xi = v + \theta\eta$, $\eta = u - v$ and $\theta \in [0, 1]$. Hence, because of CII and Hölder's inequality,

$$|[A_2(u) - A_2(v), w]| \leqq (c_q + c_r) \operatorname{mes}(Q)^{(p-1-q-r)/p} \int_0^1 |||\xi|||^{q+r-1} \, d\theta \, |||\eta||| \, \|w\|_{p,Q}.$$

Then, since $u, v \in B_\mu(0) \subset \mathcal{V}$, there is a constant $\gamma_3 = \gamma_3(p, q, r, Q)$ such that

$$|[A_2(u) - A_2(v), w]| \leqq \gamma_3 \mu^{q+r-1} |||u-v||| \, \|w\|_{L^p(\Omega)}. \tag{2.12}$$

Therefore, (2.4) follows from estimates (2.10) and (2.12) and the proof of the theorem is completed. □

The next property of A, established below, is crucial, not only in proving the existence of solutions to (2.2) but in subsequent studies of approximations.

THEOREM 2.2. *The operator $A: \mathcal{V} \to \mathcal{V}'$ defined in* (2.3), *satisfies the following nonlinear Gårding-type inequality*: $\forall u, v \in B_\mu(0) \subset \mathcal{V}$,

$$\begin{aligned} [A(u) - A(v), u - v] \geqq \alpha_0 a_0 \|u-v\|_{L^2(0,T,H_0^1(\Omega))}^2 + \alpha_1 |||u-v|||^p \\ - \alpha_2(\mu) \|u-v\|_{L^p(Q)}^{p'}. \end{aligned} \tag{2.13}$$

where $H_0^1(\Omega) = W_0^{1,2}(\Omega)$ and $\alpha_0 > 0$, $\alpha_1 > 0$, $\alpha_2(\mu) > 0$.

Proof. We observe that, $\forall u, v \in \mathcal{V}$,

$$[A(u) - A(v), u - v] \geqq [A_1(u) - A_1(v), u - v] - |[A_2(u) - A_2(v), u - v]|. \tag{2.14}$$

From the inequality in $\mathbb{R}^n$ [10],

$$(|x|^{r-2}x - |y|^{r-2}y, x - y) \geqq 2^{1-r}|x-y|^r, \qquad 2 \leqq r < \infty \tag{2.15}$$

and (2.6), it follows that

$$\begin{aligned} [A_1(u) - A_1(v), u - v] \geqq a_0 (1 + c^2(2, n) \operatorname{mes}(\Omega)^{2/n})^{-1} |||u-v|||_{L^2(0,T;W_0^{1,2}(\Omega))}^2 \\ + k_0 2^{1-p} (1 + c^p(p, n) \operatorname{mes}(\Omega)^{p/n})^{-1} |||u-v|||^p. \end{aligned} \tag{2.16}$$

On the other hand, according to (2.12) and Young's inequality for $u, v \in B_\mu(0) \subset \mathcal{V}$ and

any $b>0$,

$$|[A_2(u)-A_2(v), u-v]| \leqq \frac{b^p}{p}|||u-v|||^p + \frac{\gamma_3^{p'}\mu^{(q+r-1)p'}}{p'b^{p'}}\|u-v\|^{p'}_{L^p(Q)}. \tag{2.17}$$

Therefore, by introducing (2.16) and (2.17) into (2.14) and choosing b small enough, the desired result (2.13) is obtained. □

THEOREM 2.3. *For any data $f \in \mathcal{V}'$ and $u_0 \in L^2(\Omega)$, there exists at least one solution $u \in \mathcal{W}$ to problem* (2.2).

Proof. Theorems 2.1 and 2.2 and Aubin's compactness theorem [1] confirm that conditions in [8] are satisfied. Therefore, A is coercive and $\mathcal{W}$-pseudomonotone from $\mathcal{V} \to \mathcal{V}'$ and, by virtue of Lions [6, Chapt. 3, Thm. 1.2], the assertion of the theorem follows. □

Remark 2.1. From the proofs of Theorems 2.1 and 2.2, it is apparent that the operator A of (2.3) regarded as a map from V into V', is bounded, coercive and locally Lipschitz continuous, and satisfies the Gårding-type inequality

$$\begin{aligned}\langle A(u)-A(v), u-v\rangle \geqq \alpha_0 a_0\|u-v\|^2_{H_0^1(\Omega)} + \alpha_1\|u-v\|^p \\ -\hat{\alpha}_2(\rho)\|u-v\|^{p'}_{L^p(\Omega)} \quad \forall u, v \in B_\rho(0) \subset V.\end{aligned} \tag{2.18}$$

According to Oden [8], $A\colon V \to V'$ is necessarily V-pseudomonotone. Hence, from the theory of pseudomonotone elliptic equations (cf. [6]), A is surjective from $V \to V'$; i.e., there exists at least one solution in V to the stationary problem

$$A(u) = f, \qquad f \text{ given in } V'. \tag{2.19}$$

The evolution problem (2.2) possesses at least one equilibrium state for each $f \in V'$.

3. Sufficient conditions for uniqueness. We now proceed to determine sufficient conditions for uniqueness of solutions to the pseudomonotone diffusion problem (2.2).

In the case of monotone parabolic problems, "monotonicity" → "uniqueness" and this follows from the Carathéodory type differential inequality $d|u(t)-v(t)|^2/dt \leqq 0$, a.e. $t \in [0, T]$, $|u(0)-v(0)|^2 = 0$, the unique solution of which is $|u(t)-v(t)|^2 = 0$, u and v being solutions of the problem. This suggests that in the nonmonotone case with Gårding-type inequalities, the possibility of establishing a differential inequality of the form

$$\begin{aligned}\frac{d}{dt}|u(t)-v(t)|^s \leqq \alpha|u(t)-v(t)|^s \quad \text{a.e. } t \in [0, T], \\ |u(0)-v(0)|^s = 0,\end{aligned} \tag{3.1}$$

or equivalently,

$$|u(\tau)-v(\tau)|^s \leqq \alpha\int_0^\tau |u(t)-v(t)|^s\,dt \quad \forall \tau \in [0, T] \tag{3.2}$$

for some $\alpha \in \mathbb{R}$ and $2 \leqq s < \infty$, would be sufficient for concluding uniqueness. Indeed, from Olech and Opial [9, Thm. 3], $|u(t)-v(t)|^s = 0$, is the unique solution to (3.1). We show that in certain particular cases and, in general, for sufficiently smooth solutions, problem (2.2) falls into this class.

THEOREM 3.1. *Let $u \in \mathcal{W}$ be a solution of problem* (2.2). *Then u is unique in the following three cases*:

(3.3) i) $r = 0$ *and* $q = 1$;
ii) $r = 0$ *and* $n < p$;
iii) $a_0 > 0$ *and* $u \in L^\infty(0, T; W_0^{1,\infty}(\Omega))$.[1]

Proof. Assume that $u = u(t; f, u_0)$ and $v = v(t; f, u_0)$ are two solutions of problem (2.2) and define $\eta = u - v$. It is apparent from (2.16) that

$$\langle A_1(u(t)) - A_1(v(t)), \eta(t)\rangle \geqq \hat{\alpha}_0 a_0 \|\eta(t)\|^2_{H_0^1(\Omega)} + \hat{\alpha}_1 \|\eta(t)\|^p \quad \text{for a.e. } t \in [0, T], \tag{3.4}$$

where $\hat{\alpha}_0 > 0$ and $\hat{\alpha}_1 > 0$. Thus, from the difference of the equations satisfied by u and v, we obtain the integral inequality

$$\begin{gathered} \tfrac{1}{2}|\eta(\tau)|^2 + \hat{\alpha}_0 a_0 \int_0^\tau \|\eta(t)\|^2_{H_0^1(\Omega)}\, dt \leqq \Lambda_\tau \equiv \left| \int_0^\tau \langle A_2(u(t)) - A_2(v(t)), \eta(t)\rangle\, dt \right|, \\ \hat{\alpha}_0 > 0 \quad \forall \tau \in [0, T]. \end{gathered} \tag{3.5}$$

We now estimate the right-hand side via the formula (2.11) with $w = \eta$.

i) $r = 0$ *and* $q = 1$. In this case we have the estimate

$$\Lambda_\tau \leqq c_q \int_0^\tau |\eta(t)|^2\, dt \quad \forall \tau \in [0, T], \tag{3.6}$$

which combined with (3.5) gives the integral inequality (3.2) with $\alpha = 2c_q$ and $s = 2$. Consequently, $\eta = 0$.

ii) $r = 0$ *and* $n < p$. From the Sobolev embedding theorem, $W_0^{1,p}(\Omega)$ is continuously embedded in $C_B(\Omega) = \{v \in C(\Omega)\colon v \text{ bounded in } \Omega\}$ whenever $n < p$. Then

$$\mathcal{V} = L^p(0, T; W^{1,p}(\Omega)) \hookrightarrow L^p(0, T; L^\infty(\Omega)), \quad n < p. \tag{3.7}$$

Let μ be chosen such that $u, v \in B_\mu(0) \subset \mathcal{V}$. Then, from (2.11), with $w = \eta$ and using (3.7), we obtain, $\forall \tau \in [0, T]$,

$$\begin{aligned} \Lambda_\tau &\leqq \int_0^1 \int_0^\tau cq \|\xi(t)\|^{q-1}_{L^\infty(\Omega)} |\eta(t)|^2\, dt\, d\theta \\ &\leqq c_q \mu^{q-1} \left(\int_0^\tau |\eta(t)|^s\, dt \right)^{2/s}, \qquad 2 \leqq s = \frac{2p}{p+1-q} < p. \end{aligned} \tag{3.8}$$

Introducing this estimate into (3.5) produces the integral inequality (3.2) with

$$\alpha = (2c_q \mu^{q-1})^{s/2} \quad \text{and} \quad 2 \leqq s = \frac{2p}{(p+1-q)} < p.$$

Therefore, $\eta = 0$.

iii) $a_0 > 0$ *and* $u \in L^\infty(0, T; W_0^{1,\infty}(\Omega))$. Let $\mu > 0$ be such that $u, v \in B_\mu(0) \subset L^\infty(0, T; W_0^{1,\infty}(\Omega))$. Then, from (2.11) with $w = \eta$, we obtain $\forall \tau \in [0, T]$,

$$\Lambda_\tau \leqq \mu^{q+r-1}(c_q c(2, n) \operatorname{mes}(\Omega)^{1/n} + c_r) \int_0^\tau \|\eta(t)\|_{H_0^1(\Omega)} |\eta(t)|\, dt. \tag{3.9}$$

[1] In this case, the question of existence appears to be open.

Hence, since by hypothesis $a_0 > 0$, we can apply Young's inequality with constant b, e.g., $b = \sqrt{\alpha_0 a_0}$, to obtain the upper bound for (3.9)

$$\frac{\alpha_0 a_0}{2} \int_0^\tau \|\eta(t)\|^2_{H_0^1(\Omega)}\, dt + \frac{\alpha}{2} \int_0^\tau |\eta(t)|^2\, dt,$$

where $\alpha = \alpha(1/b^2) > 0$. Combining these results with (3.5) gives (3.2) with $s = 2$, and $\eta = 0$. □

4. Galerkin approximations. In this section, we study Galerkin approximations of the model problem (2.2) which are based on an elliptic regularization of (2.2) obtained as suggested by Lions [6]. We will establish some results on the *strong* convergence of such approximations.

For the model problem (2.2), we introduce the elliptic regularization

$$\varepsilon(\dot{u}_\varepsilon, \dot{v})_{\mathcal{H}} - (u_\varepsilon, \dot{v})_{\mathcal{H}} + (u_\varepsilon(T), v(T)) + [A(u_\varepsilon), v] = [f, v] + (u_0, v(0)), \quad \forall v \in \mathcal{U}, \tag{4.1}$$

where A is the operator defined in (2.3). Following standard techniques discussed in detail by Lions [5] and [6], it can be shown that for every $\varepsilon > 0$, there exists at least one solution $u_\varepsilon \in \mathcal{U}$ to (4.1) and, that, in the sense of $\mathcal{V}' \subset \mathcal{L}(\mathcal{D}((0, T)), W^{-1,p'}(\Omega))$, u_ε satisfies the distributional equation

$$\begin{aligned} -\varepsilon \ddot{u}_\varepsilon + \dot{u}_\varepsilon + A(u_\varepsilon) &= f \quad \text{in } \mathcal{V}', \\ -\varepsilon \dot{u}_\varepsilon(0) + u_\varepsilon(0) &= u_0 \quad \text{in } L^2(\Omega), \\ \dot{u}_\varepsilon(T) &= 0 \quad \text{in } L^2(\Omega), \end{aligned} \tag{4.2}$$

which is equivalent to (4.1). Moreover, for any sequence $\{u_\varepsilon\}_{\varepsilon > 0} \subset \mathcal{U}$ of solutions, there exists a subsequence, also denoted $\{u_\varepsilon\}_{\varepsilon > 0}$, such that, as $\varepsilon \to 0^+$, u_ε converges weakly to a solution u of (2.2) in the sense:

$$\begin{aligned} u_\varepsilon &\rightharpoonup u && \text{weakly in } \mathcal{V}, \\ \frac{\partial u_\varepsilon}{\partial t} &\rightharpoonup \frac{\partial u}{\partial t} && \text{weakly in } \mathcal{V}', \\ \varepsilon \frac{\partial u_\varepsilon}{\partial t} &\rightharpoonup 0 && \text{weakly in } \mathcal{H}, \\ A(u_\varepsilon) &\rightharpoonup A(u) && \text{weakly in } \mathcal{V}', \\ u_\varepsilon(0) &\rightharpoonup u(0) && \text{weakly in } \mathcal{H}, \\ u_\varepsilon(T) &\rightharpoonup u(T) && \text{weakly in } \mathcal{H}. \end{aligned} \tag{4.3}$$

To construct Galerkin approximations of (4.1), we introduce a family of subspaces $\{\mathcal{U}_h\}_{0 < h \leq 1}$ of $\mathcal{U}$ such that: i) $\mathcal{U}_h$ is finite-dimensional with basis functions $\{\phi_1, \phi_2, \cdots, \phi_{m_h}\}$, with dimension $m_h \to \infty$, as $h \to 0^+$; and ii) $\bigcup_h \mathcal{U}_h$ is dense in $\mathcal{U}$. A Galerkin approximation of (4.1) involves seeking a function $U_\varepsilon^h \in \mathcal{U}_h$ such that

$$\begin{aligned} &\varepsilon(\dot{U}_\varepsilon^h, \dot{\phi}_k)_{\mathcal{H}} - (U_\varepsilon^h, \dot{\phi}_k)_{\mathcal{H}} + (U_\varepsilon^h(T), \phi_k(T)) + [A(U_\varepsilon^h), \phi_k] \\ &\quad = [f, \phi_k] + (u_0, \phi_k(0)), \qquad k = 1, 2, \cdots, m_h. \end{aligned} \tag{4.4}$$

The solvability in $\mathcal{U}_h$ of (4.4) is assured by the $\mathcal{U}$-pseudomonotonicity and coercivity of the operator $\mathcal{A}_\varepsilon : \mathcal{U} \to \mathcal{U}'$ in (4.1). Similarly, as for (4.1), if $\{U_\varepsilon^h\}_{0 < h \leq 1}$ is a sequence of Galerkin approximate solutions, it can be shown that there exists a function

u_ε and a subsequence, also denoted $\{U_\varepsilon^h\}_{0<h\leqq 1}$, such that, as $h\to 0^+$,

$$
\begin{aligned}
&U_\varepsilon^h \rightharpoonup u_\varepsilon && \text{weakly in } \mathcal{V},\\
&\dot{U}_\varepsilon^h \rightharpoonup \dot{u}_\varepsilon && \text{weakly in } L^2(Q),\\
&A(U_\varepsilon^h) \rightharpoonup A(u_\varepsilon) && \text{weakly in } \mathcal{V}',\\
&U_\varepsilon^h(0) \rightharpoonup u_\varepsilon(0) && \text{weakly in } L^2(\Omega),\\
&U_\varepsilon^h(T) \rightharpoonup u_\varepsilon(T) && \text{weakly in } L^2(\Omega).
\end{aligned}
\tag{4.5}
$$

We will now demonstrate that for our model problem (2.2) much stronger results can be obtained.

THEOREM 4.1. *Let $\{u_\varepsilon\}_{\varepsilon>0}\subset \mathcal{U}$ be a weakly convergent subsequence of solutions to problem* (4.1) *and with its weak limit $u\in \mathcal{W}$ a solution of* (2.2). *Then, as $\varepsilon\to 0^+$,*

$$
\begin{aligned}
&u_\varepsilon \to u && \text{strongly in } \mathcal{V},\\
&\sqrt{\varepsilon}\dot{u}_\varepsilon \to 0 && \text{strongly in } L^2(Q),\\
&u_\varepsilon(0) \to u_0 && \text{strongly in } L^2(\Omega),\\
&u_\varepsilon(T) \to u(T) && \text{strongly in } L^2(\Omega).
\end{aligned}
\tag{4.6}
$$

Proof. We regard equation (2.2) as holding on $\mathcal{U}$ and subtract (4.1) from it. The following orthogonality condition is obtained:

$$
-(\varepsilon\dot{u}_\varepsilon, \dot{v})_{\mathcal{H}} + (u_0 - u_\varepsilon(0), v(0)) + [\dot{u}-\dot{u}_\varepsilon, v] + [A(u)-A(u_\varepsilon), v] = 0 \quad \forall v\in \mathcal{U}. \tag{4.7}
$$

According to (4.3), there is a $\mu>0$ independent of ε, such that $u_\varepsilon, u\in B_\mu(0)\subset \mathcal{V}$. Hence, using formula (1.4) and the Gårding-type inequality in (2.13), we see that

$$
\begin{aligned}
&|u_0-u_\varepsilon(0)|^2 + [\dot{u}-\dot{u}_\varepsilon, u-u_\varepsilon] + [A(u)-A(u_\varepsilon), u-u_\varepsilon]\\
&\quad \geqq \tfrac{1}{2}|u_0-u_\varepsilon(0)|^2 + \tfrac{1}{2}|u(T)-u_\varepsilon(T)|^2 + \alpha_1 |||u-u_\varepsilon|||^p - \alpha_2(\mu)\|u-u_\varepsilon\|^{p'}_{L^p(Q)}.
\end{aligned}
\tag{4.8}
$$

Next, combining these two results, we conclude that

$$
\begin{aligned}
&\tfrac{1}{2}|u_0-u_\varepsilon(0)|^2 + \tfrac{1}{2}|u(T)-u_\varepsilon(T)|^2 + \alpha_1 |||u-u_\varepsilon|||^p\\
&\quad \leqq \alpha_2(\mu)\|u-u_\varepsilon\|^{p'}_{L^p(Q)} + (u_0-u_\varepsilon(0), u_0-v(0))\\
&\qquad + [\dot{u}-\dot{u}_\varepsilon, u-v] + [A(u)-A(u_\varepsilon), u-v] + (\varepsilon\dot{u}_\varepsilon, \dot{v})_{\mathcal{H}} - |\sqrt{\varepsilon}\dot{u}_\varepsilon|^2_{\mathcal{H}} \quad \forall v\in \mathcal{U}.
\end{aligned}
\tag{4.9}
$$

Due to the compact embedding of $\mathcal{W}$ in $L^p(Q)$ (cf. [1]) and the weak convergence result (4.3), (4.6) follows. □

THEOREM 4.2. *Let $\{U_\varepsilon^h\in \mathcal{U}_h\}_{0<h\leqq 1}$ be a subsequence of Galerkin approximate solutions defined by* (4.4), *converging weakly, in the sense of* (4.5), *to a solution $u_\varepsilon\in \mathcal{U}$ of problem* (4.1). *Then, for fixed $\varepsilon>0$, as $h\to 0^+$*

$$
\begin{aligned}
&U_\varepsilon^h \to u_\varepsilon && \text{strongly in } \mathcal{V},\\
&\dot{U}_\varepsilon^h \to \dot{u}_\varepsilon && \text{strongly in } L^2(Q),\\
&U_\varepsilon^h(0) \to u_\varepsilon(0) && \text{strongly in } L^2(\Omega),\\
&U_\varepsilon^h(T) \to u_\varepsilon(T) && \text{strongly in } L^2(\Omega).
\end{aligned}
\tag{4.10}
$$

Proof. We follow similar arguments to those given previously. Restricting (4.1) to $\mathcal{U}_h$ and subtracting (4.4) from it, we obtain the orthogonality condition

$$\begin{aligned}(4.11)\qquad &\varepsilon(\dot{u}_\varepsilon - \dot{U}_\varepsilon^h, \dot{W})_{\mathcal{H}} - (u_\varepsilon - U_\varepsilon^h, \dot{W})_{\mathcal{H}} + (u_\varepsilon(T) - U_\varepsilon^h(T), W(T)) \\ &\qquad + [A(u_\varepsilon) - A(U_\varepsilon^h), W] = 0 \quad \forall W \in \mathcal{U}_h.\end{aligned}$$

Now, from (4.5) and (4.3), there is a $\mu > 0$, independent of h, such that $U_\varepsilon^h, u_\varepsilon \in B_\mu(0) \subset \mathcal{V}$. Then, by virtue of (1.4) and (2.13), it follows that

$$\begin{aligned}(4.12)\qquad &\varepsilon|\dot{u}_\varepsilon - \dot{U}_\varepsilon^h|^2 - (u_\varepsilon - U_\varepsilon^h, \dot{u}_\varepsilon - \dot{U}_\varepsilon^h)_{\mathcal{H}} + |u_\varepsilon(T) - U_\varepsilon^h(T)|^2 + [A(u_\varepsilon) - A(U_\varepsilon^h), u_\varepsilon - U_\varepsilon^h] \\ &\geqq \varepsilon|\dot{u}_\varepsilon - \dot{U}_\varepsilon^h|^2 + \tfrac{1}{2}|u_\varepsilon(0) - U_\varepsilon^h(0)|^2 + \tfrac{1}{2}|u_\varepsilon(T) - U_\varepsilon^h(T)|^2 \\ &\quad + \alpha_0\|u_\varepsilon - U_\varepsilon^h\|^2_{L^2(0,T;H_0^1(\Omega))} + \alpha_1|||u_\varepsilon - U_\varepsilon^h|||^p - \alpha_2(\mu)\|u_\varepsilon - U_\varepsilon^h\|^{p'}_{L^p(Q)}.\end{aligned}$$

Therefore, combining (4.11) and (4.12),

$$\begin{aligned}(4.13)\qquad &\varepsilon|\dot{u}_\varepsilon - \dot{U}_\varepsilon^h|^2_{\mathcal{H}} + \tfrac{1}{2}|u_\varepsilon(0) - U_\varepsilon^h(0)|^2 + \tfrac{1}{2}|u_\varepsilon(T) - U_\varepsilon^h(T)|^2 \\ &\quad + \alpha_0\|u_\varepsilon - U_\varepsilon^h\|^2_{L^2(0,T;H_0^1(\Omega))} + \alpha_1|||u_\varepsilon - U_\varepsilon^h|||^p \\ &\qquad \leqq \alpha_2(\mu)\|u_\varepsilon - U_\varepsilon^h\|^{p'}_{L^p(Q)} + \varepsilon(\dot{u}_\varepsilon - \dot{U}_\varepsilon^h, \dot{u}_\varepsilon - \dot{W})_{\mathcal{H}} \\ &\qquad\quad - (u_\varepsilon - U_\varepsilon^h, \dot{u}_\varepsilon - \dot{W})_{\mathcal{H}} + (u_\varepsilon(T) - U_\varepsilon^h(T), u_\varepsilon(T) - W(T)) \\ &\qquad\quad + [A(u_\varepsilon) - A(U_\varepsilon^h), u_\varepsilon - W] \quad \forall W \in \mathcal{U}_h.\end{aligned}$$

But $\mathcal{U}$ is compactly embedded in $L^p(Q)$ [1] and U_ε^h converges weakly to u_ε in the sense of (4.5). Hence, the right side of (4.13) $\to 0$ as $h \to 0^+$, and this proves the theorem. □

We next give an error estimate for the Galerkin approximations of the regularized elliptic problem (4.1).

THEOREM 4.3. *For fixed $\varepsilon > 0$, let $u_\varepsilon \in \mathcal{U}$ be a solution of problem* (4.1) *which is the strong limit (in the sense of* (4.10)*) of the subsequence of Galerkin approximate solutions $\{U_\varepsilon^h \in \mathcal{U}_h\}_{0<h\leqq 1}$ defined by* (4.4). *Then the following approximation error estimate holds $\forall W \in \mathcal{U}_h$:*

$$\begin{aligned}(4.14)\qquad &\tfrac{1}{2}C_1|u_\varepsilon(0) - U_\varepsilon^h(0)|^2 + \tfrac{1}{2}|u_\varepsilon(T) - U_\varepsilon^h(T)|^2 \\ &\quad + \alpha_0 a_0\|u_\varepsilon - U_\varepsilon^h\|^2_{L^2(0,T;H_0^1(\Omega))} + \tilde{\alpha}|||u_\varepsilon - U_\varepsilon^h|||^p + \tilde{\varepsilon}|\dot{u}_\varepsilon - \dot{U}_\varepsilon^h|^2_{\mathcal{H}} \\ &\qquad \leqq \alpha_2\|u_\varepsilon - U_\varepsilon^h\|^{p'}_{L^p(Q)} + C_2|u_\varepsilon(0) - W(0)|^2 \\ &\qquad\quad + C_3|u_\varepsilon - W|^2_{\mathcal{H}} + \tilde{C}|||u_\varepsilon - W|||^{p'} + C_4|\dot{u}_\varepsilon - \dot{W}|^2_{\mathcal{H}},\end{aligned}$$

where C_i, $i = 1, \cdots, 4$, α_0, $\tilde{\alpha} = \tilde{\alpha}_1(\alpha_1)$, $\alpha_2 = \alpha_2(T, \mu)$, $\tilde{\varepsilon} = \tilde{\varepsilon}(\varepsilon)$ and $\tilde{C} = \tilde{C}(C(T, \mu))$ are strictly positive constants. Here $C(T, \mu)$ is the local Lipschitz continuity constant of (2.4).

Proof. The estimate (4.14) follows directly from (4.13) upon applying formula (1.4), the local Lipschitz continuity of A, (2.4) and Young's inequality. □

5. Faedo–Galerkin approximations. We are concerned here with Faedo–Galerkin approximations of the model pseudomonotone diffusion problem (2.2). We note that this type of approximation process is not necessarily well-defined for nonmonotone parabolic problems: the corresponding weak convergence is a conditional property. We shall show that Faedo–Galerkin approximate solutions to problem (2.2) exist and are *unique*, and we shall determine sufficient conditions for weak and strong convergence.

Let $\{V_h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces approximating the space $V(=W_0^{1,p}(\Omega))$ in the following sense: (i) $\{\psi_1, \psi_2, \cdots, \psi_{m_h}\}$ denotes a basis for V_h, with dimension $m_h \to \infty$ as $h \to 0^+$; (ii) $\bigcup_h V_h$ is dense in V. A Faedo–Galerkin approximation in V_h of problem (2.2) is defined as an absolutely continuous function $U^h \in C_A([0, T]; V_h)$, which is a solution of the system

$$\begin{aligned} &\langle \dot{U}^h(t), \psi_k\rangle + \langle A(U^h(t)), \psi_k\rangle = \langle f(t), \psi_k\rangle, \qquad k = 1, 2, \cdots, m_h, \\ &U^h(0) = U_0^h, \end{aligned} \tag{5.1}$$

for a.e. $t \in [0, T]$ and where $U_0^h \to u_0$ strongly in $L^2(\Omega)$ as $h \to 0^+$. We observe that if U^h is solution of (5.1), then its time derivative $\dot{U}^h$ belongs to $L^{p'}(0, T; V_h)$.

We next establish the solvability of problem (5.1).

THEOREM 5.1. *For each $h \in (0, 1]$, the Faedo–Galerkin approximation problem* (5.1) *possesses a unique solution $U^h \in C_A([0, T]; V_h)$ continuous with respect to U_0^h.*

Proof. The local existence of solutions to (5.1) in $C_A([0, t_h]; V_h)$, $t_h > 0$, is implied by the pseudomonotonicity property of A (cf. Remark 2.1). Indeed, $f \in \mathcal{V}'$, and A is necessarily bounded and demicontinuous from $V \to V'$ and these are sufficient conditions for the vector field $\mathbf{F}(t, \mathbf{U}) = (\langle f(t), \psi_k\rangle - \langle A(U(t)), \psi_k\rangle)$ from $D = [0, T] \times \mathbb{R}^{m_h} \to \mathbb{R}^{m_h}$ to satisfy the Carathéodory conditions in D. Here $\mathbf{U} \in \mathbb{R}^{m_h}$ denotes the coordinate vector of $U \in V_h$ with respect to the reciprocal basis of V_h.

The uniqueness and continuous dependence on the initial data of local solutions to problem (5.1) follows from the condition [3] that for each compact set $w \subset D$, there is a function $g_w \in L^1(0, T)$ such that

$$|\mathbf{F}(t, \mathbf{U}) - \mathbf{F}(t, \mathbf{W})| \leqq g_w(t)|\mathbf{U} - \mathbf{W}|, \qquad (t, \mathbf{U}), (t, \mathbf{W}) \in w, \tag{5.2}$$

which is satisfied because A is locally Lipschitz continuous from $V \to V'$ (cf. Remark 2.1).

It remains to be shown that the interval of existence $[0, t_h] = [0, T]$. This is a consequence of the coercivity of A from $V \to V'$, as follows from part (1) of the proof of Theorem 5.2, given below. □

We now proceed to analyze the convergence of the Faedo–Galerkin approximation process.

THEOREM 5.2. *From the sequence of Faedo–Galerkin approximate solutions defined uniquely by* (5.1), *there is a subsequence, also denoted $\{U^h\}_{0<h\leq 1}$, and there exist functions $u \in \mathcal{W}$ and $\mathcal{X} \in \mathcal{V}'$ such that, as $h \to 0^+$,*

$$\begin{aligned} &U^h \rightharpoonup u && \text{weakly in } \mathcal{V}, \\ &U^h \rightharpoonup u && \text{weakly* in } L^\infty(0, T; L^2(\Omega)), \\ &A(U^h) \rightharpoonup \mathcal{X} && \text{weakly in } \mathcal{V}', \\ &U^h(T) \rightharpoonup u(T) && \text{weakly in } L^2(\Omega), \end{aligned} \tag{5.3}$$

and

$$\left[\frac{\partial u}{\partial t}, v\right] + [\mathcal{X}, v] = [f, v] \quad \forall v \in \mathcal{V}, \quad u(0) = u_0. \tag{5.4}$$

Moreover, the limit function u is a solution of problem (2.2) (*i.e.*, $\mathcal{X} = A(u)$) *provided one of the following conditions is satisfied*:

(5.5) i) $\dot{U}^h \in \mathcal{V}'$, $0 < h \leqq 1$, *and* $\{|||\dot{U}^h|||_*\}_{0<h\leq 1}$ *is bounded*;

(5.6) ii) $A: \mathcal{V} \to \mathcal{V}'$ *of* (2.3) *is $\mathcal{V}$-pseudomonotone.*

Proof. We follow the usual pseudonomontone method which consists of: 1) finding a priori bounds; 2) passage to the limit; and 3) the pseudomonotonicity argument.

1) From the proof of the coercivity property of A, (2.8), it is apparent that A is also coercive from $V \to V'$. Then, by standard arguments, it follows that the sequence $\{U^h\}_{0<h\leqq 1}$ turns out to be bounded in $\mathscr{V}$ and in $L^\infty(0, T; L^2(\Omega))$.

2) With the previous result and the boundedness of A from $\mathscr{V} \to \mathscr{V}'$ given by (2.5), the validity of (5.3) follows via weak compactness arguments and, then, upon the passage to the limit in equation (5.1), (5.4) is easily concluded (cf. [6, Chapt. 2]).

3) It remains to be shown that, if either (5.5) or (5.6) holds, then

$$[\mathscr{A}, v] = [A(u), v], \quad \forall v \in \mathscr{V}. \tag{5.7}$$

From (5.1), (5.3) and (5.4), we see that

$$\begin{aligned}\lim_{h\downarrow 0}\{[\dot{U}^h, U^h]+[A(U^h), U^h]\} &= \lim_{h\downarrow 0}[f, U^h] = [f, u]\\ &= [\dot{u}, u]+[\mathscr{A}, u]\\ &= \lim_{h\downarrow 0}\{[\dot{u}, U^h]+[A(U^h), u]\}.\end{aligned}$$

Therefore,

$$\lim_{h\downarrow 0}[A(U^h), U^h - u] = -\lim_{h\downarrow 0}[\dot{U}^h - \dot{u}, U^h] = -\tfrac{1}{2}\lim_{h\downarrow 0}|U^h(T)-u(T)|^2 \leqq 0. \tag{5.8}$$

Now, by the usual arguments [6], (5.7) follows from (5.8) and the first statement of (5.3) when assuming either (5.5) and using the $\mathscr{W}$-pseudomonotonicity property of A: $\mathscr{V} \to \mathscr{V}'$, or (5.6). This completes the proof of the theorem. □

We next establish that condition (5.5) is also sufficient for the strong convergence of the approximation process.

Theorem 5.3. *Suppose the condition* (5.5) *holds with bound* $\mu' > 0$. *Then the subsequence* $\{U^h\}_{0<h\leqq 1}$ *of Faedo–Galerkin approximate solutions converging weakly to a solution* $u \in \mathscr{W}$ *of problem* (2.2), *in the sense of Theorem* 5.2, *is such that, as* $h \to 0^+$,

$$\begin{aligned} &U^h \to u \quad \textit{strongly in } L^\infty(0, T; L^2(\Omega)),\\ &U^h \to u \quad \textit{strongly in } \mathscr{V}.\end{aligned} \tag{5.9}$$

In fact, the following approximation error estimates hold $\forall Z \in L^p(0, T; V_h)$:

$$\begin{aligned}|u(\tau)-U^h(\tau)| \leqq |u_0 - U_0^h| &+ \tilde{K}_1(T, \mu)\|u - U^h\|^{p'/2}_{L^p(Q)}\\ &+ \tilde{K}_2(T, \mu, \mu')|||u - Z|||^{1/2} \quad \forall \tau \in [0, T];\end{aligned} \tag{5.10}$$

$$\begin{aligned}|||u - U^h||| \leqq \tilde{K}_3|u_0 - U_0^h|^{2/p} &+ \tilde{K}_4(T, \mu)\|u - U^h\|^{1/(p-1)}_{L^p(Q)}\\ &+ \tilde{K}_5(T, \mu, \mu')|||u - Z|||^{1/p},\end{aligned} \tag{5.11}$$

where $\mu > 0$ *is a bound for* u *and* $\{U^h\}_{0<h\leqq 1}$ *in* $\mathscr{V}$.

Proof. By using formula (1.4) and the Gårding-type inequality (2.13) in $L^p(0, \tau; W_0^{1,p}(\Omega))$, $\tau \in [0, T]$, it follows that

$$\begin{aligned}&\int_0^\tau \langle \dot{u}(t) - \dot{U}^h(t) + A(u(t)) - A(U^h(t)), u(t) - U^h(t)\rangle\, dt\\ &\geqq \tfrac{1}{2}|u(\tau) - U^h(\tau)|^2 - \tfrac{1}{2}|u_0 - U_0^h|^2 + \alpha_1 \int_0^\tau \|u(t) - U^h(t)\|^p\, dt\\ &\quad - \alpha_2(\mu)\left(\int_0^\tau \|u(t) - U^h(t)\|^p_{L^p(\Omega)}\, dt\right)^{p'/p}\end{aligned} \tag{5.12}$$

and, from equations (2.2) and (5.1), the following orthogonality condition holds:

$$\int_0^\tau \langle \dot{u}(t) - \dot{U}^h(t) + A(u(t)) - A(U^h(t)), Z(t)\rangle \, dt = 0 \quad \forall Z \in L^p(0, T; V_h). \tag{5.13}$$

Hence, introducing (5.13) into (5.12) and using the local Lipschitz continuity property (2.4), we obtain

$$\begin{aligned} &\tfrac{1}{2}|u(\tau) - U^h(\tau)|^2 + \alpha_1 \int_0^\tau \|u(t) - U^h(t)\|^p \, dt \\ &\leqq \tfrac{1}{2}|u_0 - U_0^h|^2 + \alpha_2(\mu)\|u - U^h\|^{p'}_{L^p(Q)} + \{C(\mu)|||u - U^h||| + |||\dot{u} - \dot{U}^h|||_*\}|||u - Z||| \\ &\qquad \forall \tau \in [0, T], \quad \forall Z \in L^p(0, T; V). \end{aligned} \tag{5.14}$$

Therefore, the approximation error estimates (5.10) and (5.11) are implied by (5.14). Note that the strong convergence of $U^h \to u$ in $L^p(Q)$ is a consequence of the first statement of (5.3), assumption (5.5) and the compact embedding of $\mathcal{W}$ into $L^p(Q)$. □

The potential case. As a final result, we shall establish that if the bounded, coercive locally Lipschitz continuous, Gårding-type operator A of (2.3), is potential in the following sense:

CIII. A is the gradient of some Gâteaux differentiable functional $J: V \to \mathbb{R}$, for which there is a constant $\tilde{\gamma} > 0$ such that

$$J(v) \geqq \tilde{\gamma}\|v\|^p \quad \forall v \in V, \tag{5.15}$$

then, for data

$$(f, u_0) \in L^2(Q) \times V, \tag{5.16}$$

$$U_0^h \to u_0 \quad \text{strongly in } V, \tag{5.17}$$

the Faedo–Galerkin sequence of approximations defined uniquely by (5.1) is such that

$$\begin{aligned} &\{U^h\}_{0<h\leqq 1} \quad \text{is bounded in } L^\infty(0, T; V), \\ &\{\dot{U}^h\}_{0<h\leqq 1} \quad \text{is bounded in } L^2(Q). \end{aligned} \tag{5.18}$$

Since $L^2(Q) \hookrightarrow \mathcal{V}'$, the second statement of (5.18) is stronger than (5.5) and, consequently, the results of Theorems 5.2 and 5.3 are true in this potential case.

We now prove this result and establish the corresponding regularity of limit functions.

THEOREM 5.4. *Let the operator A of* (2.3) *satisfy condition* CIII *and consider problems* (2.2) *and* (5.1) *with data* (5.16), (5.17). *Then the Faedo–Galerkin sequence of approximate solutions $\{U^h\}_{0<h\leqq 1}$ is bounded in the sense of* (5.18). *Furthermore, there is a subsequence of approximations, also denoted $\{U^h\}_{0<h\leqq 1}$, converging strongly to a solution $u \in \mathcal{W}$ of problem* (2.2) *in the sense of* (5.9), *such that, as $h \to 0^+$,*

$$\begin{aligned} &U^h \rightharpoonup u \quad \textit{weakly* in } L^\infty(0, T; V), \\ &\dot{U}^h \rightharpoonup \dot{u} \quad \textit{weakly in } L^2(Q). \end{aligned} \tag{5.19}$$

Proof. Let $\{U^h\}_{0<h\leqq 1}$ be the Faedo–Galerkin sequence defined uniquely by (5.1), (5.16), (5.17), which approximates problem (2.2) with data (5.16), and suppose also that condition CIII holds. Then, by replacing ψ_k by $\dot{U}^h$ in equation (5.1), integrating with respect to time from 0 to $\tau \in [0, T]$ and, then, observing that $dJ(U^h(t))/dt = \langle A(U^h(t)), \dot{U}^h(t)\rangle$ and that $(f(t), \dot{U}^h(t)) \leqq \frac{1}{2}|f(t)|^2 + \frac{1}{2}|\dot{U}^h(t)|^2$ for a.e. $t \in (0, T)$, we

obtain

$$(5.20)\qquad \tfrac{1}{2}\int_0^\tau |\dot{U}^h(t)|^2\,dt+\tilde{\gamma}\|U^h(\tau)\|^p \leqq J(U_0^h)+\tfrac{1}{2}\int_0^\tau |f(t)|^2\,dt \qquad \forall \tau\in[0,T].$$

But, from the boundedness of A as a map from $V\to V'$ (cf. remark 2.1),

$$J(U_0^h)=J_0+\int_0^1 \langle A(sU_0^h),\, U_0^h\rangle\,ds \leqq J_0+\int_0^1 \|A(sU_0^h)\|_*\,ds\,\|U_0^h\|\leqq \text{const.}$$

Therefore, (5.18) is true.

Next observe that from Theorem 5.3, there is a subsequence of approximations $\{U^h\}_{0<h\leqq 1}$ that converges strongly to a solution u of problem (2.2) in $\mathcal{V}\cap L^\infty(0,T;L^2(\Omega))$. Hence, $u^h\to u$ weakly in $\mathcal{V}(\hookrightarrow L^1(0,T;V)$ densely) and this together with the first statement of (5.18) is equivalent to the first statement of (5.19). Also $\{U^h\}_{0<h\leqq 1}$ is bounded in $\mathcal{U}(\hookrightarrow \mathcal{V}$ densely) and this with $U^h\to u$ weakly in $\mathcal{V}$ is necessary and sufficient for $U^h\to u$ weakly in $\mathcal{U}$ (cf. [11, § V. 1]). Then the second statement of (5.19) necessarily holds and this completes the proof of the theorem. □

Conclusions. For the nonlinear evolution problems considered here, we have shown that the existence conditions of coercivity and $\mathcal{W}$-pseudomonotonicity of $A:\mathcal{V}\to\mathcal{V}'$, are satisfied and that, under conditions (3.3) uniqueness of solutions is guaranteed. The elliptic regularization ideas discussed in § 4 provide a general framework for Galerkin approximations of coercive $\mathcal{W}$-pseudomonotone problems. We have established criteria for the existence and weak convergence of such approximations, as well as strong convergence whenever a nonlinear Gårding-type inequality of the form

$$[A(v)-A(w),\,v-w]\geqq \alpha_1\|v-w\|_{\mathcal{V}}-H(\mu,\|v-w\|_{L^p(0,T;X)}) \quad \forall v,\,w\in B_\mu(0)\subset\mathcal{V},$$

holds. Here $\alpha_1>0$, and X is a Banach space continuously embedded in H and in which v is compactly embedded. Also, if in addition, $A:V\to\mathcal{V}'$ is locally Lipschitz continuous, we have shown that error estimates for Galerkin approximations can be derived.

The Faedo–Galerkin method was considered as an alternative method for constructing approximate solutions. In these cases, coercivity, boundedness and demicontinuity of A from $V\to V'$ are sufficient conditions for existence, and local Lipschitz continuity from $V\to V'$ is a sufficient condition for uniqueness. As we have seen, the convergence of this method is a conditional property in the case that A is nonmonotone; the Faedo–Galerkin method is weakly convergent if: (i) the sequence of time derivatives of the approximate solutions is bounded in $\mathcal{V}'$; or (ii) if $A:\mathcal{V}\to\mathcal{V}'$ is $\mathcal{V}$-pseudomonotone. The convergence of the method is strong if: (iii) condition (i) holds and A is locally Lipschitz continuous and satisfies a nonlinear Gårding inequality of the type given above. Furthermore, in the case in which condition (iii) is satisfied, error estimates are derivable which are compatible with the interpolation theory of finite-elements in Sobolev spaces [7], [2].

This establishes condition (i) as a fundamental convergence condition for the Faedo–Galerkin method when applied to coercive $\mathcal{W}$-pseudomonotone parabolic problems. In particular, we have shown that this condition is satisfied whenever A is, in addition: continuous and potential from $V\to V'$, its potential is coercive, and the data $(f,u_0)\in\mathcal{H}\times V$. In this potential case, the convergence condition (i) holds in $\mathcal{H}\hookrightarrow\mathcal{V}'$; furthermore, the approximate solutions form a sequence bounded in

$L^\infty(0, T; V) \hookrightarrow \mathcal{V}$, and the regularity in time result "$(u, \partial u/\partial t) \in L^\infty(0, T; V) \times \mathcal{H}$" holds for the exact solutions of the problem.

REFERENCES

[1] J. P. AUBIN, *Un théorème de compacité*, C.R. Acad. Sci. Paris, 256 (1963), pp. 5042–5044.
[2] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[3] J. K. HALE, *Ordinary Differential Equations*, Wiley–Interscience, New York, 1969.
[4] J. L. LIONS, *Sur les espaces d'interpolation; dualité*, Math. Scand., 9 (1961), pp. 147–177.
[5] ———, *Sur certaines équations paraboliques non linéaires*, Bull. Soc. Math. France, 93 (1965), pp. 155–175.
[6] ———, *Quelques Méthodes de Resolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[7] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, Wiley–Interscience, New York, 1976.
[8] J. T. ODEN, *Existence theorems for a class of problems in nonlinear elasticity*, J. Math. Anal. Appl., 69 (1979), pp. 51–83.
[9] C. OLECH AND Z. OPIAL, *Sur une inégalité différentielle*, Ann. Polon. Math., 7 (1960), pp. 247–254.
[10] J. T. ODEN, C. T. REDDY AND N. KIKUCHI, *Qualitative analysis and finite element approximation of a class of nonmonotone nonlinear Dirichlet problems*, TICOM Rep. 78-15, The University of Texas at Austin, 1978.
[11] K. YOSIDA, *Functional Analysis*, 4th Ed., Springer-Verlag, New York, Heidelberg, Berlin, 1974.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0013 $01.00/0

A TURÁN INEQUALITY ARISING IN INFORMATION THEORY*

R. J. McELIECE†, B. REZNICK‡ AND J. B. SHEARER§

Abstract. We present a strengthened version of a polynomial inequality recently obtained by Davisson et al. in a paper on information theory. It is of a type originally obtained by Turán for the family of Legendre polynomials.

Let us define the polynomials $P_n(x)$ by $P_1(x) = 1$, and for $n \geqq 2$

$$P_n(x) := \prod_{k=1}^{n-1} \left(x + \frac{k}{n} \right). \tag{1}$$

In a recent paper on information theory [1], the authors required and proved the inequality $P_{n+m}(x) \leqq (x+1)P_n(x)P_m(x)$, for $x \geqq 0$, $n, m \geqq 1$. In this note we shall prove a stronger inequality, which we find to be of independent interest:

$$P_{n-1}(x)P_{n+1}(x) < P_n(x)^2 \quad \text{for all real } x, \quad \text{all } n \geqq 2. \tag{2}$$

An inequality of this kind is called a *Turán inequality*. Turán [2] showed that (2) holds for $-1 < x < 1$ if the P_n's are the Legendre polynomials, and Karlin and Szegö [3] proved (2) for several other families of orthogonal polynomials. However, our polynomials are not orthogonal, and do not appear to satisfy higher order determinant inequalities of the type considered in [3].

Before proving (2) we note certain consequences. If we define

$$F_{n,m}(x) = \frac{P_{n+m}(x)}{P_n(x)P_m(x)}, \tag{3}$$

it follows from (2) that for any fixed $x \geqq 0$ $F_{n,m}(x)$ is a strictly decreasing function of n and m. For, omitting the fixed argument x, (2) can be rewritten as $P_{n+1}/P_n > P_{n+2}/P_{n+1}$ for $n \geqq 1$. Thus by induction $P_{n+1}/P_n > P_{n+m+1}/P_{n+m}$ for all $n, m \geqq 1$. This can be rewritten as $P_{n+m}/P_n > P_{n+m+1}/P_{n+1}$, and so by another induction $P_{n+m}/P_n > P_{n'+m}/P_{n'}$ if $n' > n$. Dividing both sides by P_m we have (cf. (3)) $F_{n,m} > F_{n',m}$ for $n' > n$. Since $F_{n,m}$ is symmetric in n and m, this shows that $F_{n,m}(x)$ is strictly decreasing in n and m, as asserted. It follows that, for fixed $x \geqq 0$, $n, m \geqq 1$,

$$F_{n,m}(x) \leqq F_{1,1}(x) = x + \tfrac{1}{2}, \tag{4}$$

$$F_{n,m}(x) > \lim_{n',m' \to \infty} F_{n',m'}(x) = \sqrt{x(1+x)}. \tag{5}$$

(The limit calculation in (5) is an easy consequence of Stirling's formula.) The inequality from [1] noted above is equivalent to $F_{n,m}(x) \leqq x + 1$, and so inequality (2) is indeed stronger. Incidentally, (4) and (5) together show that $F_{n,m}(x)$ is very nearly independent

* Received by the editors June 23, 1980, and in final revised form February 6, 1981.

† Department of Mathematics and Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801. The work of this author was partially supported by the Joint Services Electronics Program under contract N00014-78-C-0424.

‡ Department of Mathematics, University of Illinois, Urbana, IL 61801. The work of this author was partially supported by the National Science Foundation under grant MCS-80-01542.

§ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139. The work of this author was partially performed while a consultant at Bell Laboratories, Murray Hill, NJ, and was also partially supported by the U.S. Office of Naval Research under contract N0014-76-C-0366.

of n and m for large fixed values of x. For example we have $2.449 < F_{n,m}(2) \leqq 2.500$ and $10.488 < F_{n,m}(10) \leqq 10.500$, for all n, m.

We now proceed to the proof of (2). First we note that $P_n(-x) = (-1)^{n+1}P_n(x-1)$. This implies that the function $P_n(x)^2 - P_{n-1}(x)P_{n+1}(x)$ is symmetric about $x = -\frac{1}{2}$, and can be written as a polynomial in $u = (x+\frac{1}{2})^2$. In fact, we have by direct calculation

$$\begin{aligned}
P_2^2 - P_1P_3 &= \tfrac{1}{36} \\
&= 0.028, \\
P_3^2 - P_2P_4 &= \tfrac{1}{144}u + \tfrac{1}{1296} \\
&= 0.0069u + 0.00077, \\
P_4^2 - P_3P_5 &= \tfrac{1}{360}u^2 + \tfrac{329}{1440000}u + \tfrac{1}{40000} \\
&= 0.0028u^2 + 0.00023u + 0.000025, \\
P_5^2 - P_4P_6 &= \tfrac{1}{720}u^3 + \tfrac{107}{3240000}u^2 + \tfrac{209}{16200000}u + \tfrac{81}{100000000} \\
&\doteq 0.0014u^3 + 0.000033u^2 + 0.000013u + 0.00000081, \text{ and} \\
P_6^2 - P_5P_7 &\doteq (7.94E-4)u^4 - (3.62E-5)u^3 \\
&\qquad + (7.46E-6)u^2 + (4.70E-7)u + 2.69E-8).
\end{aligned}$$

From this, (2) is immediate for $n \leqq 5$, since all coefficients are positive. However, the coefficient of u^{n-3} is negative for all $n \geqq 6$, and so deeper methods of proof are required. In any event, because of this symmetry, it suffices to prove (2) for $x \geqq -\frac{1}{2}$.

Now define the polynomials $Q_n(x)$ by

$$Q_n(x) = xP_n(x) = \prod_{k=0}^{n-1}\left(x + \frac{k}{n}\right).$$

We will show that

$$Q_{n-1}(x)Q_{n+1}(x) < Q_n(x)^2, \qquad x \geqq -\tfrac{1}{2}, \quad x \neq 0, \quad n \geqq 2. \tag{6}$$

For $x \neq 0$, (6) implies (2) immediately, and by continuity it implies $P_{n-1}(0)P_{n+1}(0) \leqq P_n(0)^2$. However, equality cannot hold, since a simple algebraic manipulation of (2) for $x = 0$ yields

$$\left(1 + \frac{1}{n}\right)^n > \left(1 + \frac{1}{n-1}\right)^{n-1}, \qquad n \geqq 2,$$

a well-known inequality. We now proceed to the proof of (6).

The functions $Q_n(x)$ are related to the gamma function by

$$Q_n(x) = \frac{\Gamma((x+1)n)}{n^n\Gamma(nx)}. \tag{7}$$

(The singularities on the right side of (7) are removable.) Hence we may extend the definition of $Q_n(x)$ to nonintegral values of n using the function

$$F(x, y) = \frac{\Gamma((x+1)y)}{y^y\Gamma(xy)}.$$

We will prove (6) by proving the generalization

$$F(x, y-1)F(x, y+1) < F(x, y)^2, \qquad x \geqq -\tfrac{1}{2}, \quad x \neq 0, \quad y \geqq 0. \tag{8}$$

We must handle positive and negative x separately.

Case 1. $x>0$. For any fixed $x>0$, $F(x, y)$ is continuous and positive for all $y>0$, and so to prove (8) it suffices to show

$$\frac{\partial^2}{\partial y^2}\log F(x, y)<0 \quad \text{for all } y>0. \tag{9}$$

We have

$$\log F(x, y)=\log\Gamma((x+1)y)-\log\Gamma(xy)-y\log y,$$

$$\frac{\partial^2}{\partial y^2}\log F(x, y)=(x+1)^2\psi'((x+1)y)-x^2\psi'(x)-\frac{1}{y},$$

where $\psi(z)$ is the digamma function $\Gamma'(z)/\Gamma(z)$ and $\psi'(z)$ is its derivative, the trigamma function. If now we define

$$f(z)=z^2\psi'(z)-z, \tag{10}$$

$$z_1=(x+1)y,\ z_2=xy, \tag{11}$$

we have

$$\frac{\partial^2}{\partial y^2}\log F(x, y)=\frac{1}{y^2}(f(z_1)-f(z_2)). \tag{12}$$

Thus Case 1 will be disposed of if we can show that $f(z)$ is a decreasing function of z, i.e., $f'(z)<0$ for $z>0$.

It is known [4, item 6.4.6] that

$$\psi'(z+1)-\psi'(z)=-z^{-2}, \quad \text{and so} \tag{13}$$

$$f(z)=z^2\psi(z)-z=1+z^2\psi'(z+1)-z, \quad \text{and so}$$

$$f'(z)=z^2\psi''(z+1)+2z\psi'(z+1)-1. \tag{14}$$

It is also known that, for all $z\neq 0, -1, -2, \cdots$,

$$\psi'(z)=\frac{1}{z}+\frac{1}{2z^2}+\frac{1}{6z^3}-\sum_{n=0}^{\infty}\frac{1}{6(z+n)^3(z+n+1)^3}, \tag{15}$$

$$\psi''(z)=-\frac{1}{z^2}-\frac{1}{z^3}-\frac{1}{2z^4}+\frac{1}{6z^6}-\sum_{n=0}^{\infty}\frac{(1+2(z+n))^3}{6z^6(1+z)^6}. \tag{16}$$

Equation (15) follows from (13), since the functions $L(z)=\psi'(z)-z^{-1}-(2z^2)^{-1}-(6z^3)^{-1}$ and $R(z)=\sum_{n=0}^{\infty}(6(z+n)^3(z+n-1)^3)^{-1}$ both satisfy $L(z+1)-L(z)=R(z+1)-R(z)=-(6z^3(z+1)^3)^{-1}$ for all z, and both approach 0 as $z\to\infty$. Equation (16) follows similarly, from the known recurrence $\psi''(z+1)-\psi''(z)=2z^{-3}$ [4, ibid.].

Since the summations in (15) and (16) are positive for $z>0$, the terms preceeding the sums are upper bounds on ψ' and ψ''. Substituting these bounds into the left side of (14), we get after some calculation

$$f'(z)\leqq -\frac{1}{6(z+1)^6}(z^4+6z^3+14z^2+16z+6)<0 \quad \text{if } z>0,$$

and this completes our proof of Case 1.

Case 2. $-\frac{1}{2}\leqq x<0$. For any such (fixed) x, $F(x, y)$ is continuous for all $y>0$, but changes sign at $y=k/|x|$, $k=1, 2, \cdots$. Notice that (8) is true trivially if $F(x, y-1)$ and $F(x, y+1)$ have opposite signs. On the other hand, if $F(x, y-1)$ and $F(x, y+1)$ have

the same sign, then $F(x, \cdot)$ cannot change signs (or be zero) in the interval $(y-1, y+1)$, since the distance between consecutive zeros of $F(x, \cdot)$ is $\geqq 2$ if $x \geqq -\frac{1}{2}$. Thus to prove (8) it suffices to replace the condition (9) with

$$\frac{\partial^2}{\partial y^2} \log |F(x, y)| < 0 \quad \text{for } y > 0.$$

This leads once again to (12), and so it is sufficient to show that

$$f(z_1) < f(z_2).$$

Since (cf. (11)) z_1 is positive and z_2 is negative, and since we have already shown that $f(z)$ is strictly decreasing for positive z, it will be sufficient to prove

$$f(z) \geqq f(0) \quad \text{for } z < 0. \tag{17}$$

It is known [4, item 6.4.10] that for $z \neq 0, -1, -2, \cdots$

$$\psi'(z) = \sum_{n=0}^{\infty} \frac{1}{(z+n)^2},$$

and so (cf. (10))

$$f(z) = 1 + \sum_{n=1}^{\infty} \frac{z^2}{(z+n)^2} - z. \tag{18}$$

From (18) we see immediately that $f(0) = 1$, and that $f(z) > 1$ if $z < 0$. This proves (17), and so completes our proof of Case 2.

REFERENCES

[1] L. D. DAVISSON, R. J. McELIECE, M. B. PURSLEY AND M. S. WALLACE, *Efficient universal noiseless source codes*, IEEE Trans. Inform. Theory, IT-27 (1981).

[2] G. SZEGÖ, *On an inequality of P. Turán concerning Legendre polynomials*, Bull. Amer. Math. Soc., 54 (1948), pp. 401–405.

[3] S. KARLIN AND G. SZEGÖ, *On certain determinants whose elements are orthogonal polynomials*, J. Analyse Math., 8 (1960/61), pp. 1–157.

[4] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0014 $01.00/0

CHARACTERIZATION OF POSITIVE QUADRATURE FORMULAS*

FRANZ PEHERSTORFER†

Abstract. We give a complete description of those numerical integration formulas based on n nodes which have positive weights and are exact for polynomials of degree equal or less than $2n-1-m$, where $0 \leqq m \leqq n$.

Let integers $n, m \in \mathbb{N}_0$, $m \leqq n$ and a nonnegative weight function w defined on $[-1, +1]$ be given. If there exist nodes $x_1, \cdots, x_n$, $-1 < x_1 < x_2 < \cdots < x_n < 1$ and weights $\lambda_1, \cdots, \lambda_n$ such that

$$\int_{-1}^{+1} p(x)w(x)\,dx = \sum_{i=1}^{n} \lambda_i p(x_i) \tag{1}$$

for all $p \in P_{2n-1-m}$ (where P_{2n-1-m} is the set of polynomials of degree at most $2n-1-m$), we say that (1) is a $(2n-1-m, n, w)$ quadrature formula (qf) based on $x_1, \cdots, x_n$ with weights $\lambda_1, \cdots, \lambda_n$. If all weights λ_i are positive we call (1) a positive $(2n-1-m, n, w)$ qf. This notation was introduced in [8]. Let us note that Gaussian quadrature is the unique positive $(2n-1, n, w)$ qf. Interpolatory quadrature formulas are $(n-1, n, w)$ qf.

Bernstein, see [1, p. 93], stated a necessary condition for the positivity of a $(2n-1-m, n, w)$ qf. To the best of our knowledge a corresponding sufficient condition is not known. A full characterization is only available for $(2n-3, n, w)$ qf. It was given by Michelli and Rivlin [8]. Positive $(2n-3, n, w)$ qf were first considered by Fejer [3].

In this paper we give a full description of positive $(2n-1-m, n, w)$ qf.

In order to state our results we need the following notation. Let U be the open unit disk $\{z \,|\, |z| < 1\}$ in the complex plane. As usual we call a function $f: \mathbb{C} \to \mathbb{C}$ a Carathéodory function (C-function) if f is analytic in U and $\operatorname{Re} f(z) > 0$ for $z \in U$.

Furthermore we denote by $q_n^*(z) = z^n \bar{q}(z^{-1}) = \bar{\gamma} \prod_{i=1}^{n} (1 - \bar{z}_i z)$ the reciprocal polynomial of $q_n(z) = \gamma \prod_{i=1}^{n} (z - z_i)$, $\gamma, z, z_i \in \mathbb{C}$. T_k denotes the Chebyshev polynomial of first kind of degree k.

THEOREM 1. *A $(2n-1-m, n, w)$ qf based on the nodes $x_1, \cdots, x_n$, $-1 < x_1 < x_2 < \cdots < x_n < 1$ is positive if and only if there exists a polynomial $q_{2n-1}(z) = \prod_{i=1}^{2n-1} (z - z_i)$, $z_i \in U$, with real coefficients, such that*

$$2^{-n+1} \operatorname{Re}\{z^{-(n-1)} q_{2n-1}(z)\} = \prod_{j=1}^{n} (x - x_j),$$

$x = \frac{1}{2}(z + 1/z)$, $z = e^{i\varphi}$, $\varphi \in [0, \pi]$, and

$$\frac{q_{2n-1}^*(z) - z q_{2n-1}(z)}{q_{2n-1}^*(z) + z q_{2n-1}(z)} = 1 + \sum_{k=1}^{2n-1-m} c_k z^k + O(z^{2n-m}) \quad \text{for } z \in U,$$

where $c_k = 2 \int_{-1}^{+1} T_k(x) w(x)\,dx / \int_{-1}^{+1} w(x)\,dx$ for $k = 1, \cdots, 2n-1-m$.

* Received by the editors October 9, 1980.
† Institut für Mathematik, Universität Linz, A-4040 Linz, Austria.

Proof. Necessity. Since the qf is a positive $(2n-1-m, n, w)$ qf it follows that

$$c_k = \frac{2\int_{-1}^{+1} T_k(x)w(x)\,dx}{\int_{-1}^{+1} w(x)\,dx} = \sum_{j=1}^{n} \lambda_j \cos k \arccos x_j$$

$$= \sum_{j=1}^{n} \lambda_j \cos k\varphi_j \quad \text{for } k = 0, \cdots, 2n-1-m,$$

where $\lambda_j \in \mathbb{R}^+$ for $j = 1, \cdots, n$ and $\varphi_j := \arccos x_j$.

Now let us put

$$c_k = \sum_{j=1}^{n} \lambda_j \cos k\varphi_j \quad \text{for } k = 2n-m, \cdots, 2n-1.$$

Then we obtain that

$$R(z) := \frac{1}{2}\sum_{j=1}^{n} \lambda_j \frac{1-z^2}{1-2z\cos\varphi_j + z^2} = 1 + \sum_{k=1}^{2n-1} c_k z^k + O(z^{2n}) \quad \text{for } z \in U.$$

Since R is a (degenerate) C-function with $R(0) = 1$, there exist (compare [10, pp. 229–231]) a polynomial $q_{2n-1}(z) = \varepsilon \prod_{i=1}^{2n-1}(z - z_i)$, $z_i \in U$, with real coefficients and a $\varepsilon \in \{\pm 1\}$ such that

$$\frac{q^*_{2n-1}(z) - \varepsilon z q_{2n-1}(z)}{q^*_{2n-1}(z) + \varepsilon z q_{2n-1}(z)} = \frac{1}{2}\sum_{j=1}^{n} \lambda_j \frac{1-z^2}{1-2z\cos\varphi_j + z^2} \quad \text{for } z \in U.$$

Because of $\varphi_j \in (0, \pi)$ for $j = 1, \cdots, n$, it follows that $q^*_{2n-1}(1) + \varepsilon q_{2n-1}(1) \neq 0$. Hence $\varepsilon = 1$. Observing that $(z = e^{i\varphi})$

$$\operatorname{Re}\{z^{-(n-1)} q_{2n-1}(z)\} = \frac{z^{-n}}{2}[z q_{2n-1}(z) + q^*_{2n-1}(z)]$$

$$= 2^{n-1}\prod_{j=1}^{n}\left(\frac{1+z^2}{2z} - \cos\varphi_j\right) = 2^{n-1}\prod_{j=1}^{n}(\cos\varphi - \cos\varphi_j)$$

one part of the theorem is proved.

Sufficiency. Using the facts that the roots of q_{2n-1} are real or complex conjugate and that $z q_{2n-1}(z) + q^*_{2n-1}(z) = \prod_{j=1}^{n}(z - e^{i\varphi_j})(z - e^{-i\varphi_j})$, where $\varphi_j = \arccos x_j$, we get that (see [10, p. 230])

$$\frac{q^*_{2n-1}(z) - z q_{2n-1}(z)}{q^*_{2n-1}(z) + z q_{2n-1}(z)} = -\sum_{j=1}^{n} \lambda'_j \left(\frac{z + e^{i\varphi_j}}{z - e^{i\varphi_j}} + \frac{z + e^{-i\varphi_j}}{z - e^{-i\varphi_j}}\right)$$

$$= 2\sum_{j=1}^{n} \lambda'_j \frac{1-z^2}{1-2z\cos\varphi_j + z^2} \quad \text{for } z \in U,$$

where $\lambda'_j \in \mathbb{R}^+$ for $j = 1, \cdots, n$.

Thus we obtain by putting $\lambda_j = 4\lambda'_j$, $j = 1, \cdots, n$, that

$$\frac{1}{2}\sum_{j=1}^{n} \lambda_j \frac{1-z^2}{1-2z\cos\varphi_j + z^2} = 1 + \sum_{k=1}^{2n-1-m} c_k z^k + O(z^{2n-m}) \quad \text{for } z \in U,$$

from which we conclude that

$$c_k = \frac{2\int_{-1}^{+1} T_k(x)w(x)\,dx}{\int_{-1}^{+1} w(x)\,dx} = \sum_{j=1}^{n} \lambda_j \cos k\varphi_j$$

$$= \sum_{j=1}^{n} \lambda_j \cos k \arccos x_j \quad \text{for } k = 0, \cdots, 2n-1-m.$$

COROLLARY 1. *A* $(2n-1-m, n, w)$ qf *based on the nodes* $x_1, \cdots, x_n$, $-1 < x_1 < \cdots < x_n < 1$, *is positive if and only if there exists a polynomial* $s_{2n-2}(z) = \prod_{j=1}^{n-1}(z-e^{i\theta_j})(z-e^{-i\theta_j})$, *where* $0 < \theta_1 < \theta_2 < \cdots < \theta_{n-1} < \pi$ *such that* $0 < \varphi_1 < \theta_1 < \varphi_2 < \cdots < \theta_{n-1} < \varphi_n < \pi$ *and*

$$\frac{(1-z^2)s_{2n-2}(z)}{r_{2n}(z)} = 1 + \sum_{k=1}^{2n-1-m} c_k z^k + O(z^{2n-m}) \quad \text{for } z \in U,$$

where

$$\varphi_j = \arccos x_j, \qquad r_{2n}(z) = \prod_{j=1}^{n}(z-e^{i\varphi_j})(z-e^{-i\varphi_j})$$

and

$$c_k = \frac{2\int_{-1}^{+1} T_k(x)w(x)\,dx}{\int_{-1}^{+1} w(x)\,dx} \quad \text{for } k = 1, \cdots, 2n-1-m.$$

Proof. Necessity. In view of the proof of Theorem 1, there exists a polynomial $q_{2n}(z) = z\prod_{i=1}^{2n-1}(z-z_i)$, $z_i \in U$, with real coefficients, such that

$$\frac{q_{2n}^*(z)-q_{2n}(z)}{q_{2n}^*(z)+q_{2n}(z)} = \frac{1}{2}\sum_{j=1}^{n} \lambda_j \frac{1-z^2}{1-2z\cos\varphi_j+z^2}$$

$$= 1 + \sum_{k=1}^{2n-1-m} c_k z^k + O(z^{2n-m}),$$

where (see [10]) $\lambda_j = -2(q_{2n}^* - q_{2n})(z_j)/(z_j(q_{2n}^*+q_{2n})'(z_j)) > 0$, $z_j = e^{i\varphi_j}$, $\varphi_j \in (0, \pi)$. Setting $r_{2n}(z) = (q_{2n}^*+q_{2n})(z)$ and $s_{2n-2}(z) = (q_{2n}^*-q_{2n})(z)/(1-z^2)$ we obtain, since $z(d/dz)r_{2n}(z) = -i(d/d\varphi)r_{2n}(e^{i\varphi})$ for $z = e^{i\varphi}$, that

$$\lambda_j = \frac{2(1-e^{i2\varphi_j})s_{2n-2}(e^{i\varphi_j})}{i\dfrac{d}{d\varphi}r_{2n}(e^{i\varphi_j})}$$

$$= -4\sin\varphi_j \frac{e^{-i(n-1)\varphi_j}s_{2n-2}(e^{i\varphi_j})}{\dfrac{d}{d\varphi}e^{-in\varphi_j}r_{2n}(e^{i\varphi_j})} > 0,$$

from which the assertion follows.

Sufficiency. Partial fraction expansion gives

$$\frac{(1-z^2)s_{2n-2}(z)}{r_{2n}(z)} = -\frac{1}{4}\sum_{j=1}^{n}\lambda_j\left(\frac{z+e^{i\varphi_j}}{z-e^{i\varphi_j}}+\frac{z+e^{-i\varphi_j}}{z-e^{-i\varphi_j}}\right)$$

$$= \frac{1}{2}\sum_{j=1}^{n}\lambda_j\frac{1-z^2}{1-2z\cos\varphi_j+z^2} \quad \text{for } z\in U,$$

where $\lambda_j = (-2(1-z_j^2)s_{2n-2}(z_j))/(z_j(d/dz)r_{2n}(z_j))$, $z_j = e^{i\varphi_j}$.

Since $e^{-in\varphi}r_{2n}(e^{i\varphi})$ and $e^{-i(n-1)\varphi}s_{2n-2}(e^{i\varphi})$ are greater than zero at $\varphi = 0$, it follows that

$$\operatorname{sgn}\frac{d}{d\varphi}e^{-in\varphi_j}r_{2n}(e^{i\varphi_j}) = (-1)^j = -\operatorname{sgn}e^{-i(n-1)\varphi_j}s_{2n-2}(e^{i\varphi_j})$$

for $j = 1, \cdots, n$. Hence $\lambda_j \in \mathbb{R}^+$, $j \in \{1, \cdots, n\}$. Thus there exists a polynomial $q_{2n}(z) = z\prod_{i=1}^{2n-1}(z-z_i)$, $z_i \in U$, with real coefficients, such that

$$\frac{q_{2n}^*(z)-q_{2n}(z)}{q_{2n}^*(z)+q_{2n}(z)} = \frac{1}{2}\sum_{j=1}^{n}\lambda_j\frac{(1-z^2)}{1-2z\cos\varphi_j+z^2} = \frac{(1-z^2)s_{2n-2}(z)}{r_{2n}(z)}.$$

From Theorem 1 the assertion follows.

Notation. Let $P_l(z) = z^l + \cdots$ be that polynomial which is orthogonal on the unit circle with respect to the weight function $f(\varphi) := w(\cos\varphi)|\sin\varphi|$ for $\varphi \in [0, 2\pi)$, i.e.,

$$\int_0^{2\pi} e^{-ij\varphi}P_l(e^{i\varphi})f(\varphi)\,d\varphi = 0 \quad \text{for } j = 0, \cdots, l-1.$$

For the determination of the orthogonal polynomial P_l see [4] and [12]. The particular cases $w(x) = 1/\sqrt{1-x^2}$ and $w(x) = 1$ are of special interest. In the first case it follows immediately that $P_l(z) = z^l$, $l \in \mathbb{N}_0$. In the second (Legendre) case P_l is given by the recurrence formula $P_{l+1}(z) = zP_l(z) - a_lP_l^*(z)$, where $P_0(z) = 1$, $a_l = 0$ for l even and $a_l = -1/(2l+1)$ for l odd. Another representation of P_l can be found in [12, p. 295].

The main result of this paper is the following.

THEOREM 2. *A* $(2n-1-m, n, w)$ qf *based on the nodes* $x_1, \cdots, x_n$, $-1 < x_1 < x_2 < \cdots < x_n < 1$ *is positive if and only if there exists a polynomial* $q_m(z) = \prod_{i=1}^{m}(z-z_i)$, $z_i \in U$, *with real coefficients, such that*

$$2^{-n+1}\operatorname{Re}\{z^{-(n-1)}q_m(z)P_{2n-1-m}(z)\} = \prod_{j=1}^{n}(x-x_j),$$

$x = \frac{1}{2}(z+1/z)$, $z = e^{i\varphi}$, $\varphi \in [0, \pi]$.

Proof. Necessity. In view of Theorem 1, there exists a polynomial $p_{2n-1}(z) = \prod_{i=1}^{2n-1}(z-z_i)$, $z_i \in U$, with real coefficients, such that

$$\frac{p_{2n-1}^*(z)-zp_{2n-1}(z)}{p_{2n-1}^*(z)+zp_{2n-1}(z)} = 1 + \sum_{k=1}^{2n-1-m}c_kz^k + O(z^{2n-m}),$$

where

$$c_k = \frac{2\int_{-1}^{+1} T_k(x)w(x)\,dx}{\int_{-1}^{+1} w(x)\,dx} = \frac{2\int_0^{2\pi} e^{-ik\varphi}w(\cos\varphi)|\sin\varphi|\,d\varphi}{\int_0^{2\pi} w(\cos\varphi)|\sin\varphi|\,d\varphi}.$$

Now let Ω_l denote the polynomial of second kind with respect to the weight function $f(\varphi) = w(\cos\varphi)|\sin\varphi|$ (see [4, p. 6]).

According to [5, Theorem IX'] (see also [4, Theorem 18.2]) there exists a function $\phi\colon \mathbb{C}\to\mathbb{C}$ which is analytic in U and satisfies the inequality $|\phi(z)|<1$ for $z\in U$, such that

$$\frac{p^*_{2n-1}(z)-zp_{2n-1}(z)}{p^*_{2n-1}(z)+zp_{2n-1}(z)} = -\frac{z\Omega_{2n-1-m}(z)\phi(z)-\Omega^*_{2n-m-1}(z)}{zP_{2n-1-m}(z)\phi(z)+P^*_{2n-m-1}(z)}. \tag{1}$$

Isolating ϕ from this equality we find that ϕ can be represented by

$$\phi(z) = \frac{q_l(z)}{q_l^*(z)}, \quad \text{where } q_l(z) = \prod_{i=1}^{l}(z-z_i), \quad z_i\in U,$$

q_l has real coefficients, and $l\geqq m$.

Let us assume that $l>m$. Then it follows from (1) that $-z\Omega_{2n-1-m}\phi+\Omega^*_{2n-m-1}$ and $zP_{2n-1-m}\phi+P^*_{2n-1-m}$ have $(l-m)$ common zeros on $|z|=1$, which implies that

$$\frac{P_{2n-m-1}}{P^*_{2n-m-1}}+\frac{\Omega_{2n-m-1}}{\Omega^*_{2n-m-1}}$$

has $(l-m)$ zeros on $|z|=1$. But this is impossible, since (see [4, p. 7])

$$P_{2n-m-1}\Omega^*_{2n-m-1}+\Omega_{2n-m-1}P^*_{2n-m-1} = Kz^{2n-1-m},$$

where $K\in\mathbb{R}^+$.

With the aid of (1) we obtain that

$$2^{n-1}\prod_{j=1}^{n}(x-x_j) = \operatorname{Re}\{z^{-n+1}p_{2n-1}(z)\}$$

$$= \frac{z^{-n}}{2}(zq_m(z)P_{2n-1-m}(z)+q^*_m(z)P^*_{2n-m-1}(z))$$

$$= \operatorname{Re}\{z^{-(n-1)}q_m(z)P_{2n-1-m}(z)\} \quad \text{for } x=\frac{1}{2}\left(z+\frac{1}{z}\right),\ z=e^{i\varphi},\ \varphi\in[0,\pi].$$

Sufficiency. Let $\phi(z)=q_m(z)/q^*_m(z)$. Then, see [5, Theorem IX']

$$H(z) := -\frac{z\Omega_{2n-1-m}\phi-\Omega^*_{2n-1-m}}{zP_{2n-1-m}\phi+P^*_{2n-1-m}}$$

is a (degenerate) C-function with initial coefficients $1, c_1, \cdots, c_{2n-1-m}$, where

$$c_k = \frac{\int_0^{2\pi} e^{-ik\varphi} w(\cos\varphi)|\sin\varphi| d\varphi}{\int_0^{2\pi} w(\cos\varphi)|\sin\varphi|\, d\varphi}.$$

Thus there exists a polynomial $p_{2n-1}(z) = \prod_{i=1}^{2n-1}(z - z_i)$, $z_i \in U$, with real coefficients such that

$$\frac{p^*_{2n-1}(z) - z p_{2n-1}(z)}{p^*_{2n-1}(z) + z p_{2n-1}(z)} = H(z) = 1 + \sum_{k=1}^{2n-1-m} c_k z^k + O(z^{2n-m})$$

for $z \in U$.

The assertion follows now from Theorem 1.

The following corollary gives us a simple characterization of positive $(2n-1-m, n, (1-x^2)^{-1/2})$ qf. It completes the results of the author [9].

COROLLARY 2. *A* $(2n-1-m, n, (1-x^2)^{-1/2})$ qf *based on the nodes* $x_1, \cdots, x_n$, $-1 < x_1 < x_2 < \cdots < x_n < 1$ *is positive if and only if there exists a polynomial* $\sum_{k=0}^{m} a_k z^k$, $(a_0, \cdots, a_{m-1}) \in \mathbb{R}^m$, $a_m = 1$, *which has all zeros in* U, *such that*

$$2^{-n+1} \sum_{k=0}^{m} a_k T_{n-m+k}(x) = \prod_{j=1}^{n} (x - x_j), \qquad x \in [-1, +1].$$

Proof. Since $P_{2n-1-m}(z) = z^{2n-1-m}$ the assertion follows immediately from Theorem 2.

Notation. Let U_k, $k \in \mathbb{N}_0$, denote the Chebyshev polynomial of the second kind.

COROLLARY 3. *Let* $n, m \in \mathbb{N}_0$, $m \leqq n$, *and assume that* $\sum_{k=0}^{m} a_k z^k$, $(a_0, \cdots, a_{m-1}) \in \mathbb{R}^m$, $a_m = 1$, *has all zeros in* U. *The* qf *based on nodes which are the zeros of the polynomial* $\sum_{k=0}^{m} a_k U_{n-m+k}$ *is a positive* $(2n-1-m, n, \sqrt{1-x^2})$ qf.

Proof. Let $q_m(z) = \sum_{k=0}^{m} a_k z^k$, $s_{2n-2}(z) = -(z^{2n-m} q_m(z) - q^*_m(z))/(1-z^2)$ and $r_{2n}(z) = z^{2n+2-m} q_m(z) - q^*_m(z)/(z^2-1)$. Further let $\psi_l(\varphi) = \arg z^l q_m(z)/q^*_m(z)$ for $z = e^{i\varphi}$, $\varphi \in [0, 2\pi)$. Then it follows that $s_{2n-2}(r_{2n})$ has a zero at $z_j = e^{i\varphi_j}$, $\varphi_j \in (0, 2\pi)$, if there is a $\nu \in Z$, such that $\psi_{2n-m}(\varphi_j) = 2\nu\pi$ ($\psi_{2n+2-m}(\varphi_j) = 2\nu\pi$). Since ψ_l increases from 0 to $2(m+l)\pi$, as φ varies from 0 to 2π, we deduce that $s_{2n-2}(r_{2n})$ has all zeros on the unit disk and $n-1(n)$ zeros on the upper unit disk. Observing that $(z = e^{i\varphi})$

$$\psi_{2n-m}(\varphi) + 2\pi > \psi_{2n-m}(\varphi) + 2\arg z = \psi_{2n+2-m}(\varphi) > \psi_{2n-m}(\varphi)$$

for $\varphi \in (0, 2\pi)$, it follows that the zeros of r_{2n} and $(1-z^2)s_{2n-2}$ separate each other. Taking into consideration the facts that

$$c_k = \frac{2\int_{-1}^{+1} T_k(x)\sqrt{1-x^2}\, dx}{\int_{-1}^{+1} \sqrt{1-x^2}\, dx} = \begin{cases} -1 & \text{for } k = 2 \\ 0 & \text{for } k \in \mathbb{N}\setminus\{2\} \end{cases}$$

and

$$\frac{(1-z^2)s_{2n-2}(z)}{r_{2n}(z)} = 1 - z^2 + O(z^{2n-m}) \quad \text{for } z \in U,$$

the assertion follows from Corollary 1 and the relation

$$z^{-n}r_{2n}(z)=\frac{\operatorname{Im}\{z^{n+1-m}q_m(z)\}}{\sin\varphi}=\sum_{k=0}^{m}a_kU_{n-m+k}(x),$$

$x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi)$.

As a simple consequence of Corollary 2 and Corollary 3 we obtain a result of Micchelli [7].

COROLLARY 4. *Let $n, m\in\mathbb{N}_0$, $m\leqq n$, and assume that $a_m>a_{m-1}>\cdots>a_0>0$. The* qf *based on nodes which are the zeros of the polynomial $\sum_{k=0}^{m}a_kT_{n-m+k}$ ($\sum_{k=0}^{m}a_kU_{n-m+k}$) is a positive $(2n-1-m, n, (1-x^2)^{-1/2})((2n-1-m, n, \sqrt{1-x^2}))$* qf.

Proof. Since $a_m>\cdots>a_0>0$ it follows from Enestrom's theorem, see [6, p. 42], that $\sum_{k=0}^{m}a_kz^k$ has all zeros in U. According to Corollary 2 and Corollary 3 the assertion is proved.

With the help of a result of Shohat [11] we also get a complete characterization of those positive quadrature formulas, having nodes outside of the interval $(-1, +1)$.

THEOREM 3. *Let $n, m, s_1, s_2\in\mathbb{N}_0$, $s=s_1+s_2$ and suppose that $m+s\leqq n$. A* qf *with nodes $x_1<x_2<\cdots<x_{s_1}\leqq-1<x_{s_1+1}<\cdots<x_{n-s_2}<1\leqq x_{n-s_2+1}<\cdots<x_n$ is exact for polynomials of degree $\leqq 2n-1-(m+s)$ and has positive weights $\lambda_{s_1+1}, \cdots, \lambda_{n-s_2}$ if and only if*

$$2^{-(n-s)+1}\operatorname{Re}\{z^{-(n-s-1)}q_m(z)P'_{2(n-s)-1-m}(z)\}=\prod_{j=1}^{n-s}(x-x_{s_1+j}),$$

$x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$, where P'_l is that polynomial of degree l, which is orthogonal on the circumference with respect to the weight function

$$w(\cos\varphi)\prod_{j=1}^{s_1}(\cos\varphi-x_j)\prod_{k=1}^{s_2}(\cos\varphi-x_{n-s_2+k})|\sin\varphi|.$$

Let us note that $\operatorname{sgn}\lambda_j=(-1)^{s_1-j}$ for $j=1,\cdots,s_1$ and $\operatorname{sgn}\lambda_j=(-1)^{j-(n-s_2+1)}$ for $j=n-s_2+1,\cdots,n$.

Proof. Follows immediately from Theorem 2 and Shohat [11, pp. 468–470].

By Theorem 4.1 and Theorem 5.1 of [2] and Theorem 3 we are able to characterize that nonnegative algebraic polynomial of degree n, which has the least deviation from zero in the L^1-norm, among all nonnegative algebraic polynomials of degree n with leading coefficients $A_n, \cdots, A_{n-k}$. For example we obtain

COROLLARY 5. *Suppose $n, m\in\mathbb{N}_0$, $[(2n-m+1)/2]\leqq n\leqq 2n-m$.*

(a) *Let $R_{2n}(x)=\prod_{j=1}^{n}(x-x_j)^2=\sum_{k=2n-m}^{2n}A_kx^k+\cdots$, where $x_j\in(-1,+1)$ for $j=1,\cdots,n$. If $\int_{-1}^{+1}p(x)w(x)\,dx\geqq\int_{-1}^{+1}R_{2n}(x)w(x)\,dx$ for all nonnegative algebraic polynomials of degree $2n$ with leading coefficients $A_{2n}, \cdots, A_{2n-m}$, then there exists a polynomial $q_m(z)=\prod_{i=1}^{m}(z-z_i)$, $z_i\in U$, with real coefficients, such that*

$$2^{-n+1}\operatorname{Re}\{z^{-(n-1)}q_m(z)P_{2n-1-m}(z)\}=\prod_{j=1}^{n}(x-x_j),$$

$x=\frac{1}{2}(z+1/z)$, $z=e^{i\varphi}$, $\varphi\in[0,\pi]$.

(b) *Suppose that $q_m(z)=\prod_{i=1}^{m}(z-z_i)$, $z_i\in U$, has real coefficients and let $R_{2n}(x)=[\operatorname{Re}\{z^{-(n-1)}q_m(z)P_{2n-1-m}(z)\}]^2=\sum_{k=2n-m}^{2n}A_kx^k+\cdots$. Then $\int_{-1}^{+1}p(x)w(x)\,dx\geqq\int_{-1}^{+1}R_{2n}(x)w(x)\,dx$ for all nonnegative algebraic polynomials p of degree $2n$ with leading coefficients $A_{2n}, \cdots, A_{2n-m}$.*

REFERENCES

[1] H. BRASS, *Quadraturverfahren*, Vanderhoeck u. Ruprecht, Göttingen-Zürich, 1977.
[2] R. DE VORE, *One sided approximation of functions*, J. Approx. Theory, 1 (1968), pp. 11–25.
[3] L. FEJER, *Mechanische Quadraturen mit positiven Cotesschen Zahlen*, Math. Z., 37 (1933), pp. 287–309.
[4] J. GERONIMUS, *Polynomials orthogonal on a circle and their applications*, Zapiski Naucno-Issled. Inst. Mat. Meh. Har'kov Mat. Obsc., (4) 19 (1948), pp. 35–120. Amer. Math. Soc. Transl., (1) 3 (1962), pp. 1–78.
[5] ———, *On polynomials orthogonal on a circle, on the trigonometric moment problem and on the associated functions of Carathéodory's and Schur's types*, Mat. Sb. N.S., 15 (1944), pp. 99–130. [In Russian].
[6] N. LEVINSON AND R. M. REDHEFFER, *Complex Variables*, Holden-Day, San Francisco-Cambridge, 1970.
[7] C. A. MICCHELLI, *Some positive Cotes numbers for the Chebyshev weight function*, Aequationes Math., 21 (1980), pp. 105–109.
[8] C. A. MICCHELLI AND T. J. RIVLIN, *Numerical integration rules near gaussian quadrature*, Israel J. Math., 16 (1973), pp. 267–299.
[9] F. PEHERSTORFER, *On an extremal problem for nonnegative trigonometric polynomials and the characterization of positive quadrature formulas with Chebyshev weight function*, Acta. Math. Acad. Sci. Hungary, to appear.
[10] J. SCHUR, *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind*, J. Reine Angew. Math. 147 (1917), pp. 205–232.
[11] J. SHOHAT, *On mechanical quadratures, in particular, with positive coefficients*, Trans. Amer. Math. Soc., 42 (1937), pp. 461–496.
[12] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloquium Publications, 1967.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0006 $01,00/0

SOME TRANSFORMATIONS OF BASIC HYPERGEOMETRIC FUNCTIONS. PART I*

A. VERMA† AND V. K. JAIN‡

Abstract. q-analogues of certain formulas of Gasper are obtained. Orthogonality relations of q-Hahn and q-Racah polynomials are also obtained by a different method than a method of Askey and Wilson [SIAM J. Math. Anal., 10 (1979), pp. 1008–1016]. A bilinear generating function for q-Hahn polynomials is also discussed.

1. Recently, Gasper [4] has shown that the well-known formula of Watson [13] expressing the product of two terminating hypergeometric functions in terms of an F_4 function

$$
\begin{aligned}
&{}_2F_1\left[\begin{matrix}-n,n+a\\ c\end{matrix};z\right]{}_2F_1\left[\begin{matrix}-n,n+a\\ c\end{matrix};Z\right]\\
&\quad=\frac{(-1)^n(1+a-c)_n}{(c)_n}F_4[-n,\,n+a;\,c,\,1+a-c;\,zZ,\,(1-z)(1-Z)],
\end{aligned}
\tag{1.1}
$$

admits a generalization of the form

$$
\begin{aligned}
&{}_3F_2\left[\begin{matrix}-n,\,n+a,\,b\\ c,\,d\end{matrix};\right]{}_3F_2\left[\begin{matrix}-n,\,n+a,\,e\\ c,\,f\end{matrix};\right]\\
&\quad=\frac{(-1)^n(1+a-c)_n}{(c)_n}F\left[\begin{matrix}-n,\,n+a:\,b,\,e;\,d-b,\,f-e\\ d,\,f:\,c;\,1+a-c\end{matrix};\right].
\end{aligned}
\tag{1.2}
$$

Gasper [5] also showed that Bailey's formula [2]

$$
{}_2F_1\left[\begin{matrix}a,\,b\\ c\end{matrix};z\right]{}_2F_1\left[\begin{matrix}a,\,b\\ 1+a+b-c\end{matrix};Z\right]=F_4[a,\,b;\,c,\,1+a+b-c;\,z(1-Z),\,Z(1-z)],
\tag{1.3}
$$

which is valid inside the simply connected region surrounding $z=0$ and $Z=0$ with $|z(1-Z)|^{1/2}+|Z(1-z)|^{1/2}<1$, has a discrete analogue of the form

$$
{}_3F_2\left[\begin{matrix}a,\,b,\,-x\\ c,\,d\end{matrix};\right]{}_3F_2\left[\begin{matrix}a,\,b,\,-y\\ 1+a+b-c,\,e\end{matrix};\right]=F\left[\begin{matrix}a,\,b:\,-x,\,y+e;\,-y,\,x+d\\ d,\,e:\,c;\,1+a+b-c\end{matrix};\right],
\tag{1.4}
$$

where $x,\,y=0,\,1,\,2,\,\cdots$. Formula (1.3) is a special case of the general transformation

$$
\begin{aligned}
&F_4[a,\,b;\,c,\,c';\,z(1-Z),\,Z(1-z)]\\
&\quad=\sum_{r=0}^{\infty}\frac{(a)_r(b)_r(1+a+b-c-c')_r z^r Z^r}{(1)_r(c)_r(c')_r}{}_2F_1\left[\begin{matrix}a+r,\,b+r\\ c+r\end{matrix};z\right]{}_2F_1\left[\begin{matrix}a+r,\,b+r\\ c'+r\end{matrix};Z\right]
\end{aligned}
\tag{1.5}
$$

which is due to Burchnall and Chaundy [3]. Gasper [5] has also shown that (1.5) admits the generalization

$$
\begin{aligned}
F\left[\begin{matrix}a,\,b:\,-x,\,y+e;\,-y,\,x+d\\ d,\,e:\,c;\,c'\end{matrix};\right]&=\sum_{r=0}^{\min(x,y)}\frac{(a)_r(b)_r(1+a+b-c-c')_r(-x)_r(-y)_r}{(1)_r(c)_r(c')_r(d)_r(e)_r}\\
&\quad\cdot{}_3F_2\left[\begin{matrix}a+r,\,b+r,\,-x+r\\ c+r,\,d+r\end{matrix};\right]{}_3F_2\left[\begin{matrix}a+r,\,b+r,\,-y+r\\ c'+r,\,e+r\end{matrix};\right],
\end{aligned}
\tag{1.6}
$$

* Received by the editors October 22, 1979, and in revised form September 15, 1980.

† Mathematics Department, Roorkee University, Roorkee (U.P.), India.
‡ Mathematics Department, Bareilly College, Bareilly (U.P.), India.

where $x, y = 0, 1, 2, \cdots$. Clearly (1.6) reduces to (1.4) when $c' = 1 + a + b - c$.

The extensions (1.2), (1.4) and (1.6) were obtained by Gasper by employing an extension of the proof used by Watson for proving (1.1) and that by Bailey for proving (1.3). In this note we begin by showing that these results of Gasper follow from the classical ones by a use of a beta-function transform. Other results of Gasper viz. [5(1.7), (3.1), (4.1), (4.4)], can also be obtained by using the beta-function transform of known results. It may be remarked that Jackson in his paper [8] had remarked that the q-analogues of (1.5) and its inverse relations will be investigated in a subsequent paper. However, they were never discussed by him. In this paper we discuss in § 3 the q-analogues of (1.1)–(1.6). We also discuss in § 4 and § 5 the q-analogues of some interesting results of Gasper [5], [6]. Some interesting properties of q-Hahn and q-Racah polynomials are obtained in § 6 as applications of the transformation theory of basic hypergeometric functions discussed in § 3.

2. Notation and definitions. Let

$$[a; q]_n = (1-a)(1-aq)\cdots(1-aq^{n-1}), \quad [a; q]_0 = 1, \quad [a; q]_\infty = \prod_{j=0}^{\infty}(1-aq^j)$$

and the generalized basic hypergeometric series is defined as

$$ {}_{p+1}\phi_{p+r}\left[\begin{matrix} a_1, \cdots, a_{p+1} \\ b_1, \cdots, b_{p+r} \end{matrix}; q; x\right] $$

$$ = \sum_{n=0}^{\infty} \frac{[a_1; q]_n \cdots [a_{p+1}; q]_n x^n (-1)^{nr} q^{rn(n-1)/2}}{[q; q]_n [b_1; q]_n \cdots [b_{p+r}; q]_n}, \qquad |q| < 1 $$

which is convergent for all values of x when $r = 1, 2, \cdots$ and for $|x| < 1$ when $r = 0$.

As usual, we define the basic double hypergeometric series as:

$$ \phi\left[\begin{matrix} (a_r): (b_s); (c_t) \\ (d_u): (e_v); (f_w) \end{matrix}; x, y; q\right] $$

$$ = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{[(a_r); q]_{m+n}[(b_s); q]_m[(c_t); q]_n x^m y^n}{[q; q]_m[q; q]_n[(d_u); q]_{m+n}[(e_v); q]_m} \cdot \frac{1}{[(f_w); q]_n}. $$

3. To deduce (1.2) from (1.1), multiply both sides of (1.1) by $z^{b-1}Z^{e-1}(1-z)^{d-b-1}(1-Z)^{f-e-1}$ and integrate both sides with respect to z and Z from 0 to 1, to get (1.2) under the restrictions $Rl(b) > 0$, $Rl(d-b) > 0$, $Rl(e) > 0$, $Rl(f-e) > 0$. These conditions have arisen due to the method followed and can be removed by analytic continuation. The need for using the restrictions and then removing them may be avoided by using the equivalent contour integral taken on the path $(1+, 0+, 1-, 0-)$ in the z, Z-planes. For details see MacRobert [9, pp. 259, 364].

To deduce (1.4) from (1.3) we have to proceed in a slightly roundabout way. We begin by assuming that a is a negative integer so that (1.3) holds for all z, Z. Now multiply both sides of (1.3) by $z^{-x-1}(1-z)^{d+x-1}Z^{-y-1}Z^{-y-1}(1-Z)^{e+y-1}$ and integrate both sides with respect to z and Z from 0 to 1, to get (1.4) under the restriction that a is a negative integer and $Rl(-x) > 0$, $Rl(x+d) > 0$, $Rl(-y) > 0$, $Rl(e+y) > 0$. Under the condition that a is a negative integer both sides of the resulting expression are polynomials, and hence the formula is valid for all complex values of x, y, d and e. Then restricting x, y to $0, 1, \cdots$, we find that the resulting formula is a polynomial in a, hence is valid for all complex values of a.

(1.6) may be deduced from (1.5) by an argument similar to the one used in deducing (1.4) from (1.3); hence the details are omitted.

Next, we prove a q-analogue of (1.2) in the form

$$
\begin{aligned}
&{}_3\phi_2\left[\begin{matrix} q^{-n}, aq^n, e \\ c, f \end{matrix}; q; \frac{cf}{ae}\right] {}_3\phi_2\left[\begin{matrix} q^{-n}, aq^n, b \\ c, d \end{matrix}; q; \frac{cd}{ab}\right] \\
&= \frac{\left[\frac{aq}{c}; q\right]_n}{[c; q]_n}\left(-\frac{c}{a}\right)^n q^{-n(n+1)/2} \phi\left[\begin{matrix} q^{-n}, aq^n : b, e; \frac{f}{e}, \frac{d}{b} \\ d, f : c; \frac{aq}{c} \end{matrix}; \frac{cdf}{abe}, q; q\right]
\end{aligned} \tag{3.1}
$$

Setting $d = zb, f = eZ$ in (3.1) and letting $b, e \to 0$ gives a q-analogue of (1.1) in the form

$$
\begin{aligned}
&{}_2\phi_1\left[\begin{matrix} q^{-n}, aq^n \\ c \end{matrix}; q; \frac{cz}{a}\right] {}_2\phi_1\left[\begin{matrix} q^{-n}, aq^n \\ c \end{matrix}; q; \frac{cZ}{a}\right] \\
&= \left(-\frac{c}{a}\right)^n q^{-n(n+1)/2} \frac{\left[\frac{aq}{c}; q\right]_n}{[c; q]_n} \phi\left[\begin{matrix} q^{-n}, aq^n; -1; z, Z \\ -1 : c; \frac{aq}{c} \end{matrix}; \frac{czZ}{a}, q; q\right].
\end{aligned} \tag{3.2}
$$

To prove (3.1) we follow a procedure similar to the one used by Gasper [4] for proving (1.2). Sears [10, eq. 8.3] has shown that

$$
{}_4\phi_3\left[\begin{matrix} a, b, c, q^{-n} \\ d, e, f \end{matrix}; q; q\right] = \frac{\left[\frac{e}{c}; q\right]_n \left[\frac{de}{ab}; q\right]_n}{[e; q]_n \left[\frac{de}{abc}; q\right]_n} {}_4\phi_3\left[\begin{matrix} \frac{d}{a}, \frac{d}{b}, c, q^{-n} \\ d, \frac{cq^{1-n}}{e}, \frac{cq^{1-n}}{f} \end{matrix}; q; q\right], \tag{3.3}
$$

where $abcq^{1-n} = def$. Substituting for f and letting $n \to \infty$, (3.3) gives

$$
{}_3\phi_2\left[\begin{matrix} a, b, c \\ d, e \end{matrix}; q; \frac{de}{abc}\right] = \frac{\left[\frac{e}{c}; q\right]_\infty \left[\frac{de}{ab}; q\right]_\infty}{[e, q]_\infty \left[\frac{de}{abc}; q\right]_\infty} {}_3\phi_2\left[\begin{matrix} \frac{d}{a}, \frac{d}{b}, c \\ d, \frac{de}{ab} \end{matrix}; q; \frac{e}{c}\right]. \tag{3.4}
$$

Denote the left-hand side of (3.1) by S and transform the second ${}_3\phi_2$ series by (3.4). This gives

$$
\begin{aligned}
S = &\frac{\left[\frac{d}{b}; q\right]_\infty \left[\frac{cd}{a}; q\right]_\infty}{[d; q]_\infty \left[\frac{cd}{ab}; q\right]_\infty} \sum_{j=0}^{n} \sum_{r=0}^{\infty} \frac{[q^{-n}; q]_j [aq^n; q]_j [e; q]_j [cq^n; q]_r}{[q; q]_j [c; q]_j [f; q]_j [q; q]_r [c; q]_r} \\
&\cdot \frac{\left[\frac{c}{a} q^{-n}; q\right]_r [b; q]_r \left(\frac{d}{b}\right)^r \left(\frac{cf}{ae}\right)^j}{\left[\frac{cd}{a}; q\right]_r} \\
= &\frac{\left[\frac{d}{b}; q\right]_\infty \left[\frac{cd}{a}; q\right]_\infty \left[\frac{aq}{c}; q\right]_n}{[d; q]_\infty \left[\frac{cd}{ab}; q\right]_\infty [c; q]_n} \sum_{j=0}^{n} \sum_{r=0}^{\infty} \frac{[q^{-n}; q]_j [aq^n; q]_j [e; q]_j}{[q; q]_j [c; q]_j [f; q]_j} \\
&\cdot \frac{[cq^r; q]_n [b; q]_r \left(-\frac{cd}{ab}\right)^r \left(\frac{cf}{ae}\right)^j q^{r(r-1-2n)/2}}{[q; q]_r \left[\frac{cd}{a}; q\right]_r \left[\frac{aq}{c}; q\right]_{n-r}}.
\end{aligned} \tag{3.5}
$$

Now by the q-analogue of Vandermonde's theorem [11, eq. 3.3.27] we know that

$$\frac{[cq^r;q]_j}{[c;q]_j[q;q]_j[q;q]_r}=\sum_{s\geqq 0}\frac{q^{s(s-r-j)+rj}}{[q;q]_s[c;q]_s[q;q]_{j-s}[q;q]_{r-s}} \tag{3.6}$$

and

$$\frac{\left[\frac{q^{1-n-r}}{c};q\right]_{n-j}a^{n-j}q^{n^2-j^2}}{\left[\frac{a}{c}q^{1-r+j};q\right]_{n-j}}=\sum_{p=j-s}^{n-s}\frac{[q^{-n+j};q]_{s-j+p}[aq^{n+j};q]_{s-j+p}q^{s-j+p}}{[q;q]_{s-j+p}\left[\frac{a}{c}q^{1-r+j};q\right]_{s-j+p}}. \tag{3.7}$$

Using (3.6) and (3.7) in (3.5), setting $j=s+k$, $r=s+l$, we have after some simplification,

$$\begin{aligned}S=&\left(-\frac{c}{a}\right)^n q^{-n(n+1)/2}\frac{\left[\frac{d}{b};q\right]_\infty\left[\frac{cd}{a};q\right]_\infty\left[\frac{aq}{c};q\right]_n}{[d;q]_\infty\left[\frac{cd}{ab};q\right]_\infty[c;q]_n}\\&\cdot\sum_{s=0}^{n}\sum_{p=0}^{n-s}\frac{[q^{-n};q]_{s+p}[aq^n;q]_{s+p}}{[q;q]_s[q;q]_p},\\&\cdot\frac{[b;q]_s[e;q]_s\left(\frac{cdf}{abe}\right)^s q^p}{\left[\frac{cd}{a};q\right]_s[c;q]_s\left[\frac{aq}{c};q\right]_p[f;q]_s}\\&\cdot{}_2\phi_1\left[\begin{matrix}eq^s,q^{-p}\\fq^s\end{matrix};q;\frac{f}{e}q^p\right]{}_2\phi_1\left[\begin{matrix}bq^s,\frac{c}{a}q^{-p}\\\frac{cd}{a}q^s\end{matrix};q;\frac{d}{b}q^p\right].\end{aligned}$$

Summing the two inner ${}_2\phi_1$ series by the q-analogue of Gauss's theorem [11, eq. 3.3.2.6], we get (2.1) on simplification.

Lastly, we prove the q-analogue of the formula (1.6) in the form:

$$\begin{aligned}&\phi\left[\begin{matrix}a,b:q^{-x},eq^y;q^{-y},dq^x\\d,e:c;c'\end{matrix};\frac{cd}{ab}q^{x-y},q;q\right]\\&=\sum_{r=0}^{\min(x,y)}\frac{[a;q]_r[b;q]_r}{[q;q]_r[c;q]_r}\\&\qquad\cdot\frac{\left[\frac{abq}{cc'};q\right]_r[q^{-x};q]_r[q^{-y};q]_r(cc'd)^rq^{rx}}{[c';q]_r[d;q]_r[e;q]_r(ab)^r}\\&\qquad\cdot{}_3\phi_2\left[\begin{matrix}q^{-x+r},aq^r,bq^r\\cq^r,dq^r\end{matrix};q;\frac{cd}{ab}q^{x-r}\right]{}_3\phi_2\left[\begin{matrix}q^{-y+r},aq^r,bq^r\\c'q^r,eq^r\end{matrix};q;q\right]\end{aligned} \tag{3.8}$$

where $x, y = 0, 1, 2, \cdots$. On setting $c' = abq/c$ in (3.8), we get a q-analogue of (1.4) in the form

$$
\begin{aligned}
&{}_3\phi_2\left[\begin{matrix} a, b, q^{-x} \\ c, d \end{matrix}; q; \frac{cd}{ab}q^x\right] {}_3\phi_2\left[\begin{matrix} a, b, q^{-y} \\ e, \frac{abq}{c} \end{matrix}; q; q\right] \\
&\qquad = \phi\left[\begin{matrix} a, b: q^{-x}, eq^y; q^{-y}, dq^x \\ d, e: c; \frac{abq}{c} \end{matrix}; \frac{cd}{ab}q^{x-y}, q; q\right].
\end{aligned} \tag{3.9}
$$

Proof of (3.8). Let

$$
\begin{aligned}
S &= \frac{[d;q]_x[e;q]_y}{b^y\left[\frac{d}{a};q\right]_x\left[\frac{e}{b};q\right]_y}\phi\left[\begin{matrix} a, b: q^{-x}, eq^y; q^{-y}, dq^x \\ d, e: c; c' \end{matrix}; \frac{cd}{ab}q^{x-y}, q; q\right] \\
&= \sum_{r,s\geqq 0} \frac{[a;q]_{r+s}[b;q]_{r+s}[q^{-x};q]_r[q^{-y};q]_s(-1)^{r+s}c^r q^{r(r+1)/2-s(s+1)/2-rs+2s}}{[q;q]_r[q;q]_s[c;q]_r[c';q]_s\left[\frac{a}{d}q^{1-x};q\right]_r\left[\frac{b}{e}q^{1-y};q\right]_s b^r e^s} \\
&\qquad \cdot {}_2\phi_1\left[\begin{matrix} q^{-x+r}, aq^{r+s} \\ \frac{a}{d}q^{1+r-x} \end{matrix}; q; q\right] {}_2\phi_1\left[\begin{matrix} q^{-y+s}, bq^{r+s} \\ \frac{b}{e}q^{1+s-y} \end{matrix}; q; \frac{q^{1-r-s}}{e}\right] \\
&= \sum_{r,s\geqq 0}\sum_{j,k\geqq 0} \frac{[a;q]_{r+s+j}[b;q]_{r+s+k}[q^{-x};q]_{r+j}[q^{-y};q]_{s+k}(-1)^{r+s}c^r}{[q;q]_r[q;q]_s[q;q]_j[q;q]_k[c;q]_r[c';q]_s} \\
&\qquad \cdot \frac{q^{r(r+1)/2-s(s+1)/2-(r+s-1)k-rs+2s+j}}{\left[\frac{a}{d}q^{1-x};q\right]_{r+j}\left[\frac{b}{e}q^{1-y};q\right]_{s+k} b^r e^{s+k}}.
\end{aligned}
$$

Setting $r+j=m$ and $s+k=n$ and simplifying, we get

$$
\begin{aligned}
S &= \sum_{m,n\geqq 0} \frac{[q^{-x};q]_m[q^{-y};q]_n[a;q]_m[b;q]_n q^{m+n}}{[q;q]_m[q;q]_n\left[\frac{a}{d}q^{1-x};q\right]_m\left[\frac{b}{e}q^{1-y};q\right]_n e^n} \\
&\qquad \cdot {}_2\phi_1\left[\begin{matrix} bq^n, q^{-m} \\ c \end{matrix}; q; \frac{cq^{m-n}}{b}\right] {}_2\phi_1\left[\begin{matrix} aq^m, q^{-n} \\ c' \end{matrix}; q; q\right] \\
&= \sum_{m,n\geqq 0} \frac{[q^{-x};q]_m[q^{-y};q]_n[a;q]_m[b;q]_n\left[\frac{c}{b}q^{-n};q\right]_m\left[\frac{c'}{a}q^{-m};q\right]_n\left(\frac{a}{e}\right)^n q^{m+n+mn}}{[q;q]_m[q;q]_n\left[\frac{a}{d}q^{1-x};q\right]_m\left[\frac{b}{e}q^{1-y};q\right]_n[c;q]_m[c';q]_n}.
\end{aligned} \tag{3.10}
$$

By the q-analogue of Saalschütz's theorem [11 eq. 3.3.2.2], we have

$$
{}_3\phi_2\left[\begin{matrix} \frac{abq}{cc'}, q^{-m}, q^{-n} \\ \frac{b}{c}q^{1-m}, \frac{a}{c'}q^{1-n} \end{matrix}; q; q\right] = \frac{\left[\frac{c'}{a}q^{-m};q\right]_n\left[\frac{c}{b}q^{-n};q\right]_m q^{mn}}{\left[\frac{c}{b};q\right]_m\left[\frac{c'}{a};q\right]_n}. \tag{3.11}
$$

Substituting (3.11) in (3.10), we have

$$S=\sum_{m,n\geq 0}\frac{[q^{-x};q]_m[q^{-y};q]_n[a;q]_m[b;q]_n\left[\frac{c}{b};q\right]_m\left[\frac{c'}{a};q\right]_n q^{m+n}}{[q;q]_m[q;q]_n\left[\frac{a}{d}q^{1-x};q\right]_m\left[\frac{b}{e}q^{1-y};q\right]_n[c;q]_m[c';q]_n}\left(\frac{a}{e}\right)^n$$

$$\cdot\sum_{r=0}^{\min(m,n)}\frac{\left[\frac{abq}{cc'};q\right]_r[q^{-m};q]_r[q^{-n};q]_r q^r}{[q;q]_r\left[\frac{b}{c}q^{1-m};q\right]_r\left[\frac{a}{c'}q^{1-n};q\right]_r}.$$

Finally, setting $m=r+s$, $n=r+t$ and changing the order of summation, we have also after some simplification

$$S=\sum_{r=0}^{\min(x,y)}\frac{[q^{-x};q]_r[q^{-y};q]_r[a;q]_r[b;q]_r\left[\frac{abq}{cc'};q\right]_r(cc'q)^r}{[q;q]_r\left[\frac{a}{d}q^{1-x};q\right]_r\left[\frac{b}{e}q^{1-y};q\right]_r[c;q]_r[c';q]_r(be)^r}$$

(3.12)

$$\cdot\,{}_3\phi_2\left[\begin{matrix}q^{-x+r}, aq^r, \frac{c}{b}\\ \frac{a}{d}q^{1-x+r}, cq^r\end{matrix};q;q\right]{}_3\phi_2\left[\begin{matrix}q^{-y+r}, bq^r, \frac{c'}{a}\\ \frac{b}{e}q^{1-y+r}, c'q^r\end{matrix};q;\frac{aq}{e}\right].$$

The proof of (3.8) is completed by transforming the two ${}_3\phi_2$ series in (3.12) by the transformation

$$(3.13)\qquad {}_3\phi_2\left[\begin{matrix}b, c, q^{-n}\\ d, e\end{matrix};q;\frac{de}{bc}q^n\right]=\frac{\left[\frac{e}{c};q\right]}{[e;q]_n}{}_3\phi_2\left[\begin{matrix}\frac{d}{b}, c, q^{-n}\\ d, \frac{c}{e}q^{1-n}\end{matrix};q;q\right],$$

which is obtained from (3.3) by letting a and $f\to\infty$. To transform the first of the ${}_3\phi_2$ in (3.12) let $b\to bq^r$, $c\to aq^r$, $d\to cq^r$, $e\to dq^r$, $n=x-r$ in (3.13), and to transform the second ${}_3\phi_2$ in (3.12) let $b\to c'/a$, $c\to bq^r$, $d\to c'q^r$, $e\to(b/e)q^{1-y+r}$, $n=y-r$ in (3.13).

4. In this section we prove a q-analogue of a formula due to Gasper [5 eq. 1.7] in the form:

$$\phi\left[\begin{matrix}a, b: q^{-x}, eq^y; q^{-y}, dq^x\\ d, e: c; b\end{matrix};\frac{cd}{ab}q^{x-y}, q;q\right]$$

(4.1)

$$=\frac{a^y\left[\frac{d}{a};q\right]_x\left[\frac{e}{a};q\right]_y}{[d;q]_x[e;q]_y}\phi\left[\begin{matrix}a, q^{-x}: \frac{aq}{c}, q^{-y}; \frac{c}{b}\\ c, \frac{a}{d}q^{1-x}: \frac{a}{e}q^{1-y}; -1\end{matrix};\frac{cq}{e}, q;q\right].$$

To prove (4.1), rewrite (3.10) in the form

$$\frac{[d;q]_x[e;q]_y}{\left[\frac{d}{a};q\right]_x\left[\frac{e}{b};q\right]_y}\phi\left[\begin{matrix}a, b: q^{-x}, eq^{y}; q^{-y}, dq^{x}\\ d, e: c; b\end{matrix}; \frac{cd}{ab}q^{x-y}, q; q\right]$$

$$= b^{y}\sum_{m=0}^{x}\frac{[q^{-x};q]_m[a;q]_m\left[\frac{c}{b};q\right]_m q^{m}}{[q;q]_m\left[\frac{a}{d}q^{1-x};q\right]_m[c;q]_m}\,{}_3\phi_2\left[\begin{matrix}q^{-y}, \frac{b}{a}q^{-m}, \frac{bq}{c}\\ \frac{b}{e}q^{1-y}, \frac{b}{c}q^{1-m}\end{matrix}; q; \frac{aq}{e}\right]$$

$$= \frac{a^{y}\left[\frac{e}{a};q\right]_y}{\left[\frac{e}{b};q\right]_y}\sum_{m=0}^{x}\frac{[q^{-x};q]_m[a;q]_m\left[\frac{c}{b};q\right]_m q^{m}}{[q;q]_m\left[\frac{a}{d}q^{1-x};q\right]_m[c;q]_m}\,{}_3\phi_2\left[\begin{matrix}\frac{aq}{c}, q^{-m}, q^{-y}\\ \frac{b}{c}q^{1-m}, \frac{a}{e}q^{1-y}\end{matrix}; q; \frac{bq}{e}\right],$$

(by transforming the ${}_3\phi_2$ inside the summation by using (3.4))

$$= \frac{a^{y}\left[\frac{e}{a};q\right]_y}{\left[\frac{e}{b};q\right]_y}$$

$$\cdot\sum_{m=0}^{x}\sum_{r=0}^{m}\frac{[q^{-x};q]_m[a;q]_m\left[\frac{c}{b};q\right]_m\left[\frac{aq}{c};q\right]_r[q^{-m};q]_r[q^{-y};q]_r q^{m+r}}{[q;q]_m[q;q]_r[c;q]_m\left[\frac{a}{d}q^{1-x};q\right]_m\left[\frac{b}{c}q^{1-m};q\right]_r\left[\frac{a}{e}q^{1-y};q\right]_r}\left(\frac{b}{e}\right)^{r}.$$

Set $m = j + r$, change the order of summation and simplify to get (4.1).

Setting $c = b$ in (4.1), we have the interesting result

$$\phi\left[\begin{matrix}a, b: q^{-x}, eq^{y}; q^{-y}, dq^{x}\\ d, e: b; b\end{matrix}; \frac{d}{a}q^{x-y}, q; q\right]$$

$$= \frac{a^{y}\left[\frac{d}{a};q\right]_x\left[\frac{e}{a};q\right]_y}{[d;q_x][e;q]_y}\,{}_4\phi_3\left[\begin{matrix}a, \frac{aq}{b}, q^{-x}, q^{-y}\\ b, \frac{a}{d}q^{1-x}, \frac{a}{e}q^{1-y}\end{matrix}; q; \frac{b}{e}q\right]. \tag{4.2}$$

Another interesting special case of (4.1) is obtained by setting $c = a$:

$$\phi\left[\begin{matrix}a, b: q^{-x}, eq^{y}; q^{-y}, dq^{x}\\ d, e: a; b\end{matrix}; \frac{d}{b}q^{x-y}, q; q\right]$$

$$= \frac{a^{y}\left[\frac{d}{b};q\right]_x\left[\frac{e}{a};q\right]_y}{[d;q]_x[e;q]_y}\sum_{r=0}^{\min(x,y)}\frac{[q^{-x};q]_r[q^{-y};q]_r b^{r}q^{r}}{\left[\frac{a}{e}q^{1-y};q\right]_r\left[\frac{b}{d}q^{1-x};q\right]_r e^{r}}. \tag{4.3}$$

(4.2) and (4.3) are q-analogues of results proved earlier by Gasper [5].

It might be of interest to point out that the inverse of the relation (1.5) given by Burchnall and Chaundy [3, eq. 55] has been extended by Gasper [5] to the inverse of the relation (1.6) by following the method used by Burchnall and Chaundy [3]. Similarly

the inverse of formula (3.8) can be written in the form

$$
{}_3\phi_2\left[\begin{matrix}a, b, q^{-x}\\ c, d\end{matrix}; q; \frac{cd}{ab}q^x\right]{}_3\phi_2\left[\begin{matrix}a, b, q^{-y}\\ c', e\end{matrix}; q; q\right]
$$

$$
= \sum_{r=0}^{\min(x,y)} \frac{[a;q]_r[b;q]_r\left[\frac{cc'}{abq};q\right]_r[q^{-x};q]_r[q^{-y};q]_r d^r q^{r(1+x)}}{[q;q]_r[c;q]_r[c';q]_r[d;q]_r[e;q]_r}
$$

$$
\cdot \phi\left[\begin{matrix}aq^r, bq^r: q^{-x+r}, eq^y; q^{-y+r}, dq^x\\ dq^r, eq^r: cq^r; c'q^r\end{matrix}; \frac{cd}{ab}q^{x-y}, q; q\right],
$$

where $x, y = 0, 1, 2, \cdots$.

5. The main result of this section is

$$
{}_3\phi_2\left[\begin{matrix}a, b, q^{-x}\\ c, d\end{matrix}; q; q\right]
$$

$$
= \frac{\mu^x\left[\frac{c}{\mu};q\right]_x}{[c;q]_x} \sum_{y=0}^{x} \frac{[q^{-x};q]_y\left[\frac{\lambda d}{ab};q\right]_y[\mu;q]_y}{[q;q]_y[d;q]_y\left[\frac{\mu}{c}q^{1-x};q\right]_y}\left(\frac{abq}{c\lambda}\right)^y \tag{5.1}
$$

$$
\cdot {}_3\phi_2\left[\begin{matrix}q^{-y}, \frac{\lambda}{a}, \frac{\lambda}{b}\\ \mu, \frac{\lambda d}{ab}\end{matrix}; q; q\right]{}_3\phi_2\left[\begin{matrix}q^{y-x}, \frac{\lambda}{\mu}, \frac{ab}{\lambda}\\ dq^y, \frac{c}{\mu}\end{matrix}; q; q\right],
$$

which is a q-analogue of a result due to Gasper [6, 26]. Gasper used his result to deduce a discrete Dirichlet-Mehler formula for the Hahn polynomials

$$
{}_3F_2\left[\begin{matrix}-k, 1+\alpha+\beta+k, -x\\ 1+\alpha, -N\end{matrix}; \right],
$$

where N is a nonnegative integer and $x = 0, 1, 2, \cdots, N$. To prove (5.1), we require the following expansion formula:

$$
{}_3\phi_2\left[\begin{matrix}q^{-x}, a, b\\ c, d\end{matrix}; q; z\right]
$$

$$
= \frac{\lambda^x\left[\frac{c}{\lambda}: q\right]_x}{[c;q]_x} \sum_{j=0}^{x} \frac{[\lambda;q]_j[q^{-x};q]_j q^j}{[q;q]_j\left[\frac{\lambda}{c}q^{1-x};q\right]_j c^j}{}_3\phi_2\left[\begin{matrix}q^{-j}, a, b\\ \lambda, d\end{matrix}; q; z\right]. \tag{5.2}
$$

To prove (5.2), substitute the series definition for ${}_3\phi_2$ in the right-hand side of (5.2), change the order of summation and sum the inner ${}_2\phi_1$ by the q-analogue of Gauss' theorem [11, eq. 3.3.2.6].

Now to prove (5.1), set $z = q$ in (5.2) and transform the inner ${}_3\phi_2$ on the right-hand side by the transformation

$$(5.3)\qquad {}_3\phi_2\left[\begin{matrix} q^{-x}, a, b \\ d, e \end{matrix}; q; q\right] = \frac{\left[\frac{de}{ab}; q\right]_x}{[e; q]_x}\left(\frac{ab}{d}\right)^x {}_3\phi_2\left[\begin{matrix} q^{-x}\frac{d}{a}, \frac{d}{b} \\ d, \frac{de}{ab} \end{matrix}; q; q\right],$$

(which is obtained from (3.3) by letting c and $f \to 0$) to obtain

$$\begin{aligned} &{}_3\phi_2\left[\begin{matrix} q^{-x}, a, b \\ c, d \end{matrix}; q; q\right] \\ &= \frac{\lambda^x\left[\frac{c}{\lambda}; q\right]_x}{[c; q]_x}\sum_{j=0}^{x}\frac{[\lambda; q]_j[q^{-x}; q]_j\left[\frac{\lambda d}{ab}; q\right]_j}{[q; q]_j[d; q]_j\left[\frac{\lambda}{c}q^{1-x}; q\right]_j}\left(\frac{abq}{c\lambda}\right)^j \\ &\qquad \cdot {}_3\phi_2\left[\begin{matrix} q^{-j}, \frac{\lambda}{a}, \frac{\lambda}{b} \\ \lambda, \frac{\lambda d}{ab} \end{matrix}; q; q\right] \\ &= \frac{\lambda^x\left[\frac{c}{\lambda}; q\right]_x}{[c; q]_x}\sum_{j=0}^{x}\sum_{y=0}^{j}\frac{[q^{-x}; q]_j[q^{-j}; q]_y[\mu; q]_y\left[\frac{\lambda}{\mu}; q\right]_{j-y}\left[\frac{\lambda d}{ab}; q\right]_j}{[q; q]_j[q; q]_y[d; q]_j\left[\frac{\lambda}{c}q^{1-x}; q\right]_j} \\ &\qquad \cdot \frac{(ab\mu)^j(-1)^y q^{-y(y+1)/2+jy+y+j}}{(c\lambda)^j\mu^y}\,{}_3\phi_2\left[\begin{matrix} q^{-y}, \frac{\lambda}{a}, \frac{\lambda}{b} \\ \mu, \frac{\lambda d}{ab} \end{matrix}; q; q\right] \\ &= \frac{\lambda^x\left[\frac{c}{\lambda}; q\right]_x}{[c; q]_x}\sum_{y=0}^{x}\frac{[q^{-x}; q]_y[\mu; q]_y\left[\frac{\lambda d}{ab}; q\right]_y}{[q; q]_y[d; q]_y\left[\frac{\lambda}{c}q^{1-x}; q\right]_y}\left(\frac{abq}{c\lambda}\right)^y \\ &\qquad \cdot {}_3\phi_2\left[\begin{matrix} q^{-y}, \frac{\lambda}{a}, \frac{\lambda}{b} \\ \mu, \frac{\lambda d}{ab} \end{matrix}; q; q\right]{}_3\phi_2\left[\begin{matrix} q^{-x+y}, \frac{\lambda}{\mu}, \frac{\lambda d}{ab}q^y \\ dq^y, \frac{\lambda}{c}q^{1-x+y} \end{matrix}; q; \frac{ab\mu q}{c\lambda}\right], \end{aligned} \tag{5.4}$$

which is a q-analogue of a result of Gasper [6, 25].

Transforming the second of the two ${}_3\phi_2$ on the right-hand side of (5.4) by the transformation (3.13), we get (5.1).

6. q-Hahn polynomials of degree n are defined [7] as

$$(6.1)\qquad Q_n(x; \alpha, \beta, d; q) = {}_3\phi_2\left[\begin{matrix} q^{-n}, q^{1+\alpha+\beta+n}, q^{-x} \\ q^{1+\alpha}, q^{-d} \end{matrix}; q; q\right].$$

These polynomials satisfy the orthogonality relation (see also [1, p. 11])

$$\sum_{x=0}^{N} j(x)Q_n(x;\alpha,\beta,N;p)Q_m(x;\alpha,\beta,N;p)=\begin{cases}0 & \text{if } m\neq n,\\ \dfrac{1}{h(n)} & \text{if } m=n,\end{cases} \tag{6.2}$$

for $n, m = 0, 1, \cdots, N$, where $p = q^{-1}$,

$$j(x)=\frac{[p;p]_N[p^{1-\alpha};p]_x[p^{1+\beta};p]_{N-x}}{[p;p]_x[p;p]_{N-x}[p^{2+\alpha+\beta};p]_N}p^{(1+\alpha)(N-x)}$$

and

$$h(n)=\frac{(-1)^n[p^{-N};p]_n[p^{1+\alpha};p]_n[p^{1+\alpha+\beta};p]_n(1-p^{1+\alpha+\beta+2n})p^{-n(n+1)/2-n\alpha+nN}}{[p;p]_n[p^{1+\beta};p]_n[p^{2+\alpha+\beta+N};p]_n(1-p^{1+\alpha+\beta})}$$

To prove (6.2) rewrite the left-hand side of (6.2) (with $pq = 1$) as

$$\begin{aligned}
&\sum_{x=0}^{N}\frac{[q;q]_N[q^{1+\beta};q]_{N-x}[q^{1+\alpha};q]_x q^{(1+\beta)x}}{[q;q]_x[q;q]_{N-x}[q^{2+\alpha+\beta};q]_N}\,{}_3\phi_2\left[\begin{matrix}q^{-n},q^{1+\alpha+\beta+n},q^{-x}\\ q^{1+\alpha},q^{-N}\end{matrix};q;q^{x-N-\beta}\right]\\
&\qquad\cdot{}_3\phi_2\left[\begin{matrix}q^{-m},q^{1+\alpha+\beta+m},q^{-x}\\ q^{1+\alpha},q^{-N}\end{matrix};q;q^{x-N-\beta}\right]\\
&=\sum_{x=0}^{N}\frac{[q;q]_N[q^{1+\beta};q]_{N-x}[q^{1+\alpha};q]_x[q^{-N-m-1-\alpha-\beta};q]_m q^{(1+\beta)x}}{[q;q]_x[q;q]_{N-x}[q^{2+\alpha+\beta};q]_N[q^{-N};q]_m}\\
&\qquad\cdot{}_3\phi_2\left[\begin{matrix}q^{-n},q^{1+\alpha+\beta+n},q^{-x}\\ q^{1+\alpha},q^{-N}\end{matrix};q;q^{x-N-\beta}\right]{}_3\phi_2\left[\begin{matrix}q^{-m},q^{1+\alpha+\beta+m},q^{1+\alpha+x}\\ q^{1+\alpha},q^{2+\alpha+\beta+N}\end{matrix};q;q\right]\\
&=\frac{[q^{-N-1-m-\alpha-\beta};q]_m}{[q^{2+\alpha+\beta};q]_N[q^{-N};q]_m}\\
&\qquad\cdot\sum_{r,s\geq 0}\frac{[q^{-n};q]_r[q^{1+\alpha+\beta+n};q]_r[q^{-m};q]_s[q^{1+\alpha+\beta+m};q]_s}{[q;q]_r[q;q]_s[q^{1+\alpha};q]_s}\\
&\qquad\cdot\frac{[q^{1+\beta};q]_{N-r}[q^{1+\alpha+r};q]_s q^{r+s}}{[q^{2+\alpha+\beta+N};q]_s}\,{}_2\phi_1\left[\begin{matrix}q^{-N+r},q^{1+\alpha+r+s}\\ q^{-\beta-N+r}\end{matrix};q;q\right]\\
&=\frac{[q^{-N-1-m-\alpha-\beta};q]_m}{[q^{-N};q]_m}\sum_{r=0}^{n}\frac{[q^{-n};q]_r[q^{1+\alpha+\beta+n};q]_r q^r}{[q;q]_r[q^{2+\alpha+\beta};q]_r}\\
&\qquad\cdot{}_3\phi_2\left[\begin{matrix}q^{-m},q^{1+\alpha+\beta+m},q^{1+\alpha+\gamma}\\ q^{1+\alpha},q^{2+\alpha+\beta+r}\end{matrix};q;q\right].
\end{aligned}$$

Summing the inner ${}_3\phi_2$ by the q-analogue of Saalschütz's theorem, we readily get (6.2) (in view of $pq = 1$). Orthogonality relation for the q-Hahn polynomials could be easily completed by showing that $Q_n(x;\alpha,\beta,N;p)$ is orthogonal to a polynomial of each lower degree. The specific polynomials can be chosen so that they could be attached to the weight function, but there may be some interest in a proof that gives the orthogonality directly.

Next, we obtain a bilinear generating function for q-Hahn polynomials in the form (q-analogue of a result of Gasper [4, 3.3])

$$
\begin{aligned}
&S_z(x, y; \alpha, \beta, N, M; p) \\
&\quad = \sum_{n=0}^{z} \frac{[p^{-z/2}; p^{1/2}]_n[-p^{1+(\alpha+\beta+N)/2}; p^{1/2}]_n p^{n(z-N)/2} h(n)}{[p^{-N/2}; p^{1/2}]_n[-p^{1+1/2(\alpha+\beta+z)}; p^{1/2}]n} \\
&\qquad \cdot Q_n(x; \alpha, \beta, N; p) \cdot Q_n(y; \alpha, \beta, M; p) \\
&\quad = \frac{[p^{(1-N)/2}; p^{1/2}]_z[p^{N/2}; p^{1/2}]_z[-p^{1+(\alpha+\beta)/2}; p^{1/2}]_z}{[p^{-N/2}; p^{1/2}]_z[p^{1+(\alpha+\beta+N)/2}; p^{1/2}]_z[-p^{1/2}; p^{1/2}]_z} \cdot \\
&\qquad \cdot \phi\left[\begin{matrix} p^{-z}, p^{(2+\alpha+\beta)/2}, p^{(3+\alpha+\beta)/2} : p^{-x}, p^{-y}; p^{x-N}, p^{y-M} \\ p^{(1-N-z)/2}, p^{(2-N-z)/2}, p^{-M} : p^{1+\alpha}; p^{1+\beta} \end{matrix}; p, p^{-\alpha-x-y}; p\right].
\end{aligned} \tag{6.3}
$$

Proof of (6.3). Consider

$$
\begin{aligned}
&S_z(x, y; \alpha, \beta, d, e; p) \\
&\quad = \sum_{n=0}^{z} \frac{[q^{-z/2}; q^{1/2}]_n[-q^{-d/2}; q^{1/2}]_n[q^{1+\alpha+\beta}; q]_n[q^{1+\alpha}; q]_n}{[q; q]_n[-q^{1+(\alpha+\beta+z)/2}; q^{1/2}]_n[q^{1+(\alpha+\beta+d)/2}; q^{1/2}]_n} \\
&\qquad \cdot \frac{(1-q^{1+\alpha+\beta+2n})(-)^n}{[q^{1+\beta}; q]_n(1-q^{1+\alpha+\beta})} q^{n(n+1+z+d+2\beta)/2} \\
&\qquad \cdot {}_3\phi_2\left[\begin{matrix} q^{-n}, q^{1+\alpha+\beta+n}, q^{-x} \\ q^{1+\alpha}, q^{-d} \end{matrix}; q; q^{x-\beta-d}\right] {}_3\phi_2\left[\begin{matrix} q^{-n}, q^{1+\alpha+\beta+n}, q^{-y} \\ q^{1+\alpha}, q^{-e} \end{matrix}; q; q^{y-\beta-e}\right] \\
&\quad = \sum_{n=0}^{z} \frac{[q^{-z/2}; q^{1/2}]_n[-q^{-d/2}; q^{1/2}]_n[q^{3+\alpha+\beta}; q^2]_n[q^{1+\alpha+\beta}; q]_n q^{n(z+d)/2}}{[q; q]_n[-q^{1+(\alpha+\beta+z)/2}; q^{1/2}]_n[q^{1+(\alpha+\beta+d)/2}; q^{1/2}]_n[q^{1+\alpha+\beta}; q^2]_n} \cdot \\
&\qquad \cdot \phi\left[\begin{matrix} q^{-n}, q^{1+\alpha+\beta+n} : q^{-y}, q^{-x}; q^{x-d}, q^{y-e} \\ q^{-d}, q^{-e} : q^{1+\alpha}; q^{1+\beta} \end{matrix}; q^{x+y-\beta-e-d}, q; q\right]
\end{aligned}
$$

(on using (3.1))

$$
\begin{aligned}
&= \sum_{r,s \geqq 0} \frac{[q^{-z/2}; q^{1/2}]_{r+s}[-q^{-d/2}; q^{1/2}]_{r+s}[q^{2+\alpha+\beta}; q]_{2r+2s}[q^{-x}; q]_r}{[q; q]_r[q; q]_s[-q^{1+(\alpha+\beta+z)/2}; q^{1/2}]_{r+s}[q^{1+(\alpha+\beta+d)/2}; q^{1/2}]_{r+s}} \cdot \\
&\quad \cdot \frac{[q^{-y}; q]_r[q^{x-d}; q]_s[q^{y-e}; q]_s(-1)^{r+s} q^{-(r+s)(r+s+1-z-d)/2+r(x+y-\beta-e-d)+s}}{[q^{-e}; q]_{r+s}[q^{-d}; q]_{r+s}[q^{1+\alpha}; q]_r[q^{1+\beta}; q]_s} \cdot \\
&\quad \cdot \Bigg\{ \sum_{n \geqq 0} \frac{[q^{1+\alpha+\beta+2r+2s}; q]_n[q^{3+\alpha+\beta+2r+2s}; q^2]_n[-q^{(-d+r+s)/2}; q^{1/2}]_n}{[q; q]_n[q^{1+\alpha+\beta+2r+2s}; q^2]_n[q^{1+(\alpha+\beta+d+r+s)/2}; q^{1/2}]_n} \\
&\qquad \cdot \frac{[q^{(-z+r+s)/2}; q^{1/2}]_n q^{n(2+d)/2}}{[-q^{1+(\alpha+\beta+z+r+s)/2}; q^{1/2}]_n q^{n(r+s)}} \Bigg\} \\
&= \frac{[q^{(1-d)/2}; q^{1/2}]_z[q^{d/2}; q^{1/2}]_z[-q^{1+(\alpha+\beta)/2}; q^{1/2}]_z}{[q^{-d/2}; q^{1/2}]_z[q^{1+(\alpha+\beta+d)/2}; q^{1/2}]_z[-q^{1/2}; q^{1/2}]_z} \\
&\quad \cdot \phi\left[\begin{matrix} q^{-z}, q^{1+(\alpha+\beta)/2}, q^{(3+\alpha+\beta)/2} : q^{-x}, q^{-y}; q^{x-d}, q^{y-e} \\ q^{(1-z-d)/2}, q^{1-(d+z)/2}, q^{-e} : q^{1+\alpha}; q^{1+\beta} \end{matrix}; q^{x+y-\beta-e-d}, q; q\right]
\end{aligned}
$$

(on summing the inner series by the summation theorem [12, 5.8])

$$= \frac{p^{(1-d)/2}; p^{1/2}]_z [p^{d/2}; p^{1/2}]_z [-p^{1+(\alpha+\beta)/2}; p^{1/2}]_z}{[p^{-d/2}; p^{1/2}]_z [p^{1+(\alpha+\beta+d)/2}; p^{1/2}]_z [-p^{1/2}; p^{1/2}]_z}$$

$$\cdot \phi\left[\begin{matrix} p^{-z}, p^{1+(\alpha+\beta)/2}, p^{(3+\alpha+\beta)/2} : p^{-x}, p^{-y}; p^{x-d}, p^{y-e} \\ p^{(1-d-z)/2}, p^{(2-d-z)/2}, p^{-e} : p^{1+\alpha}; p^{1+\beta} \end{matrix}; p, p^{-x-y-\alpha}; p\right]. \tag{6.4}$$

Setting $d = N$, $e = M$ in (6.4), we get (6.3).

In (6.3) let $z \to N$ to get

$$S_N(x, y; \alpha, \beta, N, M; p)$$

$$= \frac{[p^{2+\alpha+\beta}; p]_N (-1)^N p^{N(N-1)/2}}{[p^{-M}; p]_N p^{N(\alpha+x+y)}}$$

$$\cdot \sum_{r \geq 0} \frac{[p^{-x}; p]_r [p^{-y}; p]_r [p^{x-N}; p]_{N-r} [p^{y-m}; p]_{N-r} p^{r(1+\alpha)}}{[p; p]_r [p; p]_r [p^{1+\alpha}; p]_{N-r} [p^{1+\beta}; p]_r [p^{1+\beta}; p]_{N-r} p^{-r(x+y)}}$$

$$= \frac{(-1)^N [p^{2+\alpha+\beta}; p]_N [p^{-y}; p]_x [p^{-x}; p]_x [p^{-N+x}; p]_{N-x} [p^{y-M}; p]_{N-x}}{[p; p]_x [p; p]_{N-x} [p^{-M}; p]_N [p^{1+\alpha}; p]_x [p^{1+\beta}; p]_{N-x}}$$

$$\cdot \frac{p^{N(N-1)/2+x(1+\alpha+x+y)}}{p^{N(\alpha+x+y)}}.$$

Hence

$$S_N(x, y; \alpha, \beta, N, N; p) = \begin{cases} 0 & \text{if } y \neq x, \\ \dfrac{1}{j(x)} & \text{if } y = x. \end{cases} \tag{6.5}$$

Formula (6.5) gives the dual orthogonality relation for q-Hahn polynomials. This is a special case ($b = 0$) of the orthogonality for the q-Racah polynomials which was given by Askey and Wilson in [1, eq. 1.7]. The q-Racah polynomials are defined by

$$P_n(\mu(x); a, b, c, d; q) = P_n(\mu(x)) = {}_4\phi_3\left[\begin{matrix} q^{-n}, abq^{1+n}, q^{-x}, cdq^{1+x} \\ aq, bdq, cq \end{matrix}; q; q\right] \tag{6.6}$$

($\mu(x) = q^{-x} + cdq^{1+x}$ and aq, bdq or cq is of the form q^{-N}). They satisfy the orthogonality relation (see [1] for details)

$$\sum_{x=0}^{N} j(x) P_m(\mu(x)) P_n(\mu(x)) = \begin{cases} 0 & \text{if } m \neq n, \\ \dfrac{1}{h(n)} & \text{if } m = n, \end{cases} \tag{6.7}$$

for $n, m = 0, 1, 2, \cdots, N$ where

$$j(x) = \frac{[aq; q]_x [cq; q]_x [bdq; q]_x [cdq; q]_x (1 - cdq^{1+2x})}{[q; q]_x [dq; q]_x \left[\dfrac{cq}{b}; q\right]_x \left[\dfrac{cdq}{a}; q\right]_x (1 - cdq)(abq)^x}$$

and

$$h(n)=\frac{[aq;q]_n[cq;q]_n[bdq;q]_n[abq;q]_n(1-abq^{2n+1})}{[q;q]_n[bq;q]_n\left[\frac{aq}{d};q\right]_n\left[\frac{abq}{c};q\right]_n(1-abq)(cdq)^n}$$

$$\cdot\frac{[dq;q]_\infty\left[\frac{cq}{b};q\right]_\infty\left[\frac{cdq}{a};q\right]_\infty\left[\frac{1}{abq};q\right]_\infty}{\left[\frac{d}{a};q\right]_\infty\left[\frac{c}{ab};q\right]_\infty[cdq^2;q]_\infty\left[\frac{1}{b};q\right]_\infty}.$$

A proof of (6.7) can be given by following a procedure very similar to one used in proving (6.2). Indeed transforming the second of the two ${}_4\phi_3$'s in (6.7) by (3.3), we get

$$\sum_{x=0}^{N} j(x)P_m(\mu(x))P_n(\mu(x))$$

$$=\sum_{x=0}^{N}\frac{[aq;q]_x[cq;q]_x[bdq;q]_x[cdq;q]_x}{[q;q]_x[dq;q]_x\left[\frac{cq}{b};q\right]_x\left[\frac{cdq}{a};q\right]_x}$$

$$\cdot\frac{(1-cdq^{2x+1})\left[\frac{abq}{c};q\right]_n\left[\frac{aq}{d};q\right]_n\left(\frac{cd}{a}\right)^n}{(1-cdq)[bdq;q]_n[cq;q]_n(abq)^x}\,{}_4\phi_3\left[\begin{matrix}aq^{1+x},\frac{a}{cd}q^{-x},abq^{1+n},q^{-n}\\ aq,\frac{aq}{d},\frac{abq}{c}\end{matrix};q;q\right]$$

$$\cdot\,{}_4\phi_3\left[\begin{matrix}q^{-m},abq^{1+m},q^{-x},cdq^{1+x}\\ aq,bdq,cq\end{matrix};q;q\right]$$

$$=\frac{(cd)^n\left[\frac{aq}{d};q\right]_n\left[\frac{abq}{c};q\right]_n}{a^n[bdq;q]_n[cq;q]_n}\sum_{r,s\geqq 0}\frac{[abq^{1+n};q]_s[q^{-n};q]_s[q^{-m};q]_r[abq^{1+m};q]_r}{[q;q]_r[q;q]_s[aq;q]_s\left[\frac{aq}{d};q\right]_s\left[\frac{abq}{c};q\right]_s}$$

$$\cdot\frac{\left[\frac{a}{cd};q\right]_s[cdq^2;q]_{2r}[aq^{1+r};q]_s(-1)^rq^{-r(r+1+2s)/2}}{\left[\frac{cd}{a}q^{1-s};q\right]_r[dq;q]_r\left[\frac{cq}{b};q\right]_r(abq)^r}$$

$$\cdot\,{}_6\phi_5\left[\begin{matrix}cdq^{1+2r},q^{3/2+r}\sqrt{cd},-q^{3/2+r}\sqrt{cd},bdq^{1+r},cq^{1+r},aq^{1+r+s}\\ q^{1/2+r}\sqrt{cd},-q^{1/2+r}\sqrt{cd},\frac{c}{b}q^{1+r},dq^{1+r},\frac{cd}{a}q^{1+r-s}\end{matrix};q;\frac{q^{-r-s}}{abq}\right]$$

(using the summation theorem [11, 3.3.1.4])

$$=\frac{\left[\frac{abq}{c};q\right]_n\left[\frac{aq}{d};q\right]_n\left(\frac{cd}{a}\right)^n[cdq^2;q]_\infty\left[\frac{1}{b};q\right]_\infty\left[\frac{c}{ab};q\right]_\infty\left[\frac{d}{a};q\right]_\infty}{[bdq;q]_n[cq;q]_n[dq;q]_\infty\left[\frac{cdq}{a};q\right]_\infty\left[\frac{1}{abq};q\right]_\infty\left[\frac{cq}{b};q\right]_\infty}$$

$$\cdot\sum_{r\geqq 0}\frac{[q^{-m};q]_r[abq^{1+m};q]_rq^r}{[q;q]_r[abq^2;q]_r}\,{}_3\phi_2\left[\begin{matrix}aq^{1+r},abq^{1+n},q^{-n}\\ aq,abq^{2+r}\end{matrix};q;q\right].$$

Summing the inner ${}_3\phi_2$ by the q-analogue of Saalschütz's theorem, we readily get (6.7).

The proof of the orthogonality relation in [1] is different. There $P_n(\mu(x))$ was shown to be orthogonal to a polynomial of each lower degree. The specific polynomials were chosen so they could be attached to the weight function. The proof in [1] is more elementary than the proof we gave, but there may be some interest in a proof that gives the orthogonality directly.

REFERENCES

1. R. ASKEY AND J. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, SIAM J. Math. Anal., 10 (1979) pp. 1008–1016.
2. W. N. BAILEY, *A reducible case of the fourth type of Appell's hypergeometric function of two variables*, Quart. J. Math. (Oxford), 4 (1933), pp. 305–308.
3. J. L. BURCHNALL AND T. W. CHAUNDY, *Expansion of Appell's double hypergeometric functions*, Quart. J. Math. (Oxford), 11 (1940), pp. 249–70.
4. G. GASPER, *Non-negativity of discrete Poisson kernel for the Hahn polynomials*. J. Math. Anal. Appl., 42 (1973), pp. 438–51.
5. ———, *Product of terminating ${}_3F_2$ series*, Pacific J. Math., 56 (1975), pp. 87–95.
6. ———, *Formulas of the Dirichlet-Mehler type*, Proc. Conference on Fractional Calculus and its Application to the Mathematical Sciences, Springer-Verlag, New York, 1974.
7. W. HAHN, *Über orthogonal Polynome, die q-Differenzenglichungen qenügen*, Math. Nach., 2 (1949), pp. 4–34.
8. F. H. JACKSON, *On basic double hypergeometric functions*, Quart. J. Math. (Oxford), 13 (1942), pp. 69–82.
9. T. M. MACROBERT, *Functions of a Complex Variable*, (5th edition), Macmillan, New York, 1962.
10. D. B. SEARS, *On the transformation theory of basic hypergeometric functions*. Proc. London Math. Soc. (2) 53 (1951), pp. 158–80.
11. L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.
12. A. VERMA AND V. K. JAIN, *Transformations between basic hypergeometric series on different bases and identities of Rogers-Ramanujan type*, J. Math. Anal. Appl., 76 (1980).
13. G. N. WATSON *Product of two hypergeometric functions*, Proc. London Math. Soc., (2) 20 (1922), pp. 189–95.

SIAM J. MATH. ANAL.
Vol. 12, No. 6, November 1981

0036-1410/81/1206-0006 $01.00/0

SOME TRANSFORMATIONS OF BASIC HYPERGEOMETRIC FUNCTIONS. PART II*

V. K. JAIN†

Abstract. q-analogues of some quadratic transformations are obtained. We also derive the generalizations of these transformations of basic hypergeometric series which on specialization yield some known as well as new summation theorems.

1. Carlitz [5] obtained the q-analogue of the terminating version of the following quadratic transformation:

$$(1.1)\qquad {}_3F_2\left[\begin{matrix}a, b, c\\ 1+a-b, 1+a-c\end{matrix}; x\right] = (1-x)^{-a}\,{}_3F_2\left[\begin{matrix}\frac{a}{2}, \frac{1}{2}(1+a), 1+a-b-c\\ 1+a-b, 1+a-c\end{matrix}; -\frac{4x}{(1-x)^2}\right]$$

where $a=-n$. In § 3 of this note we obtain q-analogues of the quadratic transformations

$$(1.2)\qquad {}_2F_1\left[\begin{matrix}a, b\\ 2b\end{matrix}; 2z\right] = (1-z)^{-a}\,{}_2F_1\left[\begin{matrix}\frac{a}{2}, \frac{1}{2}(1+a)\\ b+\frac{1}{2}\end{matrix}; \frac{z^2}{(1-z)^2}\right],$$

$$(1.3)\qquad {}_2F_1\left[\begin{matrix}2a, a+b\\ 2a+2b\end{matrix}; z\right] = (1-z)^{-a}\,{}_2F_1\left[\begin{matrix}a, b\\ a+b+\frac{1}{2}\end{matrix}; -\frac{z^2}{4(1-z)}\right],$$

$$(1.4)\qquad {}_2F_1\left[\begin{matrix}2a, 2b\\ a+b+\frac{1}{2}\end{matrix}; z\right] = {}_2F_1\left[\begin{matrix}a, b\\ a+b+\frac{1}{2}\end{matrix}; 4z(1-z)\right],$$

and discuss some of their generalizations. As applications of these results we obtain some of the known, as well as new summation formulae for basic hypergeometric series.

2. Definitions and notation. If we let

$$|q|<1,\quad [a;q]_n=(1-a)(1-aq)\cdots(1-aq^{n-1}),\quad [a;q]_0=1$$

and $[a;q]_\infty=\prod_{r=0}^{\infty}(1-aq^r)$ then we may define the basic hypergeometric series as

$${}_{p+1}\phi_{p+r}\left[\begin{matrix}a_1, a_2, \cdots, a_{p+1}\\ b_1, b_2, \cdots, b_{p+r}\end{matrix}; q; x\right] = \sum_{n=0}^{\infty}\frac{[a_1;q]_n\cdots[a_{p+1};q]_n x^n(-1)^{nr}q^{rn(n-1)/2}}{[q;q]_n[b_1;q]_n\cdots[b_{p+r};q]_n},$$

where the series ${}_{p+1}\phi_{p+r}(x)$ converges for all positive integral values of r and for all x; when $r=0$, it converges only for $|x|<1$.

3. A q-analogue of the transformation (1.2) is

$$(3.1)\qquad {}_3\phi_2\left[\begin{matrix}a, b, -b\\ b^2, az\end{matrix}; q; -z\right] = \frac{[z;q]_\infty}{[az;q]_\infty}\,{}_2\phi_1\left[\begin{matrix}a, aq\\ b^2q\end{matrix}; q^2; z^2\right],\qquad |z|<1.$$

Proof of (3.1). In view of the q-analogue of Vandermonde's theorem [9, § 3.3.2.7], we have

$$(3.2)\qquad {}_2\phi_1\left[\begin{matrix}q^{-n}, q^{1-n}\\ b^2q\end{matrix}; q^2; q^2\right] = \frac{[b^2;q^2]_n q^{-n(n-1)/2}}{[b^2;q]_n}.$$

* Received by the editors May 30, 1978, and in final revised form December 30, 1980.
† Department of Mathematics, Bareilly College, Bareilly (U.P.) India.

Using (3.2) we may rewrite the left-hand side of (3.1) in the form

$$\sum_{n=0}^{\infty} \frac{[a;q]_n(-z)^n q^{n(n-1)/2}}{[q;q]_n[az;q]_n} {}_2\phi_1\left[\begin{matrix} q^{-n}, q^{1-n} \\ b^2q \end{matrix}; q^2; q^2\right]$$

$$= \sum_{r=0}^{\infty} \frac{[a;q]_{2r}z^{2r}}{[q^2;q^2]_r[b^2q;q^2]_r[az;q]_{2r}} {}_1\phi_1\left[\begin{matrix} aq^{2r} \\ azq^{2r} \end{matrix}; q; z\right]. \tag{3.3}$$

Summing the resulting inner series by a limiting case of the q-analogue of Gauss's theorem [9; § 3.3.2.5], we get the right-hand side of (3.1).

Rewrite (3.1) in the form

$${}_2\phi_1\left[\begin{matrix} a, aq \\ b^2q \end{matrix}; q^2; z^2\right] = \sum_{n=0}^{\infty} \frac{[a;q]_n[b^2;q^2]_n(-z)^n}{[q;q]_n[b^2;q]_n} \sum_{r=0}^{\infty} \frac{[aq^n;q]_r z^r}{[q;q]_r}, \tag{3.4}$$

replace z by zt and multiply by $t^{c-1}(1-tq)_{d-c-1}$, and take the q-beta integral [7] on both sides to obtain

$${}_4\phi_3\left[\begin{matrix} a, aq, c, cq \\ b^2q, d, dq \end{matrix}; q^2; z^2\right] = \sum_{n=0}^{\infty} \frac{[a;q]_n[b^2;q^2]_n[c;q]_n(-z)^n}{[q;q]_n[b^2;q]_n[d;q]_n} {}_2\phi_1\left[\begin{matrix} aq^n, cq^n \\ dq^n \end{matrix}; q; z\right], \tag{3.5}$$

$|z|<1$. (3.5) for $c=q^{-N}$ (N a nonnegative integer), $z=q$ yields the following q-analogue of a transformation due to Bailey [3; § 4.41]:

$${}_4\phi_3\left[\begin{matrix} a, aq, q^{1-N}, q^{-N} \\ b^2q, d, dq \end{matrix}; q^2; q^2\right] = \frac{a^N\left[\frac{d}{a};q\right]_N}{[d;q]_N} {}_4\phi_2\left[\begin{matrix} a, b, -b, q^{-N} \\ b^2, \frac{a}{d}q^{1-N} \end{matrix}; q; -\frac{q}{d}\right]. \tag{3.6}$$

It may be remarked that using (3.2), we can rewrite

$${}_4\phi_3\left[\begin{matrix} a^2, b^2, c, -c \\ ab\sqrt{q}, -ab\sqrt{q}, c^2 \end{matrix}; q; q\right]$$

in the form

$$\sum_{n=0}^{\infty} \frac{[a^2;q]_n[b^2;q]_n q^{n(n+1)/2}}{[q;q]_n[a^2b^2q;q^2]_n} {}_2\phi_1\left[\begin{matrix} q^{-n}, q^{1-n} \\ c^2q \end{matrix}; q^2; q^2\right]$$

$$= \sum_{r=0}^{\infty} \frac{[a^2;q]_{2r}[b^2;q]_{2r}q^{2r}}{[q^2;q^2]_r[c^2q;q^2]_r[a^2b^2q;q^2]_{2r}} {}_2\phi_2\left[\begin{matrix} a^2q^{2r}, b^2q^{2r} \\ abq^{1/2+2r}, -abq^{1/2+2r} \end{matrix}; q; -q\right] \tag{3.7}$$

$$= \frac{[a^2q;q^2]_\infty[b^2q;q^2]_\infty}{[a^2b^2q;q^2]_\infty[q;q^2]_\infty} {}_2\phi_1\left[\begin{matrix} a^2, b^2 \\ c^2q \end{matrix}; q^2; q^2\right].$$

Now, if a or b is of the form q^{-N}, the resulting ${}_2\phi_1$ may be summed by the q-analogue of Vandermonde's theorem to yield a q-analogue of Watson's summation theorem due to Andrews [2] (since if a or b is of the form $q^{-N-1/2}$, the above ${}_4\phi_3$ is equal to zero).

Next, we prove a q-analogue of the quadratic transformation (1.3) in the form

$${}_3\phi_2\left[\begin{matrix} a^2, ab, -ab \\ a^2b^2, -za^2 \end{matrix}; q; z\right] = \frac{[a^2z^2;q^2]_\infty}{[-za^2;q]_\infty[z;q]_\infty} {}_2\phi_2\left[\begin{matrix} a^2, b^2 \\ a^2b^2q, z^2a^2 \end{matrix}; q^2; a^2z^2q\right]. \tag{3.8}$$

Proof of (3.8). Transforming the ${}_2\phi_1$ in the right-hand side of (3.1) by the formula

$$(3.9)\qquad {}_2\phi_1\left[\begin{matrix}a, b\\ c\end{matrix}; q; z\right]=\frac{[az;q]_\infty}{[z;q]_\infty}{}_2\phi_2\left[\begin{matrix}a, \frac{c}{b}\\ c, az\end{matrix}; q; bz\right]$$

due to Jackson [6] and replacing a, b and z by a^2, ab and $-z$ respectively, we get (3.8).

Lastly, a q-analogue of (1.4) is

$$(3.10)\qquad {}_3\phi_2\left[\begin{matrix}a^2, b^2, z\\ ab\sqrt{q}, -ab\sqrt{q}\end{matrix}; q; q\right]={}_3\phi_2\left[\begin{matrix}a^2, b^2, z^2\\ a^2b^2q, 0\end{matrix}; q^2; q^2\right],$$

where a or b is of the form q^{-N}.

To prove (3.10), we first prove the following transformations:

$$(3.11)\qquad {}_4\phi_4\left[\begin{matrix}a^2, b^2, c, d\\ ab\sqrt{q}, -ab\sqrt{q}, f, g\end{matrix}; q; -z\right]=\sum_{n=0}^{\infty}\frac{[a^2;q^2]_n[b^2;q^2]_n[c;q]_n[d;q]_nz^n}{[q;q]_n[a^2b^2q;q^2]_n[f;q]_n[g;q]_n}\cdot{}_3\phi_3\left[\begin{matrix}q^{-n}, cq^n, dq^n\\ -q, fq^n, gq^n\end{matrix}; q; -zq^n\right]$$

and

$$(3.12)\qquad {}_4\phi_4\left[\begin{matrix}a, \frac{q}{a}, c, d\\ -q, b, f, g\end{matrix}; q; -\frac{bz}{q}\right]=\sum_{n=0}^{\infty}\frac{\left[\frac{b}{a};q^2\right]_n\left[\frac{ab}{q};q^2\right]_n[c;q]_n[d;q]_nz^n}{[q^2;q^2]_n[b;q]_n[f;q]_n[g;q]_n}\cdot{}_3\phi_3\left[\begin{matrix}b^{-1}q^{1-n}, cq^n, dq^n\\ -q, fq^n, gq^n\end{matrix}; q; -bzq^{n-1}\right].$$

Proof of (3.11). In view of the q-analogue of Saalschütz's summation theorem [9, § 3.3.2.2], the left-hand side of (3.11) can be rewritten as

$$(3.13)\qquad S=\sum_{n=0}^{\infty}\frac{[a^2;q^2]_n[b^2;q^2]_n[c;q]_n[d;q]_nz^n}{[q;q]_n[a^2b^2q;q^2]_n[f;q]_n[g;q]_n}{}_3\phi_2\left[\begin{matrix}q^{-n}, q^{1-n}, a^{-2}b^{-2}q^{1-2n}\\ a^{-2}q^{2-2n}, b^{-2}q^{2-2n}\end{matrix}; q^2; q^2\right].$$

Rearranging the two series and then diagonalizing, we get the right-hand side of (3.11).

Similarly (3.12) is proved by observing that its left-hand side is

$$(3.14)\qquad S=\sum_{n=0}^{\infty}\frac{\left[\frac{b}{a};q^2\right]_n\left[\frac{ab}{q};q^2\right]_n[c;q]_n[d;q]_nz^n}{[q^2;q^2]_n[b;q]_n[f;q]_n[g;q]_n}\cdot{}_3\phi_2\left[\begin{matrix}q^{-2n}, b^{-1}q^{1-n}, b^{-1}q^{2-n}\\ ab^{-1}q^{2-2n}, a^{-1}b^{-1}q^{3-2n}\end{matrix}; q^2; q^2\right].$$

If a, b, c or d is of the form q^{-N}, $z=-gq$, $f=-cd$ and then $g\to\infty$, (3.11) yield[1] (we sum the inner series by the q-analogue of Saalschütz's summation theorem):

$$(3.15)\qquad {}_4\phi_3\left[\begin{matrix}a^2, b^2, c, d\\ ab\sqrt{q}, -ab\sqrt{q}, -cd\end{matrix}; q; q\right]={}_4\phi_3\left[\begin{matrix}a^2, b^2, c^2, d^2\\ a^2b^2q, -cd, -cdq\end{matrix}; q^2; q^2\right].$$

(3.15) for $d=0$ and $c=z$ reduces to (3.10), whereas for $d=-c$, (3.15) yields

[1] I am indebted to Professor R. Askey for drawing my attention to the formula (3.15) which he and J. Wilson have obtained recently.

$$(3.16)\qquad {}_4\phi_3\left[\begin{matrix} a^2, b^2, c, -c \\ ab\sqrt{q}, -ab\sqrt{q}, c^2 \end{matrix}; q; q\right] = {}_3\phi_2\left[\begin{matrix} a^2, b^2, c^2 \\ a^2b^2q, c^2q \end{matrix}; q^2; q^2\right].$$

In (3.16), if a or b is of the form q^{-N}, we can sum the ${}_3\phi_2$ on the right-hand side by the q-analogue of Saalschütz's theorem to obtain the q-analogue of Watson's summation theorem due to Andrews [2]. On the other hand, if $c = q^{-N}$ we get

$$(3.17)\qquad {}_4\phi_3\left[\begin{matrix} a^2, b^2, -q^{-N}, q^{-N} \\ ab\sqrt{q}, -ab\sqrt{q}, q^{-2N} \end{matrix}; q; q\right] = \frac{[a^2q; q^2]_N[b^2q; q^2]_N}{[a^2b^2q; q^2]_N[q; q^2]_N},$$

which is the q-analogue of a result of Bailey [4]. (3.17) for $N \to \infty$ gives the q-analogue of the Gauss's second summation theorem due to Andrews [1]. It may be remarked that (3.17) also yields the q-analogue of Gauss's second summation theorem due to Andrews if $c = d = f = g$, $z = q$ and summing the inner series on the right-hand side by a limiting case of the q-analogue of Gauss's theorem and using the q-analogue of Gauss's theorem.

On the other hand, if $d = q^{-N}$, $f = -(c/b)q^{1-N}$, $z = -(g/b)q^2$ and then $g \to \infty$, (3.12) gives the transformation

$$(3.18)\qquad {}_4\phi_3\left[\begin{matrix} a, \frac{q}{a}, c, q^{-N} \\ -q, b, -\frac{c}{b}q^{1-N} \end{matrix}; q; q\right] = \frac{[-b; q]_N\left[-\frac{q}{c}; q\right]_N}{\left[-\frac{b}{c}; q\right]_N[-q; q]_N} {}_4\phi_3\left[\begin{matrix} \frac{b}{a}, \frac{ab}{q}, c^2, q^{-2N} \\ b^2, -cq^{-N}, -cq^{1-N} \end{matrix}; q^2; q^2\right].$$

Formula (3.18) may also be deduced from (3.15) on replacing a^2, b^2 and d by b/a, ab/q and q^{-N} and transforming the left-hand side by a result due to Sears [8, § 8.3]. Equation (3.18) for $c = -q^{-N}$ yields the q-analogue of a terminating version of Whipple's summation theorem for ${}_3F_2$ (1) due to Bailey [4]:

$$(3.19)\qquad {}_4\phi_3\left[\begin{matrix} a, \frac{q}{a}, -q^{-N}, q^{-N} \\ -q, b, b^{-1}q^{1-2N} \end{matrix}; q; q\right] = \frac{[ab; q^2]_N\left[\frac{bq}{a}; q^2\right]_N}{[b; q]_{2N}}.$$

In (3.19), letting $N \to \infty$, we get the q-analogue of Bailey's summation theorem due to Andrews [1]. This may also be deduced from (3.12) if $c = d = f = g$, $z = q$, summing the inner series by a limiting case of the q-analogue of Gauss's theorem and then summing by the q-analogue of Gauss's theorem.

Formulae (3.17) and (3.19) are different in nature from the q-analogues of the Watson and Whipple theorems given by Andrews [2] (see also Bailey [4]).

Acknowledgment. I am grateful to Dr. A. Verma for suggesting the problem and for his helpful discussions during the preparation of this paper.

REFERENCES

[1] G. E. ANDREWS, *On the q-analogue of Kummer's theorem and applications*, Duke Math. J., 40 (1973), pp. 525–528.
[2] ———, *On q-analogues of the Watson and Whipple summation*, this Journal, 7 (1976), pp. 332–336.
[3] W. N. BAILEY, *Transformations of generalized hypergeometric series*. Proc. London Math. Soc. 29 (1929), pp. 495–502.
[4] ———, *On the sum of a terminating* ${}_3F_2$ (1). Quart. J. Math., 4 (1953), pp. 237–240.
[5] L. CARLITZ, *Some formulas of F. H. Jackson*. Monatshefte für Mathematike, 73 (1969), pp. 193–198.
[6] F. H. JACKSON, *Transformations of q-series*, Mess. Math., 39 (1910), pp. 145–153.
[7] ———, *q-difference equations*, Amer. J. Math., 32 (1910), pp. 305–314.
[8] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. London Math. Soc., (2) 53 (1951), pp. 158–180.
[9] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.