

## THE NATURE OF RESONANCE IN A SINGULAR PERTURBATION PROBLEM OF TURNING POINT TYPE\*

P. P. N. DE GROEN†

**Abstract.** On the interval  $(a, b)$  with  $a < 0 < b$  we study the boundary value problem

$$-\varepsilon u'' + xp(x, \varepsilon)u' + xq(x, \varepsilon)u = r(\varepsilon)u, \quad u(a) = A, \quad u(b) = B, \quad 0 < \varepsilon \ll 1.$$

The related eigenvalue problem has a discrete set of eigenvalues for each  $\varepsilon > 0$ . We expand each eigenvalue in a formal asymptotic series in integral powers of  $\varepsilon$  and we prove the validity of the expansion with the aid of the Rayleigh quotient characterizations of the eigenvalues. If  $r(\varepsilon)$  is not equal to an eigenvalue, the solution exists and is unique; we prove that it decays exponentially for  $\varepsilon \rightarrow +0$ , provided the distance between  $r(\varepsilon)$  and the nearest eigenvalue is larger than  $\exp(-\gamma/\varepsilon)$  for some positive  $\gamma$  depending on  $p$ . If  $r(\varepsilon)$  is equal to an eigenvalue, no solution exists (in general) and, if  $r(\varepsilon)$  is near enough to an eigenvalue, the dominant term in the solution is a multiple of the corresponding eigenfunction. From a spectral point of view the "Ackerberg-O'Malley resonance" is the familiar effect that the nearest free mode of the equation is amplified by the inverse of the distance from  $r(\varepsilon)$  to the corresponding eigenvalue.

### 1. Introduction.

**1.1. The problem.** In this paper we study the singularly perturbed two-point boundary value problem of turning point type on the real interval  $[a, b]$

$$(1.1a) \quad L_\varepsilon u := -\varepsilon u'' + xp(x, \varepsilon)u' + xq(x, \varepsilon)u = r(\varepsilon)u, \quad (' = d/dx),$$

$$(1.1b) \quad u(a) = A, \quad u(b) = B, \quad a < 0 < b,$$

where  $\varepsilon$  is a small positive parameter, and where  $p, 1/p, q$  and  $r$  are sufficiently smooth functions with respect to both parameters  $x$  and  $\varepsilon$ . We shall treat the case  $p > 0$  only, since the analysis for  $p < 0$  is analogous. Without loss of generality we can assume  $p(0,0) = 1$  and

$$(1.2) \quad \Delta := \int_0^b tp(t, 0) dt \leq \int_a^0 tp(t, 0) dt.$$

This problem has some intriguing features due to the fact that the coefficient of  $u'$  in equation (1.1a) changes sign in the interval. In the easier and well-analyzed case where the coefficient of  $u'$  is of one sign and is positive (negative) throughout the interval, the contribution to the solution coming from the prescribed boundary value at the right (left) endpoint is exponentially small outside a small boundary layer near that endpoint; cf. [11] or [12]. We note that "exponentially small" means "of the order  $\mathcal{O}(\exp(-\gamma/\varepsilon))$ ,  $\varepsilon \rightarrow +0$ , for some  $\gamma > 0$ ". The analysis in this easier case transferred to problem (1.1) suggests that the contribution from the boundary value at both endpoints is exponentially small; hence the solution of problem (1.1) is exponentially small uniformly in every compact subinterval of  $(a, b)$  and boundary layers are located at both endpoints. However, this suggestion is not always true, as can be seen from the following example,

$$(1.3) \quad -\varepsilon u'' + xu' - ru = 0, \quad u(a) = A, \quad u(1) = B, \quad a \leq 1,$$

which can be solved exactly in terms of parabolic cylinder functions or in terms of the confluent hypergeometric functions  ${}_1F_1(-\frac{1}{2}r, \frac{1}{2}, x^2/2\varepsilon)$  and  $x{}_1F_1(\frac{1}{2}-\frac{1}{2}r, \frac{3}{2}, x^2/2\varepsilon)$ , cf. [5, § 2]. By well-known asymptotic formulas for these functions we indeed find exponential

\* Received by the editors October 1, 1977 and in revised form July 24, 1978.

† Department of Mathematics, Eindhoven University of Technology, Eindhoven, The Netherlands.

decay if  $r$  is not a nonnegative integer,

(1.4a)

$$u_\varepsilon(x) \sim A \exp\{(a-x)/\varepsilon\} + B \exp\{(x-1)/\varepsilon\}, \quad \varepsilon \rightarrow +0 \text{ and } r \neq 0, 1, 2, \dots,$$

where  $\sim$  means “asymptotically equivalent”.

However, if  $r$  is a nonnegative integer, one of the confluent hypergeometric functions is equal to the  $r$ th Hermite polynomial and we find for  $\varepsilon \rightarrow +0$  and  $r = 0, 1, 2, \dots$ :

$$(1.4b) \quad u_\varepsilon(x) \sim \begin{cases} Bx^r + A \exp\{(a-x)/\varepsilon\}, & \text{if } a < -1 \\ \frac{1}{2}(B + (-1)^r A)x^r + \frac{1}{2}(B - (-1)^r A) \exp\{-(x-1)^2/2\varepsilon\}, & \text{if } a = -1. \end{cases}$$

We see that the solution of (1.3) does not decay at all in the exceptional case where  $r = 0, 1, 2, \dots$  and that in general (if  $a \neq 1$ ) one of the boundary layers disappears.

**1.2. History.** In [2] Ackerberg and O'Malley draw attention to problem (1.1). They establish exponential decay of its solution in the case  $r(0) \neq 0, 1, 2, \dots$ . For nonnegative integral values of  $r(0)$  they construct by the WKB method a formal approximation, which does not decay for  $\varepsilon \rightarrow +0$ . This approximation converges to a solution of the reduced equation  $xpu' + xqu = ru$ , whose magnitude is fixed by the boundary condition  $u(b) = B$  if equality in (1.2) does not hold and by  $u(b) = \frac{1}{2}(B + (-1)^r A)$  otherwise. This phenomenon, that the solution of (1.1) does not decay exponentially and converges to a definite nonzero solution of the reduced equation, Ackerberg and O'Malley have called *resonance*. Their publication has drawn much interest and has been followed by a large number of papers which study this phenomenon of “resonance”; e.g. see [3], [8], [9], [10] and the references there. These papers steadily propose better approximations to the “resonant” solution of (1.1) and more refined criteria for “resonance” to occur, most derived by formal methods only and not supported by proofs. For a review of these papers we refer to the introduction of [10]. Olver constructs in [10] an approximation by linking together uniform approximations of two pairs of independent solutions of the equation. The boundary conditions at  $a$  and  $b$  and the continuity conditions across the turning point yield four linear equations which can be solved under certain conditions on  $r(\varepsilon)$ . His final conclusion is that for each nonnegative number  $n$  a function  $r(\varepsilon)$  exists such that the approximation and hence the solution itself shows “resonance” (in the sense of Ackerberg and O'Malley); moreover the “resonant” approximation remains valid if  $r(\varepsilon)$  is changed by an amount not exceeding  $e^{-\gamma/\varepsilon}$  with  $\gamma > \Delta$  and  $\Delta$  as in (1.2). We remark (1) that the same conclusion can be drawn from [5, Thm. 4.4 and Cor. 4.5], and (2) that the existence proof does not (and cannot, as we shall explain later on) yield a method for construction of such an  $r(\varepsilon)$ .

**1.3. Re-evaluation of the problem.** In the papers cited above the secondary question “Under what conditions does the solution of (1.1) show resonance in the sense of Ackerberg and O'Malley?” has obscured the original question “Can we find an asymptotic approximation to the solution of (1.1) and how does it look like if  $r(0)$  is a nonnegative integer?”; Ackerberg–O'Malley-resonance has been considered as a fundamental property of the solutions of equation (1.1). However, from the example (1.3) we can read that the first question is not the best one to ask. If  $a = -1$ , the solution of (1.3) is

$$(1.5) \quad u_\varepsilon(x, r) = \frac{1}{2}(A + B) \frac{{}_1F_1(-\frac{1}{2}r; \frac{1}{2}; x^2/2\varepsilon)}{{}_1F_1(-\frac{1}{2}r; \frac{1}{2}; 1/2\varepsilon)} + \frac{1}{2}(A - B) \frac{x {}_1F_1(\frac{1}{2} - \frac{1}{2}r; 3/2; x^2/2\varepsilon)}{{}_1F_1(\frac{1}{2} - \frac{1}{2}r; 3/2; 1/2\varepsilon)},$$

provided the denominators are nonzero. These denominators, considered as functions of  $r$ , have denumerably many simple zeros for each  $\varepsilon > 0$  and the zeros converge to the nonnegative integers for  $\varepsilon \rightarrow +0$ . Our first conclusion from this example is that a solution of problem (1.1) need not exist, a fact that is not mentioned in any of the papers cited above. The second conclusion is that it is not very interesting to ask for the conditions under which the solution of (1.1) (if it exists) converges to a definite solution of the reduced equation, since for every multiple of this limit we can ask the same question. As a matter of fact, for any point  $x_0 \in (0, 1)$ , any nonnegative integer  $n$  and any real number  $C$  we can find a function  $r(\varepsilon)$  with  $r(0) = n$  such that  $u_\varepsilon$  in the example (1.5) satisfies  $u_\varepsilon(x_0, r(\varepsilon)) = C$ , because  $n$  is the limit of a zero of a denominator; since the restriction of problem (1.3) to  $(x_0, 1)$  has no turning points, the well-known analysis implies that  $u_\varepsilon(x, r(\varepsilon))$  converges on  $(x_0, 1)$  (pointwise) to that solution of the reduced equation which takes the value  $C$  at  $x_0$ . Clearly the interesting question is how the mechanism works that provides solutions of any magnitude.

The answer to this question also can be read from the example (1.3) with  $a = -1$ . The zeros of the denominators in (1.5) are the eigenvalues of the operator  $-\varepsilon d^2/dx^2 + xd/dx$  in  $\mathcal{H}_0^1(-1, 1) \cap \mathcal{H}^2(-1, 1)$ . Let us denote these eigenvalues and the corresponding eigenfunctions by  $(\pi_k(\varepsilon), \tilde{\psi}_k(\cdot, \varepsilon))$  and let us assume that the eigenvalues are ordered in increasing sense (i.e.  $\lambda_{k+1} > \lambda_k$ ); they satisfy the relations

$$(1.6) \quad -\varepsilon \tilde{\psi}_k'' + x \tilde{\psi}_k' = \pi_k \tilde{\psi}_k \quad \text{and} \quad \lim_{\varepsilon \rightarrow +0} \pi_k(\varepsilon) = k.$$

We define  $Z_\varepsilon$  to be the ordinary boundary layer terms, as given in (1.4a),

$$Z_\varepsilon(x) := A \exp\{-(x+1)/\varepsilon\} + B \exp\{(x-1)/\varepsilon\}$$

and we expand the residue  $L_\varepsilon(u_\varepsilon - Z_\varepsilon)$  in the eigenfunctions,

$$(1.7) \quad L_\varepsilon(u_\varepsilon - Z_\varepsilon) = \sum \beta_k \tilde{\psi}_k.$$

The solution  $u_\varepsilon$  satisfies

$$(1.8) \quad u_\varepsilon = Z_\varepsilon + \sum_{k=0}^{\infty} \frac{\beta_k \tilde{\psi}_k}{\pi_k - r}.$$

If  $\pi_k - r$  is bounded away from zero for all  $k$ , the infinite sum in (1.8) is small, as is the residue in (1.7). However, if  $r(0) = n$  for some integer  $n$ , the  $n$ th term can be quite large and we obtain the approximation

$$u_\varepsilon \sim Z_\varepsilon + \frac{\beta_n(\varepsilon) \tilde{\psi}_n(\cdot, \varepsilon)}{\pi_n(\varepsilon) - r(\varepsilon)}, \quad \varepsilon \rightarrow +0 \quad \text{and} \quad r(0) = n.$$

This formula displays the mechanism at work in a resonant situation and it explains why the solution is so extremely sensitive for small variations in  $r(\varepsilon)$ . It is clear that an analogous formula can be given for the solution of (1.1). Problem (1.1) can be considered as the equation for the steady state of a vibrating system and in such a setting the phenomenon, that the solution grows beyond bound in the vicinity of an eigenvalue, is commonly called resonance. From this point of view the phenomenon, which Ackerberg and O'Malley have called resonance by chance (?), is a quite familiar spectral effect. We have pointed at this connection to the spectrum already in [13, § 9].

**1.4. Outline of the paper.** The purpose of this paper is to construct a uniformly valid approximation to the solution of problem (1.1), if it exists. The explanation of the phenomenon of resonance clearly indicates the road to follow in order to arrive at such

an approximation. First we have to determine the eigenvalues of the operator  $L_\varepsilon$  acting on  $\mathcal{H}^2(a, b) \cap \mathcal{H}_0^1(a, b)$ . Next we have to construct uniform approximations to the corresponding eigenfunctions. Finally we have to estimate the coefficients in an eigenfunction expansion of type (1.8) and we have to approximate the sum of the series, since the infinite series itself hardly can be considered as a satisfactory approximation. The techniques we shall use in our analysis are quite classical, namely the Rayleigh quotient characterization of eigenvalues, Sturm–Liouville theory for eigenfunction expansions, matched asymptotic expansions for the construction of approximations of the eigenfunctions and the maximum principle for the proof of their validity; cf. [4] and [12].

Our first result concerns the location of the eigenvalues. The eigenvalues are the values of  $\lambda$  for which the problem

$$(1.9) \quad L_\varepsilon u = -\varepsilon u'' + xpu' + xqu = \lambda u, \quad u(a) = u(b) = 0,$$

has a nontrivial solution. Sturm–Liouville theory implies that a denumerable set of eigenvalues and eigenfunctions

$$\{(\lambda_k(\varepsilon), \tilde{e}_k(\cdot, \varepsilon)) \mid k = 0, 1, 2, \dots\} \quad \text{with } L_\varepsilon \tilde{e}_k = \lambda_k \tilde{e}_k$$

exists; ordering these eigenvalues in an increasing sequence we find

$$(1.10) \quad \lambda_k(\varepsilon) = k + \mathcal{O}(\varepsilon), \quad \text{for } \varepsilon \rightarrow +0.$$

This result is already contained in [5] and [6], but the proof there is fairly complicated. Here we shall present an easier proof, based only on the minimax and maximin characterizations of the eigenvalues by Rayleigh’s quotient; cf. [4]. We transform equation (1.9) to a selfadjoint form and we construct formal approximations of its eigenfunctions. The maximum of Rayleigh’s quotient over the span of the first  $k$  of these approximate eigenfunctions yields an upper estimate for  $\lambda_{k-1}$  and the minimum over the orthogonal complement yields a lower estimate of  $\lambda_k$ . A good estimate of the maximum is derived easily since the maximum is taken over a finite dimensional space. An estimate of the minimum over the orthogonal complement, which is of infinite dimension, is more complicate since the estimates of the eigenfunctions are not uniform. We split this space into two subspaces such that in one of them Rayleigh’s quotient is large enough to be estimated from below by the Rayleigh’s quotient of the Hermite operator, cf. (1.3), whose eigenvalues are known, and such that the other subspace is of finite dimension.

Once the convergence of the eigenvalues to well-separated limits is established, we can expand the eigenvalues and the corresponding eigenfunctions of the symmetrized problem in formal power series in powers of  $\varepsilon$ . If  $p$  and  $q$  are  $\mathcal{C}^\infty$  we can compute all terms of these series by a formal asymptotic method which is analogous to the “suppression of secular terms” in celestial mechanics. The coefficients in the power series expansion of  $\lambda_k(\varepsilon)$  are uniquely determined by the condition that nonpolynomial solutions (which are exponentially large) have to be suppressed in every step of the iteration. The validity of these series is proved by expansion of the residue of the approximate eigenfunction in the true eigenfunctions of the symmetrized problem and by use of well-known estimates for the coefficients of such eigenfunction expansions. Transforming back to the original nonselfadjoint form we find an approximation of the eigenfunction which is uniformly valid in the interior boundary layer of width  $\mathcal{O}(\varepsilon^{1/2})$ .

At both sides of this interior boundary layer we can match the interior expansion to the regular expansion, whose lowest order term is the solution of the reduced equation  $xpu' + xqu = \lambda u$ . Both regular expansions are matched to the boundary conditions

$u(a) = u(b) = 0$  in ordinary boundary layers. The validity of the approximation on  $[a, -\varepsilon^{1/2}m]$  and  $[\varepsilon^{1/2}m, b]$  for some  $m > 0$  is proved by the maximum principle. We shall restrict our computation of an asymptotic approximation of the eigenfunction to a first order approximation, which outside the boundary layers has a relative error of the order  $\mathcal{O}(\varepsilon^{1/2})$ .

If  $r(\varepsilon)$  is not equal to any eigenvalue, problem (1.1) has a unique solution  $U_\varepsilon$ . By “matched asymptotic expansions” we construct a formal approximation  $Z_\varepsilon$ , which satisfies the boundary conditions (1.1b), and which is exponentially small in the interior of the interval. Assuming that  $n$  is the nonnegative integer nearest to  $r(0)$ , and using the eigenfunction expansion as in (1.8) we finally obtain the result

$$(1.11) \quad U_\varepsilon = Z_\varepsilon + \frac{\beta_n(\varepsilon)\tilde{e}_n(\cdot, \varepsilon)}{\lambda_n(\varepsilon) - r(\varepsilon)} + \text{an exponentially small error,}$$

where  $\beta_n$  is the coefficient of  $\tilde{e}_n$  in the eigenfunction expansion of  $(L_\varepsilon - r)(U_\varepsilon - Z_\varepsilon)$ . The magnitude of the (resonant) eigenfunction term in (1.11) can be read from the formula

$$(1.12) \quad \max_{a \leq x \leq b} |\beta_n(\varepsilon)\tilde{e}_n(x, \varepsilon)| = C\varepsilon^{-n-1/2} e^{-\Delta/\varepsilon} (1 + \mathcal{O}(\varepsilon^{1/2})), \quad (\varepsilon \rightarrow +0),$$

where  $C$  does not depend on  $\varepsilon$  and  $\Delta$  is given by (1.2). We see that the magnitude of the resonant part of (1.11) is of order unity if the distance from  $r(\varepsilon)$  to the nearest eigenvalue  $\lambda_n(\varepsilon)$  is of the same order as (1.12) and that the resonant part vanishes if the distance is of larger order.

Formula (1.11) together with the approximation of the eigenfunction  $\tilde{e}_n$  and the estimate of the coefficient  $\beta_n$  give a precise picture of the asymptotic behavior of the solution of problem (1.1) in the neighborhood of an eigenvalue. Unfortunately this picture inevitably contains the distance from  $r(\varepsilon)$  to the nearest eigenvalue. Since in general no better approximation for an eigenvalue can be obtained than an asymptotic (nonconvergent) power series in  $\varepsilon$ , the exponentially small orders in the distance cannot be detected (by asymptotic methods). Hence, if the asymptotic series of  $\lambda_n$  and  $r$  do not agree, the solution of (1.1) decays exponentially, but, if they agree, the magnitude of the resonant part cannot be determined in general. Only in the exceptional case where a solution of the equation  $L_\varepsilon u = ru$  happens to be known, which is normalized by  $|u(0)| + |u'(0)| = 1$  and which is bounded by some negative power of  $\varepsilon$  uniformly with respect to  $\varepsilon$  and  $x$ , can the magnitude of the resonant part be determined. Examples of such a case are problem (1.3) and problem (1.1) with  $xq - r \equiv 0$ . Moreover, if in a problem of type (1.1) the resonant part of the solution is of order unity, small changes in  $\varepsilon^{1/2}p, q$  and  $r$  do not affect the magnitude of the resonant part in first order, provided those changes are of an order smaller than (1.12) is, uniformly in  $x$ .

The methods employed here admit considerable generalizations, to the case where the sign of  $p$  is negative, to the case where there are several turning points, where a turning point is located at the boundary or where it is of higher order and to analogous (elliptic) problems in several dimensions; cf. [6] and [7].

**1.5. Notations.**  $\mathbb{N}, \mathbb{N}_0, \mathbb{R}$  and  $\mathbb{C}$  are the sets of natural, nonnegative integral, real and complex numbers. If  $I$  is an (open) interval in  $\mathbb{R}$ ,  $\mathcal{L}^2(I)$  denotes the set of square integrable functions on  $I$  and  $\mathcal{H}^k(I)$  the subset of functions in  $\mathcal{L}^2(I)$  whose  $k$ th derivative is still square integrable ( $k \in \mathbb{N}$ ).  $\mathcal{H}_0^1(I)$  is the subset of  $\mathcal{H}^1(I)$  of functions which are zero at the endpoints of the interval  $I$ . If  $I$  refers to the interval  $(a, b)$  it is dropped: in that case we shall write  $\mathcal{L}^2$  instead of  $\mathcal{L}^2(a, b)$ , etc. The inner product in  $\mathcal{L}^2$

is denoted by  $(\cdot, \cdot)$  and the norm by  $\|\cdot\|$ :

$$(u, v) := \int_a^b u(x)\bar{v}(x) dx, \quad \|u\| := (u, u)^{1/2}.$$

If  $\mathcal{V}$  is a subspace of  $\mathcal{L}^2(I)$ , then  $\mathcal{V}^\perp$  denotes its orthogonal complement:

$$\mathcal{V}^\perp = \{u \in \mathcal{L}^2(I) \mid (u(x), v(x)) = 0 \text{ for all } v \in \mathcal{V}\}.$$

**2. The eigenvalues and Rayleigh's quotient.** For the study of its eigenvalues problem (1.9) does not have a very suitable form, since the differential equation is not symmetric. This is amended by the transformation

$$(2.1) \quad v(x, \varepsilon) = u(x, \varepsilon)J_\varepsilon(x), \quad J_\varepsilon(x) := \exp \left\{ -\frac{1}{2\varepsilon} \int_0^x tp(t, \varepsilon) dt \right\};$$

it results in the equation

$$(2.2) \quad -\varepsilon v'' + \{x^2 p^2/4\varepsilon + xq - \frac{1}{2}p - \frac{1}{2}xp'\}v = \lambda v, \quad v(a) = v(b) = 0.$$

We recall that  $p$  and  $q$  are  $\mathcal{C}^\infty$ -functions of  $x$  and  $\varepsilon$  such that

$$(2.3) \quad p(x, \varepsilon) \geq p_0 > 0, \quad p(0, 0) = 1,$$

that because of assumption (1.2)  $J_\varepsilon$  satisfies the inequality

$$J_\varepsilon(a) \leq J_\varepsilon(b) = e^{-\Delta/(2\varepsilon)}$$

and that  $\lambda$  is a complex and  $\varepsilon$  a small positive real parameter.

Although the transformation (2.1) makes  $v$  exponentially small with respect to  $u$  for all  $x \neq 0$ , it is clear that  $u$  is an eigenfunction of (1.9) if and only if  $v$  is an eigenfunction of (2.2); hence the eigenvalues of (1.9) and (2.2) coincide. Let us denote the differential operator connected with equation (2.2) by  $T_\varepsilon$ :

$$(2.4) \quad T_\varepsilon u := -\varepsilon u'' + \{x^2 p^2/4\varepsilon + xq - \frac{1}{2}p - \frac{1}{2}xp'\}u \quad \text{for all } u \in \mathcal{H}_0^1 \cap \mathcal{H}^2.$$

It is well-known that the (symmetric) eigenvalue problem (2.2) has a denumerable set of real eigenvalues for each  $\varepsilon > 0$  and that this set is bounded from below. We shall denote the eigenvalues of (2.2) by  $\lambda_k(\varepsilon)$  with  $k \in \mathbb{N}_0$ , arranged in increasing order such that  $\lambda_{k-1} < \lambda_k$  for all  $k \in \mathbb{N}$ .

Rayleigh's quotient for problem (2.2) is the quotient

$$R_\varepsilon(u) := (T_\varepsilon u, u)/(u, u).$$

Integrating the denominator once, we see that it is defined for all  $u \in \mathcal{H}_0^1$ , provided  $u \neq 0$ . The eigenvalues of (2.2) can be computed from Rayleigh's quotient by the following minimax and maximin characterizations:

$$(2.5a) \quad \lambda_k(\varepsilon) = \inf_{\mathcal{G} \subset \mathcal{H}_0^1, \dim \mathcal{G} \geq k+1} \sup_{u \in \mathcal{G}, u \neq 0} R_\varepsilon(u),$$

$$(2.5b) \quad \lambda_k(\varepsilon) = \sup_{\mathcal{F} \subset \mathcal{L}^2, \dim \mathcal{F} \leq k} \inf_{u \in \mathcal{F}^\perp \cap \mathcal{H}_0^1, u \neq 0} R_\varepsilon(u).$$

In the minimax characterization (2.5a) the maximum of Rayleigh's quotient in a  $k+1$ -dimensional subspace is minimized over all such subspaces and in the maximin form (2.5b) the minimum of Rayleigh's quotient in the orthogonal complement of a  $k$ -dimensional subspace is maximized. The proof of these characterizations is straightforward using the (orthogonal) eigenfunctions; cf. [4, Chap. 6, § 1.4]. We remark that it

is not necessary to maximize Rayleigh's quotient in (2.5a) over all  $u \in \mathcal{E}$ ; because of linearity it suffices to maximize over all  $u \in \mathcal{E}$  satisfying  $\|u\| = t$  for some  $t > 0$ . The same is true for the minimum in (2.5b). Moreover we remark that the maximum of Rayleigh's quotient over a subspace  $\mathcal{E}$  and the minimum over the orthogonal complement of a subspace  $\mathcal{F}$  yield an upper and a lower bound for the eigenvalue under consideration for each choice of  $\mathcal{E}$  and  $\mathcal{F}$ . The bounds become better as  $\mathcal{E}$  and  $\mathcal{F}$  are better approximations of the span of the first  $k + 1$  and  $k$  eigenfunctions.

The minimum over all subspaces  $\mathcal{E}$  in (2.5a) is attained by the span of the eigenfunctions belonging to the first  $k + 1$  (counting from zero on) eigenvalues and the maximum in (2.5b) by the span of the first  $k$  eigenfunctions. If  $\Pi_\varepsilon$  is a second operator of the form (2.4), which satisfies

$$(2.6) \quad (\Pi_\varepsilon u, u) \leq (T_\varepsilon u, u) \quad \text{for all } u \in \mathcal{H}_0^1,$$

whose sets of eigenvalues and eigenfunctions are the sets

$$\{\pi_k(\varepsilon) \mid k \in \mathbb{N}_0\} \quad \text{and} \quad \{\psi_k(x, \varepsilon) \mid k \in \mathbb{N}_0\}$$

such that  $\pi_k < \pi_{k+1}$  and  $\Pi_\varepsilon \psi_k = \pi_k \psi_k$ , then we have by (2.5b):

$$(2.7) \quad \begin{aligned} \lambda_k(\varepsilon) &\cong \inf_{u \in \text{span}\{\psi_0, \dots, \psi_{k-1}\}^\perp \cap \mathcal{H}_0^1, \|u\|=1} (T_\varepsilon u, u) \\ &\cong \inf_{u \in \text{span}\{\psi_j \mid j \geq k\}, \|u\|=1} (\Pi_\varepsilon u, u) = \pi_k(\varepsilon). \end{aligned}$$

**3. Approximate eigenfunctions.** Since estimates of eigenvalues by Rayleigh's quotient require approximations of the eigenfunctions, we define the functions  $\chi_n$  by

$$(3.1) \quad \chi_n(x, \varepsilon) := \exp(-x^2/4\varepsilon) H_n(x/\sqrt{2\varepsilon}),$$

where  $H_n$  is the  $n$ th Hermite polynomial. These functions are "approximate eigenfunctions" (or better: formal approximations of the eigenfunctions). We show first that they are approximately orthogonal:

LEMMA 1. *The functions  $\chi_n$  satisfy for all  $n, m \in \mathbb{N}_0$*

$$(3.2) \quad (\chi_n, \chi_m) = (2\pi\varepsilon)^{1/2} 2^n n! \{\delta_{n,m} + \mathcal{O}(\varepsilon^{1/2-n/2-m/2} \exp(-b^2/2\varepsilon))\},$$

where  $\delta_{nm}$  is the Kronecker delta. If  $w$  is strictly positive and has a piecewise continuous first derivative, they satisfy for all  $n, m \in \mathbb{N}_0$  ( $m \leq n$ )

$$(3.3) \quad (\chi_n, w\chi_m) = w(0)(2\pi\varepsilon)^{1/2} 2^n n! \{\delta_{nm} + \mathcal{O}(\varepsilon^{1/2}(n+1))\},$$

and if  $w$  has a piecewise continuous second derivative they satisfy for all  $n, m \in \mathbb{N}$  ( $m \leq n$ ) with  $|n - m| \neq 1$

$$(3.4) \quad (\chi_n, w\chi_m) = w(0)(2\pi\varepsilon)^{1/2} 2^n n! \{\delta_{nm} + \mathcal{O}(\varepsilon n^2 + \varepsilon)\}.$$

*Proof.* The well-known recurrence relations for the Hermite polynomials imply

$$(3.5) \quad x\chi_n = (2\varepsilon)^{1/2}(n\chi_{n-1} + \frac{1}{2}\chi_{n+1}) \quad \text{and} \quad \chi_n' = (2\varepsilon)^{-1/2}(n\chi_{n-1} - \frac{1}{2}\chi_{n+1})$$

and their orthogonality on  $\mathbb{R}$  implies

$$(3.6) \quad \begin{aligned} \int_{-\infty}^{\infty} \chi_n(x, \varepsilon) \chi_m(x, \varepsilon) dx &= (2\varepsilon)^{1/2} \int_{-\infty}^{\infty} \exp(-x^2) H_n(x) H_m(x) dx \\ &= (2\pi\varepsilon)^{1/2} 2^n n! \delta_{nm}. \end{aligned}$$

Since in the left-hand side the integral over the tails  $x < a$  and  $x > b$  (with  $0 < b \leq |a|$ ) is of the order

$$\mathcal{O}(\varepsilon^{1-n/2-m/2} \exp(-b^2/2\varepsilon)),$$

this proves formula (3.2).

If the weight function  $w$  has a piecewise continuous derivative, it satisfies  $w(x) = w(0) + \mathcal{O}(x)$ , ( $x \rightarrow 0$ ); hence (3.5) and (3.6) imply

$$\begin{aligned} (w\chi_n, \chi_m) - w(0)(2\pi\varepsilon)^{1/2} n! 2^n \delta_{n,m} &= ((w - w(0))\chi_n, \chi_m) + \mathcal{O}(\cdot) \\ &= \mathcal{O}(\|x\chi_n\| \|\chi_m\|) \end{aligned}$$

and this implies (3.3). Formula (3.4) is proved in the same way; we remark that (3.4) is not true for  $|n - m| = 1$ . Q.E.D.

Next we show that  $\chi_k$  is a formal approximation of an eigenfunction of  $T_\varepsilon$ :

LEMMA 2. For every  $n, k \in \mathbb{N}_0$ ,  $n \geq k$ , the approximate eigenfunctions satisfy

$$(3.7) \quad \|T_\varepsilon \chi_n - n\chi_n\|^2 = \mathcal{O}(\varepsilon(n^3 + 1)\|\chi_n\|^2),$$

$$(3.8) \quad (T_\varepsilon \chi_n - n\chi_n, \chi_k) = \begin{cases} \mathcal{O}(\varepsilon(n^4 + 1)\|\chi_n\| \|\chi_k\|) & \text{if } n - k \neq 1, \\ \mathcal{O}(\varepsilon^{1/2}(n^3 + 1)\|\chi_n\| \|\chi_k\|) & \text{if } n - k = 1. \end{cases}$$

*Proof.* Since  $\chi_n$  satisfies the equation

$$-\varepsilon u'' + x^2 u / 4\varepsilon - \frac{1}{2}u = nu$$

we find from the recurrence relations (3.5) by straightforward calculations

$$(3.9) \quad \begin{aligned} T_\varepsilon \chi_n - n\chi_n &= (p^2 - 1)\left(\frac{1}{2}n(n-1)\chi_{n-2} + \frac{1}{4}\chi_{n+2}\right) \\ &\quad + \left\{(n + \frac{1}{2})(p^2 - 1) + \frac{1}{2}(1 - p) + x(p - \frac{1}{2}p')\right\}\chi_n. \end{aligned}$$

Since  $p = 1 + \mathcal{O}(\varepsilon) + \mathcal{O}(x)$ , Lemma 1 implies the estimates (3.7)–(3.8). Q.E.D.

*Remark 1.* Strictly speaking, the function  $\chi_n$  is not in  $\mathcal{H}_0^1$ , since it is nonzero at the endpoints  $a$  and  $b$  of the interval. However, it is of the orders  $\mathcal{O}(\varepsilon^{-n/2} \exp(-a^2/4\varepsilon))$  and  $\mathcal{O}(\varepsilon^{n/2} \exp(-b^2/4\varepsilon))$  there and we can easily amend this drawback by adding suitable boundary layer corrections. The corrected function  $\hat{\chi}_n$  is defined by

$$(3.10) \quad \begin{aligned} \hat{\chi}_n(x, \varepsilon) &:= \chi_n(x, \varepsilon) - \chi_n(b, \varepsilon)\rho(bx) \exp\{b(x-b)/2\varepsilon\} \\ &\quad - \chi_n(a, \varepsilon)\rho(ax) \exp\{a(x-a)/2\varepsilon\}, \end{aligned}$$

where  $\rho$  is an infinitely differentiable cut-off function satisfying  $\rho(x) \equiv 0$  if  $x < \frac{1}{4}$  and  $\rho(x) \equiv 1$  if  $x > \frac{3}{4}$ . The correction is of exponentially small order and can be disregarded in the computations above; more precisely we find:

$$(3.11) \quad (\hat{\chi}_n, \hat{\chi}_n) = (2\pi\varepsilon)^{1/2} 2^n n! \{\delta_{nm} + \mathcal{O}(\varepsilon^{-n/2-m/2} \exp(-b^2/2\varepsilon))\},$$

$$(3.12) \quad \|\varepsilon \hat{\chi}_n'' + (x^2/4\varepsilon - n - \frac{1}{2})\hat{\chi}_n\|^2 = \mathcal{O}(\varepsilon^{1-n} \exp(-b^2/2\varepsilon)),$$

$$(3.13) \quad (-\varepsilon \hat{\chi}_n'' + (x^2/4\varepsilon - n - \frac{1}{2})\hat{\chi}_n, \hat{\chi}_n) = tb(2b^2/\varepsilon)^n \exp(-b^2/2\varepsilon)(1 + \mathcal{O}(\varepsilon)),$$

where  $t = 1$  if  $b < -a$  and  $t = 2$  if  $b = -a$ .

*Remark 2.* Since it is expedient to have an orthogonal set of approximate eigenfunctions, we orthogonalize the set  $\{\hat{\chi}_n \mid n \in \mathbb{N}_0\}$  by the Gram–Schmidt process, resulting in the set  $\{\tilde{\chi}_n \mid n \in \mathbb{N}_0\}$ . In view of formula (2.6) this orthogonalization adds to  $\hat{\chi}_n$  only terms of the same exponentially small order, such that the Lemmas 1 and 2 remain valid if  $\chi_n$  is replaced by  $\hat{\chi}_n$  or  $\tilde{\chi}_n$ .



*Remark 3.* In view of the proof of convergence of the eigenvalues (Theorem 1) we have chosen the functions  $\chi_n$  such that they are approximate eigenfunctions for all operators of type (2.4) at once. In § 7 we shall construct approximations of higher order, which depend on the operator given.

**4. An upper bound for the eigenvalues.** In the minimax characterization (2.5a) we can use as trial space  $\mathcal{E}$  the span  $\mathcal{V}_k$  of the first  $k+1$  approximate eigenfunctions

$$(4.1) \quad \mathcal{V}_k := \text{span} \{ \tilde{\chi}_0, \tilde{\chi}_1, \dots, \tilde{\chi}_k \}$$

and for this choice we can compute an upper bound for  $\lambda_k$ .

LEMMA 3. *The  $k$ th eigenvalue  $\lambda_k(\varepsilon)$  satisfies the upper estimate*

$$(4.2) \quad \lambda_k(\varepsilon) \leq k + C_1 \varepsilon (k+1)^6$$

for some constant  $C_1$  and for all  $k \in \mathbb{N}_0$ .

*Proof.* The lowest eigenvalue satisfies by (3.8):

$$\lambda_0 \leq (T_\varepsilon \tilde{\chi}_0, \tilde{\chi}_0) \leq C_1 \varepsilon$$

for some constant  $C_1$ . As induction hypothesis we assume that the supremum of Rayleigh's quotient over  $\mathcal{V}_{k-1}$  is bounded by

$$\sup_{u \in \mathcal{V}_{k-1}, \|u\|=1} (T_\varepsilon u, u) \leq k - 1 + Ck^6 \varepsilon.$$

A function  $v \in \mathcal{V}_k$  can be written uniquely as the sum  $u + t\tilde{\chi}_k$  for some  $t \in \mathbb{C}$  such that  $u \in \mathcal{V}_{k-1}$  and  $\|v\|^2 = \|u\|^2 + \|t\tilde{\chi}_k\|^2$ . Formula (3.8) yields a constant  $C$  such that

$$(T_\varepsilon \tilde{\chi}_k, \tilde{\chi}_k) \leq (k + C\varepsilon k^4) \|\tilde{\chi}_k\|^2$$

and

$$2t(u, T_\varepsilon \tilde{\chi}_k) \leq 2tC\varepsilon^{1/2}(k^3 + 1)\|u\|\|\tilde{\chi}_k\| \leq \|u\|^2 + \varepsilon t^2 C^2(1 + k^3)^2 \|\tilde{\chi}_k\|^2.$$

Hence we can reduce the supremum of Rayleigh's quotient over  $\mathcal{V}_k$  to a supremum over  $\mathcal{V}_{k-1}$ :

$$\begin{aligned} \sup_{v \in \mathcal{V}_k} R_\varepsilon(v) &= \sup_{t \in \mathbb{C}} \sup_{u \in \mathcal{V}_{k-1}, \|u\|=1} R_\varepsilon(u + t\tilde{\chi}_k) \\ &\leq \sup_{t \in \mathbb{C}} \left\{ \sup_{u \in \mathcal{V}_{k-1}, \|u\|=1} (T_\varepsilon u, u) + 1 \right. \\ &\quad \left. + (k + \varepsilon Ck^4 + \varepsilon C^2(1 + k^3)^2) \|t\tilde{\chi}_k\|^2 \right\} / (1 + \|t\tilde{\chi}_k\|^2) \\ &\leq k + \varepsilon \tilde{C}(k+1)^6. \end{aligned}$$

This proves the estimate (4.2). Q.E.D.

**5. The differential equation of Hermite.** Before deriving a lower bound for the eigenvalues of  $T_\varepsilon$  we shall study first the eigenvalues of the particular turning point problem

$$(5.1) \quad -\varepsilon u'' + xu' = \lambda u, \quad u(a) = u(b) = 0;$$

we remark that the differential equation becomes Hermite's differential equation by the stretching  $x = \xi\sqrt{2\varepsilon}$ . By transformation (2.1) we obtain the symmetrized form

$$(5.2) \quad \Pi_\varepsilon v := -\varepsilon v'' + x^2 v / 4\varepsilon - \frac{1}{2}v = \lambda v, \quad v(a) = v(b) = 0.$$

Denoting the set of its eigenvalues by  $\{\pi_k(\varepsilon) \mid k \in \mathbb{N}_0\}$ , arranged in increasing order, we find by analogy to Lemma 3 from the estimates (3.11)–(3.13) the better upper bound for  $\pi_k(\varepsilon)$ :

LEMMA 4. *A constant  $C$  exists such that*

$$\pi_k(\varepsilon) \leq k + C\varepsilon^{-n-1/2} e^{-b^2/(2\varepsilon)}$$

for all  $\varepsilon > 0$ .

For a lower bound we apply the stretching  $x = \xi\sqrt{\varepsilon}$  to (5.2) and we obtain on the interval  $(a/\sqrt{\varepsilon}, b/\sqrt{\varepsilon})$  the eigenvalue problem

$$(5.3) \quad -\varepsilon\ddot{v} + \frac{1}{4}\xi^2 v - \frac{1}{2}v = \lambda v, \quad v(a/\sqrt{\varepsilon}) = v(b/\sqrt{\varepsilon}) = 0, \quad (\cdot = d/d\xi),$$

whose eigenvalues are identical to those of (5.2). We introduce the notations

$$(u, v)_\varepsilon := \int_{a/\sqrt{\varepsilon}}^{b/\sqrt{\varepsilon}} u(\xi)\bar{v}(\xi) d\xi,$$

$$\mathcal{H}_\varepsilon := \mathcal{H}_0^1(a/\sqrt{\varepsilon}, b/\sqrt{\varepsilon}) \quad \text{and} \quad \mathcal{L}_\varepsilon := \mathcal{L}^2(a/\sqrt{\varepsilon}, b/\sqrt{\varepsilon});$$

moreover, we continue all elements of  $\mathcal{H}_\varepsilon$  and  $\mathcal{L}_\varepsilon$  by zero outside the interval  $(a/\sqrt{\varepsilon}, b/\sqrt{\varepsilon})$ , such that we have the inclusions  $\mathcal{H}_\varepsilon \subset \mathcal{H}_\delta$  and  $\mathcal{L}_\varepsilon \subset \mathcal{L}_\delta$  provided  $0 < \delta < \varepsilon$ . Rayleigh's quotient for (5.3) is

$$Q_\varepsilon(u) := (u', u')_\varepsilon + (\frac{1}{4}\xi^2 u - \frac{1}{2}u, u)_\varepsilon, \quad u \in \mathcal{H}_\varepsilon$$

Its value does not change if (for fixed  $u \in \mathcal{H}_\varepsilon$ ) the interval of integration is enlarged, i.e. it satisfies

$$(5.4) \quad Q_\varepsilon(u) = Q_\delta(u) \quad \text{for all } u \in \mathcal{H}_\varepsilon \text{ and } \delta \in (0, \varepsilon).$$

In conjunction with the maximin characterization (2.5b) and the previous lemma we obtain

LEMMA 5. *For every  $k \in \mathbb{N}_0$  we have the inequality*

$$(5.5) \quad k \leq \pi_k(\varepsilon) \leq k + C\varepsilon^{-k-1/2} e^{-b^2/2\varepsilon}.$$

*Proof.* Assume  $0 < \delta < \varepsilon$ . If  $\mathcal{F}$  is a  $k$ -dimensional subspace of  $\mathcal{L}_\delta$  then its restriction to  $\mathcal{L}_\varepsilon$  cannot have a larger dimension; moreover, if  $u \in \mathcal{H}_\varepsilon$  is orthogonal to the restriction of  $\mathcal{F}$  to  $\mathcal{L}_\varepsilon$ , it is orthogonal to  $\mathcal{F}$  in  $\mathcal{L}_\delta$  too; hence  $\mathcal{F}^\perp \cap \mathcal{H}_\varepsilon \subset \mathcal{F}^\perp \cap \mathcal{H}_\delta$ . Consequently formula (5.4) implies that the minimum of  $Q_\delta(u)$  as  $u$  ranges over  $\mathcal{F}^\perp \cap \mathcal{H}_\delta$  cannot be larger than the minimum over  $\mathcal{F}^\perp \cap \mathcal{H}_\varepsilon$ . Taking the maxima over all these minima we find

$$\pi_k(\delta) = \sup_{\mathcal{F} \subset \mathcal{L}_\delta, \dim \mathcal{F} \leq k} \inf_{u \in \mathcal{F}^\perp \cap \mathcal{K}_\delta, u \neq 0} Q_\delta(u)$$

$$\leq \sup_{\mathcal{F} \subset \mathcal{L}_\delta, \dim \mathcal{F} \leq k} \inf_{u \in \mathcal{F}^\perp \cap \mathcal{K}_\varepsilon, u \neq 0} Q_\varepsilon(u) = \pi_k(\varepsilon);$$

hence  $\pi_k(\varepsilon)$  cannot increase as  $\varepsilon$  decreases.

In the limit for  $\varepsilon \rightarrow +0$  Rayleigh's quotient  $Q_\varepsilon(u)$  of (5.3) tends to the Rayleigh quotient of Hermite's operator (which is well-known as the "harmonic oscillator" in quantum mechanics), whose eigenvalues are known to be the nonnegative integers. This implies that  $\pi_k(\varepsilon)$  is bounded below by  $k$ . Q.E.D.

We define the function  $\psi_k(x, \varepsilon)$  to be the normalized eigenfunction of problem (5.2) associated with the eigenvalue  $\pi_k(\varepsilon)$ , i.e.

$$\Pi_\varepsilon \psi_k = \pi_k \psi_k \quad \text{and} \quad \|\psi_k\| = 1.$$

It is well-known from Sturm–Liouville theory that they form a complete orthonormal set in  $\mathcal{L}^2$ ; in conjunction with the estimates (3.11)–(3.13) this implies:

LEMMA 6. For each  $k \in \mathbb{N}_0$  the eigenvalue and eigenfunction satisfy the estimates

$$(5.6a) \quad \pi_k(\varepsilon) = k + (t/k!)(2\pi)^{-1/2} b^{2k+1} \varepsilon^{-k-1/2} \exp(-b^2/2\varepsilon)(1 + \mathcal{O}(\varepsilon)),$$

$$(5.6b) \quad \|\hat{\chi}_k - (\hat{\chi}_k, \psi_k)\psi_k\|^2 = \mathcal{O}(\varepsilon^{1-k} \exp(-b^2/2\varepsilon)).$$

*Proof.* We expand  $\hat{\chi}_k$  in the eigenfunctions of  $\Pi_\varepsilon$ ,

$$\hat{\chi}_k = \sum_{j=0}^{\infty} (\hat{\chi}_k, \psi_j)\psi_j \quad \text{and} \quad \|\hat{\chi}_k\|^2 = \sum_{j=0}^{\infty} |(\hat{\chi}_k, \psi_j)|^2.$$

Since the previous lemma implies  $|k - \pi_j(\varepsilon)| \geq \frac{1}{2}$  if  $j \neq k$ , we find from formula (3.12):

$$\begin{aligned} \|\hat{\chi}_k - (\hat{\chi}_k, \psi_k)\psi_k\|^2 &= \sum_{j=0, j \neq k}^{\infty} |(\hat{\chi}_k, \psi_k)|^2 \\ &\leq 2 \sum_{j=0}^{\infty} |(\pi_j(\varepsilon) - k)(\hat{\chi}_k, \psi_k)|^2 = \|(\Pi_\varepsilon - k)\hat{\chi}_k\|^2 \\ &= \mathcal{O}(\varepsilon^{1-n} \exp(-b^2/2\varepsilon)). \end{aligned}$$

This proves formula (5.6b); moreover, it shows that  $\|\hat{\chi}_k\|^2 - (\hat{\chi}_k, \psi_k)^2$  is of the same order, hence

$$\begin{aligned} (\Pi_\varepsilon \hat{\chi}_k, \hat{\chi}_k) &= \sum_{j=0}^{\infty} \pi_j(\varepsilon) (\hat{\chi}_k, \psi_k)^2 \\ &= \pi_k(\varepsilon) \|\hat{\chi}_k\|^2 + \mathcal{O}(\varepsilon^{1-n} \exp(-b^2/2\varepsilon)). \end{aligned}$$

In conjunction with (3.11) and (3.13) this implies (5.6a). Q.E.D.

*Remark 1.* Formula (5.6) agrees with [5, Formula (2.6)], which was derived by different means.

*Remark 2.* The estimate (5.6b) implies that Lemma 2 remains valid if  $\psi_n$  is substituted for  $\chi_n$  in the estimates (3.7)–(3.8).

**6. A lower bound for the eigenvalues.** According to the inequalities (2.6)–(2.7) the lower bound on the eigenvalues of Hermite’s operator is shared by the eigenvalues of all operators whose Rayleigh quotient is larger than Rayleigh’s quotient of Hermite’s operator. This property we shall use in order to derive a lower bound for  $\lambda_k(\varepsilon)$ .

Explicitly we have

$$(6.1) \quad (T_\varepsilon u, u) = \varepsilon \|u'\|^2 + ((x^2 p^2/4\varepsilon + xq - \frac{1}{2}p - \frac{1}{2}xp')u, u).$$

Since we assumed  $p(x, \varepsilon) = 1 + \mathcal{O}(\varepsilon) + \mathcal{O}(x)$ , the coefficient in the second term in the right-hand side has a local minimum (provided  $\varepsilon$  is small enough) at a point  $\alpha(\varepsilon)$  near  $x = 0$ , where it has the value  $-\frac{1}{2} + \beta(\varepsilon)$ ,

$$\beta(\varepsilon) := x^2 p^2/4\varepsilon + xq - \frac{1}{2}p + \frac{1}{2} - \frac{1}{2}xp' \Big|_{x=\alpha(\varepsilon)};$$

$\alpha$  and  $\beta$  are both of the order  $\mathcal{O}(\varepsilon)$  and the second derivative of the coefficient at  $\alpha(\varepsilon)$  is equal to  $1 + \mathcal{O}(\varepsilon)$ . Without loss of generality we can assume  $\alpha(\varepsilon) = 0$ , since we can shift

the  $x$ -variable over a distance  $\alpha(\varepsilon)$ ; the endpoints  $a$  and  $b$  are then shifted over the same distance, but this does not change our asymptotic estimates. Thus we find that the function  $\tilde{p}$ ,

$$\tilde{p}(x, \varepsilon) := 4\varepsilon x^{-2}(x^2 p^2/4\varepsilon + xq - \frac{1}{2}p + \frac{1}{2} - \frac{1}{2}xp' - \beta(\varepsilon)),$$

satisfies

$$\tilde{p}(x, \varepsilon) = 1 + \mathcal{O}(x) + \mathcal{O}(\varepsilon) \quad \text{and} \quad \tilde{p}(x, \varepsilon) \geq \frac{1}{2}p_0 > 0 \quad (\text{if } \varepsilon \text{ is small enough}).$$

This implies

$$\begin{aligned} (6.2) \quad (T_\varepsilon u, u) &= \varepsilon \|u\|^2 + (x^2 \tilde{p}u/4\varepsilon - \frac{1}{2}u + \beta u, u) \\ &\geq \frac{1}{2}p_0 \{ \varepsilon \|u\|^2 + (x^2 u/4\varepsilon - \frac{1}{2}u, u) \} + (\frac{1}{4}p_0 - \frac{1}{2} + \beta(\varepsilon)) \|u\|^2 \\ &= \frac{1}{2}p_0 (\Pi_\varepsilon u, u) + \mathcal{O}(\varepsilon \|u\|^2). \end{aligned}$$

Dividing by  $\|u\|^2$  we find in the right-hand side of the inequality the Rayleigh quotient of  $\Pi_\varepsilon$ . Using this estimate we can find a satisfactory lower bound for  $\lambda_k(\varepsilon)$ :

**THEOREM 1.** *For every  $k \in \mathbb{N}_0$  the eigenvalue  $\lambda_k(\varepsilon)$  of  $T_\varepsilon$  satisfies the estimate*

$$(6.3) \quad \lambda_k(\varepsilon) = k + \mathcal{O}(\varepsilon k^6 + \varepsilon).$$

*Proof.* We define the spaces  $\mathcal{V}_k$  and  $\mathcal{W}_{k,n}$  by

$$\mathcal{V}_k := \text{span} \{ \psi_j \mid j \in \mathbb{N}_0, j \geq k \} \quad \text{and} \quad \mathcal{W}_{k,n} := \text{span} \{ \psi_j \mid j \in \mathbb{N}_0, k \leq j < n \}.$$

A lower bound for  $\lambda_k(\varepsilon)$  is obtained by minimizing Rayleigh's quotient over  $\mathcal{V}_k$ , since  $\mathcal{V}_k$  is (by definition) orthogonal to a  $k$ -dimensional space. We choose  $n$  to be the smallest integer such that

$$\frac{1}{2}np_0 + \frac{1}{4}p_0 - \frac{1}{2} + \beta(\varepsilon) \geq k + 1.$$

Each  $u \in \mathcal{V}_k$  can be written as the orthogonal sum  $u = u_1 + u_2$  such that  $u_1 \in \mathcal{W}_{k,n}$  and  $u_2 \in \mathcal{V}_n$ . By Lemma 5 and (6.2) we find

$$(6.4) \quad \inf_{u_2 \in \mathcal{V}_n} R_\varepsilon(u_2) \geq \frac{1}{2}np_0 + \frac{1}{4}p_0 - \frac{1}{2} + \beta_\varepsilon \geq k + 1.$$

By analogy to formula (4.2) we can prove by induction

$$(6.5) \quad \inf_{u_1 \in \mathcal{W}_{k,n}} R_\varepsilon(u_1) \geq k - C_1 \varepsilon (k^6 + 1)$$

and Lemma 2 and the second remark following Lemma 6 imply

$$(6.6) \quad \begin{aligned} 2(T_\varepsilon u_1, u_2) &\geq -C_2 \varepsilon^{1/2} (k^3 + 1) \|u_1\| \|u_2\| \\ &\geq -C_2^2 \varepsilon (k^3 + 1)^2 \|u_1\|^2 - \|u_2\|^2. \end{aligned}$$

Formulae (6.4)–(6.6) now imply

$$\begin{aligned} \lambda_k(\varepsilon) &\geq \inf_{u \in \mathcal{V}_k, \|u\|=1} (T_\varepsilon u, u) \\ &\geq \inf_{t \in [0,1]} \inf_{\substack{u_1 \in \mathcal{W}_{k,n} \\ \|u_1\|^2 = t}} \inf_{\substack{u_2 \in \mathcal{V}_n \\ \|u_2\|^2 = 1-t}} \{ tR_\varepsilon(u_1) + (1-t)R_\varepsilon(u_2) + 2(T_\varepsilon u_1, u_2) \} \\ &\geq k - C_1 \varepsilon (k^6 + 1) - C_2 \varepsilon (k^3 + 1)^2. \end{aligned}$$

This proves the lower estimate for  $\lambda_k(\varepsilon)$ . **Q.E.D.**

COROLLARY. Let  $e_k(x, \varepsilon)$  be the normalized eigenfunction of  $T_\varepsilon$  associated with the eigenvalue  $\lambda_k(\varepsilon)$ ; then

$$(6.7) \quad \|\chi_k - (\chi_k, e_k)e_k\| = \mathcal{O}(\varepsilon^{1/2}k^{3/2}).$$

*Proof.* The proof is analogous to the proof of Lemma 6.

*Remark.* From the proof of Theorem 1 we easily derive the following stability property of the eigenvalues. If the coefficients  $p$  and  $xq - r$  of  $L_\varepsilon$  are changed by amounts which are of the orders  $\mathcal{O}(\varepsilon^{1/2}\sigma(\varepsilon))$  and  $\mathcal{O}(\sigma(\varepsilon))$  respectively uniformly in  $x$  with  $\varepsilon^{1/2}\sigma(\varepsilon) = o(1)$ ,  $\varepsilon \rightarrow +0$ , then Rayleigh's quotient  $T_\varepsilon$  and hence the eigenvalues change by the order  $\mathcal{O}(\sigma(\varepsilon))$  at most.

**7. Higher order approximations of eigenvalues and eigenfunctions.** Approximations of higher order of the eigenvalues and eigenfunctions can be computed most easily from the original nonsymmetric equation (1.9). Since the leading term of the asymptotic expansion of the  $n$ th eigenfunction  $\tilde{e}_n = J_\varepsilon^{-1}e_n$  of (1.2) is equal to  $H_n(x/\sqrt{2\varepsilon})$  (modulo a constant factor), we can choose all approximants to be polynomials in  $\varepsilon$  and  $x/\sqrt{2\varepsilon}$ ; doing so, we need not bother about the boundary conditions in view of Remark 1 in § 3. However, in order to prove that these formal computations yield the correct result, we have to apply transformation (2.1) to the approximants and to operate with the symmetric equation (2.2) as before.

In the differential equation (1.9) we introduce the substitution  $x = \xi\sqrt{2\varepsilon}$  and the (formal) asymptotic expansions

$$(7.1) \quad \begin{aligned} p(x, \varepsilon) &= 1 + \sum_{i,j=0}^{\infty} p_{ij}x^{i+1}\varepsilon^j, & q(x, \varepsilon) &= \sum_{i,j=0}^{\infty} q_{ij}x^i\varepsilon^j, \\ \tilde{e}_n(\xi/\sqrt{2\varepsilon}, \varepsilon) &= s \sum_{j=0}^{\infty} e_{nj}(\xi)\varepsilon^{j/2}, & \lambda_n(\varepsilon) &= \sum_{j=0}^{\infty} \lambda_{nj}\varepsilon^j, \end{aligned}$$

where  $e_{n,0} := H_n$ ,  $\lambda_{n,0} := n$  and  $s$  is a scaling factor. Collecting equal powers of  $\varepsilon^{1/2}$  and setting their coefficients equal to zero we obtain the recursive system of equations ( $\dot{\phantom{x}} = d/d\xi$ ):

$$(7.2) \quad \begin{aligned} \ddot{e}_{nm} - 2\xi\dot{e}_{nm} + 2ne_{nm} &= - \sum_{j=1}^{\frac{1}{2}m} 2\lambda_{nj}e_{n,m-2j} \\ &+ \sum_{i=0}^{m-1} \sum_{j=0}^{\frac{1}{2}(m-1-i)} 2\xi^{i+1} \left( p_{ij}\xi \frac{d}{d\xi} + q_{ij} \right) e_{n,m-2j-i-1}, \end{aligned}$$

with the side condition that the solution  $e_{nm}$  has to be a polynomial. Since the leading term  $e_{n,0} := H_n$  is a polynomial of degree  $n$  which is even or odd if  $n$  is even or odd, we see by induction (1), that the right-hand side of (7.2) is a polynomial of degree  $n+m$  which is even or odd if  $n+m$  is even or odd, (2) that this right-hand side can be expanded in a finite sum of Hermite polynomials, which does not contain  $H_n$  if  $m$  is odd and (3) that  $\lambda_{nm}$  can be chosen such that the coefficient of  $H_n$  in the expansion of the right-hand side is zero, if  $m$  is even. We conclude from this that for each  $m \in \mathbb{N}$  a unique scalar  $\lambda_{nm}$  exists such that the equation (7.2) has a polynomial solution (which is unique too). This procedure of solving  $e_{nm}$  and  $\lambda_{nm}$  recursively from (7.2) under the side condition that the solution has to be a polynomial is known in other contexts as the "suppression of secular terms".

In order to prove that we have obtained the correct asymptotic series for eigenvalue and eigenfunction, we apply the transformation (2.1) and we define the partial

sums  $\Lambda_{nk}$  and  $E_{nk}$  by

$$\Lambda_{nk}(\varepsilon) := \sum_{j=0}^k \lambda_{nj} \varepsilon^j,$$

$$E_{nk}(x, \varepsilon) := s \sum_{j=0}^k \varepsilon^{j/2} e_{nj}(x/\sqrt{2\varepsilon}) J_\varepsilon(x)$$

and we choose the scaling factor  $s$  such that  $\|E_{n,k}\| = 1$ . From the construction of the functions  $e_{nj}$  we see that the partial sums satisfy

$$(7.3) \quad \|(T_\varepsilon - \Lambda_{nk}(\varepsilon))E_{n,2k+1}(\cdot, \varepsilon)\| = \mathcal{O}(\varepsilon^{k+1}), \quad (\varepsilon \rightarrow +0),$$

since the remainder is a polynomial in  $x/\sqrt{2\varepsilon}$  of degree  $n + 2k + 1$  multiplied by  $\varepsilon^{k+1}$  and by the exponential. Expanding  $E_{n,2k+1}$  in the set of orthonormal eigenfunctions  $\{e_j | j \in \mathbb{N}_0\}$  of  $T_\varepsilon$ ,

$$E_{n,2k+1} = \sum_{j=0}^{\infty} \gamma_{nkj} e_j \quad \text{with} \quad \sum_{j=0}^{\infty} |\gamma_{nkj}|^2 = \|E_{n,2k+1}\|^2,$$

we find by Theorem 1:

$$\begin{aligned} \|(T_\varepsilon - \Lambda_{nk})E_{n,2k+1}\|^2 &= \sum_{j=0}^{\infty} |\lambda_j - \Lambda_{nk}|^2 |\gamma_{nkj}|^2 \\ &\leq 2 \sum_{j=0, j \neq n}^{\infty} |\gamma_{nkj}|^2 + |\lambda_n - \Lambda_{nk}|^2 |\gamma_{nkj}|^2 = \mathcal{O}(\varepsilon^{2k+2}). \end{aligned}$$

Since  $\|E_{n,2k+1}\|$  is of order unity this implies

$$(7.4) \quad \lambda_n(\varepsilon) = \Lambda_{nk}(\varepsilon) + \mathcal{O}(\varepsilon^{k+1}) \quad \text{and} \quad \|E_{n,2k+1} - (E_{n,2k+1}, e_j)e_j\| = \mathcal{O}(\varepsilon^{k+1})$$

for all  $k, n \in \mathbb{N}_0$ . Since each  $u \in \mathcal{H}^1$  satisfies (Sobolev)

$$(7.5a) \quad \max_{a < x < b} |u(x)|^2 \leq 2\|u\|\|u'\| + 2\|u\|^2/(b-a)$$

and since a positive constant  $C$  exists, such that

$$(7.5b) \quad \|u'\|^2 \leq 4\varepsilon^{-1}\|u\|\{\|T_\varepsilon u - \lambda u\| + (C/\varepsilon + |\lambda|)\|u\|\}$$

for all  $u \in \mathcal{H}^2$  and for all  $\lambda \in \mathbb{C}$ , cf. [6, chap. 2], the estimate of the error in  $E_{n,2k+1}$  is valid in the maximum norm too. Summing up we have derived:

**THEOREM 2.** *The eigenvalues and eigenfunctions  $\lambda_n(\varepsilon)$  and  $e_n(x, \varepsilon)$  of the operator  $T_\varepsilon$  have for  $\varepsilon \rightarrow +0$  the asymptotic series expansions*

$$(7.6) \quad \lambda_n(\varepsilon) = n + \sum_{j=1}^{\infty} \varepsilon^j \lambda_{nj},$$

$$(7.7) \quad e_n(x, \varepsilon) = s J_\varepsilon(x) \left\{ H_n(x/\sqrt{2\varepsilon}) + \sum_{j=1}^{\infty} \varepsilon^{j/2} e_{nj}(x/\sqrt{2\varepsilon}) \right\},$$

where the coefficients are determined recursively from the system of equations (7.2). Explicit computation shows

$$\lambda_{n\varepsilon} = n + \varepsilon \{3n^2 p_{10} + (2n+1)q_{10} - 12n^2 p_{00}^2 - (12n+2)p_{00}q_{00} - 2q_{00}^2\} + \mathcal{O}(\varepsilon^2).$$

The formal series expansion of  $\tilde{e}_n$  in (7.1), from which (7.7) is derived, is not asymptotic in the whole interval  $[a, b]$ . Since the  $j$ th coefficient  $e_{nj}$  is a polynomial in

$\xi = x/\sqrt{2\varepsilon}$  of degree  $n+j$ , the  $j$ th term is of the order  $\mathcal{O}(s\varepsilon^{-n/2}x^{n+j})$  and hence all terms are of the same order of magnitude for fixed  $x \neq 0$  and for  $\varepsilon \rightarrow +0$ . In (7.7) it is the exponential factor  $J_\varepsilon$  that makes the series asymptotic. The formal series expansion of  $\tilde{e}_n$  is asymptotic only in an  $\varepsilon$ -dependent neighborhood of the point  $x=0$  whose diameter shrinks to zero for  $\varepsilon \rightarrow +0$ . Theorem 2 implies that this series is asymptotically correct in a neighborhood whose diameter is of the order  $\mathcal{O}(\varepsilon^{1/2})$  only.

For a better approximation of  $\tilde{e}_n$  outside a neighborhood of  $x=0$  we construct the regular expansions in the subdomains  $[a, -\varepsilon^\delta]$  and  $[\varepsilon^\delta, b]$  for some  $\delta \in (0, \frac{1}{2}]$ . In these regions we expand  $\tilde{e}_n$  and the coefficients of the differential equation into the formal power series

$$\tilde{e}_n(x, \varepsilon) = \sum_{k=0}^{\infty} \varepsilon^k v_{nk}, \quad p(x, \varepsilon) = \sum_{k=0}^{\infty} \varepsilon^k p_k(x), \quad q(x, \varepsilon) = \sum_{k=0}^{\infty} \varepsilon^k q_k(x);$$

substituting them in the differential equation and collecting equal powers of  $\varepsilon$  we obtain the system of equations

$$(7.8) \quad (xp_0 d/dx + xq_0 - n)v_{nk} = v''_{n,k-1} - \sum_{j=1}^k (xp_j d/dx + xq_j - \lambda_{nj})v_{n,k-j}.$$

The constants of integration are obtained from matching to the inner expansion obtained before. The lowest order term  $v_{n,0}$  is

$$v_{n,0}(x) = c_{n,0} x^n \exp \left\{ \int_0^x (n - np_0(t) - tq_0(t)) dt / tp_0(t) \right\};$$

because  $p_0(0) = 1$  this function is  $\mathcal{C}^\infty$  and satisfies

$$(7.9) \quad v_{n,0}(x) = c_{n,0} x^n (1 + \mathcal{O}(x)) \quad (x \rightarrow 0).$$

For matching we substitute the intermediate variable  $\zeta := x\varepsilon^{-\delta} = \xi\varepsilon^{1/2-\delta}$  with  $\delta \in (0, \frac{1}{2})$  in both expansions for  $\tilde{e}_n$  and we expand both series again into powers of  $\varepsilon$ . Since the leading terms of both series must agree, we find

$$(7.10) \quad c_{n,0} \varepsilon^{n\delta} \zeta^n = s 2^{1/2n} \zeta^n \varepsilon^{n\delta-n/2} \Rightarrow c_{n,0} = 2^{n/2} s \varepsilon^{n/2}.$$

The regular expansion of  $\tilde{e}_n$  is matched to the boundary conditions  $\tilde{e}_n(a, \varepsilon) = \tilde{e}_n(b, \varepsilon) = 0$  in ordinary boundary layers. In the boundary layer at  $x=b$  we substitute the local variable  $\theta := \sigma_b(x)/\varepsilon$ , where  $\sigma_b$  is  $\mathcal{C}^\infty$  and satisfies

$$(7.11a) \quad \sigma'_b > 0 \quad \text{and} \quad \sigma_b(x) = x - b + \mathcal{O}((x-b)^2) \quad \text{for } x \rightarrow b,$$

and we expand the solution and the coefficients of the differential equation in (formal) power series in  $\varepsilon$ :

$$\tilde{e}_n(x, \varepsilon) = \sum_{j=0}^{\infty} \varepsilon^j w_j(\theta), \quad xp(x, \varepsilon) = \sum_{j=0}^{\infty} \tilde{p}_j(\theta) \varepsilon^j, \quad xq(x, \varepsilon) = \sum_{j=0}^{\infty} \tilde{q}_j(\theta) \varepsilon^j.$$

This results in the system of differential equations

$$(7.11b) \quad -w''_{nk} + \tilde{p}_0 w'_{nk} = - \sum_{j=1}^k (\tilde{p}_j d/d\theta + \tilde{q}_{j-1} + \lambda_{n,j-1}) w_{n,k-j} \quad ({}' = d/d\theta)$$

with the matching conditions

$$(7.11c) \quad w_{nk}(0) = -v_{nk}(b) \quad \text{and} \quad \lim_{\theta \rightarrow -\infty} w_{nk}(\theta) = 0.$$

Since  $\tilde{p}_0 = bp(b, 0)$ , the lowest order term of this expansion is

$$w_{n0}(\theta) = -v_{n0}(b) \exp\{bp(b, 0)\theta\}.$$

*Remark.* We could have chosen  $\theta = (x - b)/\varepsilon$  as the boundary layer variable; however, in order to have a better control over the decay of the boundary layer correction  $w$  in a neighborhood of the boundary layer we prefer to have some extra freedom in  $\theta$ . Outside the boundary layer we cut the correction off multiplying it by  $\rho(bx)$ , where  $\rho$  is a  $\mathcal{C}^\infty$ -function satisfying  $\rho(x) \equiv 0$  if  $x < \frac{1}{4}$  and  $\rho(x) \equiv 1$  if  $x > \frac{3}{4}$ .

In the boundary layer at  $x = a$  we construct analogously the formal expansion  $e_n = \sum \varepsilon^i \hat{w}_{nj}$ ; clearly we find

$$\hat{w}_{n0}(\eta) = -v_{n0}(a) \exp\{-ap(a, 0)\eta\}, \quad \varepsilon\eta := \sigma_a(x) = x - a + \mathcal{O}((x - a)^2).$$

Thus we have constructed a formal approximation for the  $n$ th eigenfunction  $\tilde{e}_n$  of (1.2). The lowest order term of this approximation is  $F_{n0}$ ,

$$F_{n0}(x, \varepsilon) := sH_n(x/\sqrt{2\varepsilon}) + v_{n0}(x) - c_{n0}x^n + \rho(bx)w_{n0}(\sigma_b(x)/\varepsilon) + \rho(ax)\hat{w}_{n0}(\sigma_a(x)/\varepsilon);$$

the term  $c_{n0}x^n$  is subtracted since it is contained in  $sH_n$  and in  $v_{n0}$  and it is counted twice otherwise. We shall prove the validity of this approximation with the aid of the following consequence of the maximum principle:

LEMMA 7. *Let  $n \in \mathbb{N}$  and  $r \in \mathbb{R}$  satisfy  $r \leq n$  and let  $m \in \mathbb{R}$  be larger than the largest zero of  $H_n(x/\sqrt{2})$ . If a constant  $M$  exists such that the function  $z$  satisfies*

$$(7.12a) \quad |-\varepsilon z'' + xpz' + xqz - rz| \leq M\varepsilon^{-n/2}x^n \quad \text{for all } x \in [m\varepsilon^{1/2}, b],$$

$$(7.12b) \quad |z(\varepsilon^{1/2}m)| \leq Mm^n \quad \text{and} \quad |z(b)| \leq M\varepsilon^{-n/2}b^n,$$

then a constant  $N$  exists such that

$$(7.12c) \quad |z(x)| \leq \begin{cases} N\varepsilon^{-n/2}x^n & \text{if } r \neq n, \\ N\varepsilon^{-n/2}x^n |\log \varepsilon| & \text{if } r = n, \end{cases}$$

for all  $x \in [m\varepsilon^{1/2}, b]$ .

*Proof.* We choose the barrier function  $W_r$ :

$$(7.13a) \quad \begin{aligned} W_r(x, \varepsilon) &:= sH_n(x/\sqrt{2\varepsilon}) + v_{n0}(x) - c_{n0}x^n & \text{if } n \neq r, \\ W_n(x, \varepsilon) &:= (sH_n(x/\sqrt{2\varepsilon}) + v_{n0}(x) - c_{n0}x^n) \log(2\varepsilon^{-1/2}x/M). \end{aligned}$$

From the computations above we easily find positive constants  $d$  and  $D$  such that

$$(7.13b) \quad dx^n \varepsilon^{-n/2} \leq W_r(x, \varepsilon) \leq \begin{cases} sDx^n \varepsilon^{-n/2} |\log \varepsilon| & \text{if } r = n, \\ sDx^n \varepsilon^{-n/2} & \text{if } r < n, \end{cases}$$

$$(7.13c) \quad (L_\varepsilon - r)W_r \geq \begin{cases} sdx^n \varepsilon^{-n/2} & \text{if } r = n, \\ sd(n - r)x^n \varepsilon^{-n/2} & \text{if } r < n, \end{cases}$$

provided  $\varepsilon$  is sufficiently small and  $\varepsilon^{1/2}m \leq x \leq b$ . According to the maximum principle it follows from (7.12a) and (7.13c) that  $(-MW_n \pm sdz)/W_n$  cannot have positive maxima in  $(\varepsilon^{1/2}m, b)$ . Since (7.12b) and (7.13b) imply that they are negative at  $x = \varepsilon^{1/2}m$  and at  $x = b$ , they are negative everywhere. If  $r \neq n$  we use the same argument. Q.E.D.

THEOREM 3. *A constant  $C$  exists, such that the  $n$ -th eigenfunction  $\tilde{e}_n$  of problem (1.2) satisfies the estimate*

$$(7.14) \quad |\tilde{e}_n(x, \varepsilon) - F_{n0}(x, \varepsilon)| \leq sC(1 + x^n \varepsilon^{-n/2})\varepsilon^{1/2} |\log \varepsilon|$$

uniformly for all  $x \in [a, b]$ .



*Proof.* Theorem 2 and formula (7.9) imply that for each  $m > 0$  a constant  $C_m$  exists, such that

$$|\tilde{e}_n(x, \varepsilon) - F_{n0}(x, \varepsilon)| \leq C_m \varepsilon^{1/2}, \quad \text{provided } |x| \leq m\varepsilon^{1/2};$$

moreover, since  $e_n(b, \varepsilon) = F_{n0}(b, \varepsilon) = 0$ , condition (7.12b) is satisfied. From the construction of the approximation it follows that

$$(L_\varepsilon - \lambda_n)(\tilde{e}_n - F_{n0} - \varepsilon w_{n1}) = \mathcal{O}(s\varepsilon^{1/2-n/2} x^n)$$

and that  $w_{n1}$  is of the same order as  $w_{n0}$  is; hence to the subinterval  $(\varepsilon^{1/2}m, b)$  we can apply the previous lemma (with  $r = n$ ). To the subinterval  $(a, -\varepsilon^{1/2}m)$  we can apply the same argument. Q.E.D.

In order to compute higher order terms of the expansion of  $\tilde{e}_n$  we must solve (7.8) (and (7.11), but this is well-known) recursively and match each term to the inner expansion by ‘‘intermediate matching’’; cf. Eckhaus [12]. Having computed the regular expansion up to the index  $j-1$ , we must verify that the  $j$ th equation has a solution which is  $\mathcal{C}^\infty$  at  $x = 0$ ; this is guaranteed by the fact that the coefficient of  $x^n$  in the Taylor series expansion at  $x = 0$  of the right-hand side in the equation (7.8) is made zero by the choice of  $\lambda_{nk}$  in (7.2); otherwise the solution would contain a term of the order  $\mathcal{O}(x^n \log x)$  ( $x \rightarrow 0$ ). For the matching we substitute the intermediate variable  $\zeta = x\varepsilon^{-\delta} = \xi\varepsilon^{1/2-\delta}$  with  $0 < \delta < \frac{1}{2}$  in  $\sum_{k=0}^{\infty} s\varepsilon^{k/2} e_{nk}$  and in  $\sum_{k=0}^j \varepsilon^k v_{nk}$  and we expand the new series in powers of  $\varepsilon$  up to the order  $\mathcal{O}(s\varepsilon^{j-n\delta})$ ; the constant of integration, which is in the term of the order  $\mathcal{O}(s\varepsilon^{j-n\delta})$  is now determined by the condition that both series must agree up to this order. The proof of validity is analogous to the proof given above.

The approximation for  $\tilde{e}_n$ , we have constructed, is such that the *relative* error is uniform outside the boundary layers, i.e. if  $a + m\varepsilon < x < -\varepsilon^{1/2}M$  and if  $\varepsilon^{1/2}M < x < b - m\varepsilon$  for sufficiently large constants  $M$  and  $m$ . Hence we obtain by transformation (2.1) an approximation of  $e_n$  with a good relative error, which is better than (7.7) is. However, its Rayleigh quotient does not yield a better approximation of the corresponding eigenvalue, since it differs from (7.6) by exponentially small terms only, which are too small to be proved correct, unless the asymptotic series happens to converge. In Lemma 6 we have given an example in which the dominant asymptotic series of the eigenvalues terminates, such that exponentially small terms can be computed.

**8. Exponential decay and resonance.** Having established the conditions under which the solution of the boundary value problem (1.1) exists and is unique, we can study the asymptotic behavior of this solution.

The construction of a formal asymptotic approximation to the solution  $U_\varepsilon$  of (1.1) is analogous to the construction of the approximation of  $\tilde{e}_n$  in the preceding section. Now we assume that the inner and the regular expansions are zero and hence that the approximation consist of boundary layer terms only. As in (7.11) we substitute in the boundary layer at  $x = b$  the local variable  $\theta := \sigma_b(x)/\varepsilon$  and we expand everything in formal power series in  $\varepsilon$ :

$$U_\varepsilon(x) = \sum \varepsilon^j z_j(\theta), \quad xp = \sum \varepsilon^j \tilde{p}_j, \quad xq = \sum \varepsilon^j \tilde{q}_j, \quad r(\varepsilon) = \sum \varepsilon^j r_j.$$

Hence, we obtain the system of differential equations

$$(8.1) \quad -z_k'' + \tilde{p}_0 z_k' = - \sum_{j=1}^k (\tilde{p}_j d/d\theta + \tilde{q}_{j-1} - r_{j-1}) z_{k-j}$$

with the boundary conditions

$$z_0(b) = B, \quad z_k(b) = 0 \quad (k \geq 1) \quad \text{and} \quad \lim_{\theta \rightarrow -\infty} z_k(\theta) = 0 \quad (k \geq 0).$$

The lowest order term is

$$(8.2) \quad z_0(\theta) = B \exp\{bp(b, 0)\theta\}$$

and higher order terms are computed easily; since  $p_j$  and  $q_j$  are polynomials in  $\xi$  of degree  $j$ ,  $z_k$  is equal to a polynomial in  $\xi$  of degree  $2k$  multiplied by  $\exp(bp(b, 0)\theta)$  and constants  $C_k$  exists such that each partial sum satisfies for all  $\theta \leq 0$ :

$$(8.3) \quad \left| (L_\varepsilon - r(\varepsilon)) \sum_{j=0}^k \varepsilon^j z_j(\theta) \right| \leq \varepsilon^k C_k |B| (1 + \theta^{2k}) \exp(bp(b, 0)\theta).$$

In the same way we construct at  $x = a$  the boundary layer expansion

$$(8.4) \quad \begin{aligned} U_\varepsilon(x) &= \sum_{j=0}^{\infty} \varepsilon^j \hat{z}_j(\eta) \quad \text{with } \varepsilon\eta := \sigma_a(x) = x - a + \mathcal{O}((x-a)^2), \\ \hat{z}_0(\eta) &= A \exp\{ap(a, 0)\eta\}, \end{aligned}$$

which satisfies an estimate analogous to (8.6). So we have constructed the formal approximation  $Z_\varepsilon^k$  of  $U_\varepsilon$ :

$$(8.5) \quad Z_\varepsilon^k(x) := \sum_{j=0}^k \varepsilon^j (\rho(bx)z_j(x) + \rho(ax)\hat{z}_j(x)),$$

where  $\rho$  is a  $\mathcal{C}^\infty$  cut-off function ( $\rho(x) = 0$  if  $x < \frac{1}{4}$  and  $\rho(x) \equiv 1$  if  $x > \frac{3}{4}$ ). Exploiting the relation  $T_\varepsilon J_\varepsilon u = J_\varepsilon L_\varepsilon u$  between  $T_\varepsilon$  and  $L_\varepsilon$  and the eigenfunction expansion of  $T_\varepsilon$  we prove the validity of this formal approximation:

**THEOREM 4.** *Let  $n \in \mathbb{N}_0$  be the nonnegative integer that is nearest to  $r(0)$  and let  $U_\varepsilon$  be the solution of problem (1.1). The formal approximation  $Z_\varepsilon^k$  satisfies*

$$(8.6) \quad U_\varepsilon(x) = Z_\varepsilon^k(x) + \frac{\tilde{e}_n(x, \varepsilon)}{\lambda_n(\varepsilon) - r(\varepsilon)} \{Bbp(b)J_\varepsilon^2(b)v_{n0}(b) + A|a|p(a)J_\varepsilon^2(a)v_{n0}(a)\} (1 + \mathcal{O}(\sqrt{\varepsilon}))$$

$$+ \begin{cases} \mathcal{O}(A\varepsilon^k J_\varepsilon(a) + \mathcal{O}(B\varepsilon^k J_\varepsilon(b)J_\varepsilon^{-1}(x))) & \text{if } x \geq 0, \\ \mathcal{O}(B\varepsilon^k J_\varepsilon(b) + \mathcal{O}(A\varepsilon^k J_\varepsilon(a)J_\varepsilon^{-1}(x))) & \text{if } x \leq 0, \end{cases}$$

where  $\tilde{e}_n = J_\varepsilon^{-1} e_n$  is the  $n$ -th eigenfunction of problem (1.9) and where  $v_{n0}$  is the lowest order term of the regular expansion of  $\tilde{e}_n$ , cf. (7.9),

$$v_{n0}(x) = (n! \sqrt{2\pi\varepsilon})^{-1/2} \varepsilon^{-n/2} x^n \exp \left\{ \int_0^x (n - np(t, 0) - tq(t, 0)) dt / tp(t, 0) \right\} (1 + \mathcal{O}(\varepsilon))$$

for  $x \neq 0$  and  $\varepsilon \rightarrow +0$ .

*Proof.* Let  $U_\varepsilon^B$  be the solution of (1.1) if  $A = 0$ . The construction (8.1) implies that the error  $D_\varepsilon^k$ ,

$$D_\varepsilon^k(x) := U_\varepsilon^B(x) - \sum_{j=0}^k \varepsilon^j \rho(bx)z_j(\sigma_b(x)/\varepsilon)$$

is an element of  $\mathcal{H}^1 \cap \mathcal{H}^2$ . Hence,  $J_\varepsilon D_\varepsilon^k$  can be expanded in the eigenfunctions of  $T_\varepsilon$  and

its component orthogonal to  $e_n$  satisfies by formula (8.3) and Theorem 1:

$$\begin{aligned} \|J_\varepsilon D_\varepsilon^k - (J_\varepsilon D_\varepsilon^k, e_n)e_n\|^2 &= \sum_{j=0, j \neq n}^{\infty} |(J_\varepsilon D_\varepsilon^k, e_j)|^2 \\ &\leq \sum_{j=0, j \neq n}^{\infty} |(J_\varepsilon(L_\varepsilon - r)D_\varepsilon^k, e_j)/(\lambda_j - r)|^2 \\ &\leq \|J_\varepsilon(L_\varepsilon - r)D_\varepsilon^k\|^2 = \mathcal{O}(\varepsilon^{2k+1} J_\varepsilon^2(b)). \end{aligned}$$

Sobolev's inequality (7.5) now implies existence of a constant  $C$  such that

$$|J_\varepsilon(x)D_\varepsilon^k(x) - (J_\varepsilon D_\varepsilon^k, e_n)e_n(x, \varepsilon)| \leq C\varepsilon^k J_\varepsilon(b)$$

for all  $x \in [a, b]$ . In particular, this is true if  $x \leq e^{1/2}m$  for some  $m \in \mathbb{R}$ , where we have  $J_\varepsilon(e^{1/2}m) = \mathcal{O}(1)$  for  $\varepsilon \rightarrow +0$ ; since we also have  $(L_\varepsilon - r)D_\varepsilon^k = 0$  for  $x \leq 0$  we can apply Lemma 7 to the restriction of  $D_\varepsilon^k$  to  $[a, -\varepsilon^{1/2}m]$ . Hence, the component of  $D_\varepsilon^k$  orthogonal to  $J_\varepsilon e_n$  satisfies the estimate

$$(8.7) \quad D_\varepsilon^k(x) - (D_\varepsilon^k, J_\varepsilon e_n)e_n(x, \varepsilon) = \begin{cases} \mathcal{O}(B\varepsilon^k J_\varepsilon(b) J_\varepsilon^{-1}(x)) & \text{if } x \geq 0, \\ \mathcal{O}(B\varepsilon^k J_\varepsilon(b)) & \text{if } x \leq 0, \end{cases}$$

uniformly for all  $x \in [a, b]$ .

In order to compute the inner product  $(J_\varepsilon D_\varepsilon^k, e_n)$  we choose the function  $\sigma_b$  in the boundary layer variable as follows:

$$(8.8a) \quad \sigma_b(x) = x - b - \mu(x - b)^2 \quad \text{with } \mu = \nu - \frac{1}{2}(p(b, 0) + bp'(b, 0))/bp(b, 0).$$

If  $\nu$  is a sufficiently large positive number this implies

$$(8.8b) \quad \int_x^b tp(t, 0) dt + bp(b, 0)\sigma_b(x) = -\nu(x - b)^2 + \mathcal{O}((x - b)^3) < 0$$

for all  $x \in [0, b]$ . Hence, we find by (7.14)

$$\begin{aligned} (8.9) \quad (J_\varepsilon D_\varepsilon^k, e_n) &= -(J_\varepsilon(L_\varepsilon - r)z_0(\sigma_b/\varepsilon), e_n)/(\lambda_n(\varepsilon) - r(\varepsilon))(1 + \mathcal{O}(\varepsilon)) \\ &= -\frac{Bbp(b)J_\varepsilon^2(b)v_{n0}(b)}{\varepsilon(\lambda_n(\varepsilon) - r(\varepsilon))} \int_a^b e^{-\nu(x-b)^2/\varepsilon} \{2\nu(x-b) + \mathcal{O}(\varepsilon + |x-b|^2)\} dx \\ &= Bbp(b)J_\varepsilon^2(b)v_{n0}(b)/(\lambda_n(\varepsilon) - r(\varepsilon))(1 + \mathcal{O}(\sqrt{\varepsilon})). \end{aligned}$$

For the solution  $U_\varepsilon^A$  of (1.1) with  $B = 0$  we can derive estimates analogous to (8.7) and (8.9); since  $U_\varepsilon = U_\varepsilon^A + U_\varepsilon^B$ , this implies formula (8.6). Q.E.D.

*Remark.* In fact we have used in the proof the generalized eigenfunction expansion in the biorthogonal series  $\{J_\varepsilon e_n\}$  and  $\{J_\varepsilon^{-1} e_n\}$  of eigenfunctions of  $L_\varepsilon$  and its adjoint  $L_\varepsilon^*$ .

This theorem gives all information we want about the solution  $U_\varepsilon$ . We see from (8.6) and (7.14) that  $U_\varepsilon$  decays exponentially fast in the interior of the interval if the distance between  $r(\varepsilon)$  and the nearest eigenvalue of  $T_\varepsilon$  satisfies condition (1.4). Moreover, it gives a good estimate of the magnitude (and the form) of the resonance and it displays exactly how the resonant part of the solution explodes if  $r(\varepsilon)$  approaches the eigenvalue sufficiently fast. Unfortunately it is in general not possible to determine exponentially small terms in the asymptotic expansion of  $\lambda_n(\varepsilon)$ ; hence, in general it remains unknown whether or not the denominator  $\lambda_n(\varepsilon) - r(\varepsilon)$  in (8.6) is smaller than the numerator.

In the special case of the Hermite operator (5.1) the exact solution can be determined, e.g. in confluent hypergeometric functions. Its asymptotic expansion agrees with formulae (5.6a) and (8.6); cf. [5, formula (2.7a, b, c, d)].

Another example in which we can approximate accurately the resonant solution occurs near the smallest eigenvalue, when the coefficient  $q$  is equal to zero. In the particular eigenvalue problem

$$(8.10) \quad -\varepsilon u'' + xpu' = \lambda u, \quad u(a) = u(b) = 0$$

the inner and regular expansions of  $\tilde{e}_0$  reduce to only one term, namely  $\tilde{e}_0 = \text{constant}$ . By Theorem 3 we then find the uniform approximation  $\tilde{e}_0 = F_{00}(1 + \mathcal{O}(\varepsilon))$ , where

$$F_{00}(x, \varepsilon) = s\{1 - \rho(bx) \exp(bp(b, 0)\sigma_b(x)/\varepsilon) - \rho(ax) \exp(ap(a, 0)\sigma_a(x)/\varepsilon)\}.$$

Rayleigh's quotient of  $J_\varepsilon F_{00}$  is (by analogy to (8.9))

$$\begin{aligned} (T_\varepsilon J_\varepsilon F_{00}, J_\varepsilon F_{00}) / \|J_\varepsilon F_{00}\|^2 &= (2\pi\varepsilon)^{-1/2} (L_\varepsilon F_{00}, J_\varepsilon^2 F_{00}) (1 + \mathcal{O}(\sqrt{\varepsilon})) \\ &= (2\pi\varepsilon)^{-1/2} \{bp(b)J_\varepsilon^2(b) + |a|p(a)J_\varepsilon^2(a)\} (1 + \mathcal{O}(\sqrt{\varepsilon})), \end{aligned}$$

if the functions  $\sigma_a$  and  $\sigma_b$  in the boundary layer variable are chosen as in (8.8). Since  $F_{00}$  satisfies

$$\|T_\varepsilon J_\varepsilon F_{00}\|^2 = \|J_\varepsilon L_\varepsilon F_{00}\|^2 = \mathcal{O}(\varepsilon J_\varepsilon^2(b)) = \mathcal{O}(\varepsilon^{1/2} J_\varepsilon^2(b) \|F_{00}\|^2),$$

we find from the eigenfunction expansion of  $J_\varepsilon F_{00}$

$$(8.11) \quad \lambda_0(\varepsilon) = (2\pi\varepsilon)^{-1/2} \{bp(b)J_\varepsilon^2(b) + |a|p(a)J_\varepsilon^2(a)\} (1 + \mathcal{O}(\sqrt{\varepsilon}))$$

in the same way as in Lemma 6. By formula (8.6) we find for the solution  $U_\varepsilon$  of the boundary value problem

$$\varepsilon u'' + xp(x, \varepsilon)u' = 0, \quad u(a) = A, \quad u(b) = B$$

the result

$$(8.12a) \quad U_\varepsilon(x) = B + (B - A) \exp\{ap(a, 0)(x - a)\} + \mathcal{O}(\sqrt{\varepsilon}),$$

provided

$$\int_0^a tp(t, 0) dt > \int_0^b tp(t, 0) dt,$$

cf. (1.2), and

$$(8.12b) \quad \begin{aligned} U_\varepsilon(x) &= \{bp(b)B - ap(a)A + ap(a)(A - B) \exp(bp(b, 0)(x - b)) \\ &\quad + bp(b)(A - B) \exp(ap(a, 0)(x - a))\} / (bp(b) - ap(a)) + \mathcal{O}(\sqrt{\varepsilon}). \end{aligned}$$

provided both integrals are equal.

### 9. Generalizations and related problems.

a. Imposing on problem (1.1) the condition “ $p$  strictly negative” instead of “ $p$  positive” we obtain a problem which is intimately related to problem (1.1). Such a type of problem is represented by the adjoint of (1.1a)

$$(9.1a) \quad L_\varepsilon^* u := -\varepsilon u'' - xpu' + (xq - p - xp')u = ru.$$

$$(9.1b) \quad u(a) = A \quad \text{and} \quad u(b) = B.$$

Clearly its eigenvalues are equal to the eigenvalues of (1.2) and the eigenfunction

connected to  $\lambda_k(\varepsilon)$  is  $J_\varepsilon e_k$ . If  $r(0) \neq n$  the solution  $u_\varepsilon$  of (9.1) satisfies

$$(9.2) \quad u_\varepsilon(x) = \begin{cases} A \exp \left\{ \int_a^x w(t) dt \right\} (1 + \mathcal{O}(\varepsilon x^{-2})), & \text{if } x < 0, \\ B \exp \left\{ \int_b^x w(t) dt \right\} (1 + \mathcal{O}(\varepsilon x^{-2})), & \text{if } x > 0, \end{cases}$$

where  $w(t) := \{tq(t, 0) - p(t, 0) - tp'(t, 0) - r(0)\} / tp(t, 0)$ ; cf. [7, Thm. 3.15]. If  $r(0) = n$ , we have to add a multiple of  $J_\varepsilon e_n / ((\lambda_n(\varepsilon)) - r(\varepsilon))$  as before. Due to the exponentially decaying nature of  $J_\varepsilon$  this resonant part is dominant only in a subinterval (containing  $x = 0$ ) whose diameter depends on the magnitude of  $1/|\lambda_n - r|$ ; if  $1/|\lambda_n - r| = \mathcal{O}(\varepsilon^{-\beta})$  for some  $\beta > 0$ , then the diameter of this subinterval is of the order  $\mathcal{O}(\varepsilon^{1/2} \log \varepsilon)$ .

**b.** We can add to the differential equations (1.1) and (9.1) an inhomogeneous term  $f$  and construct an asymptotic approximation to the solution, provided  $r(0)$  is not equal to the limit of an eigenvalue.

In (9.1) the leading term of the outer expansion is the solution of the reduced equation, which satisfies the boundary values at  $a$  and  $b$ . In order to prove convergence for  $r(0) > n \geq 0$  we have to embed the problem in the negative Sobolev space  $\mathcal{H}^{-n-1}$  and to prove first convergence in weak sense; afterwards we can show convergence in stronger sense by interpolation; cf. [5] and [6].

In (1.1) the leading term of the outer expansion is that solution of the reduced equation that is continuous at  $x = 0$ . This solution is an analytic function of  $r(0)$  which can be continued analytically in the positive halfplane up to the line  $\Re r(0) = n$ , provided  $f$  has  $n$  derivatives at  $x = 0$  and which has poles at the points  $r(0) = k \in \mathbb{N}_0$  (this continuation is the smoothest solution of the reduced equation). In order to prove convergence for  $r(0) > n \geq 0$  we have to restrict the problem to the positive Sobolev space  $\mathcal{H}^{n+1}$  (i.e. to prove convergence of the  $n$ th derivative first); cf. [6] and [1]. Alternatively we can use the technique by which Theorem 4 has been proved: transform the error by (2.1), expand it in the eigenfunctions of  $T_\varepsilon$  resulting in a max-norm estimate in an  $\mathcal{O}(\varepsilon^{1/2})$ -neighborhood around  $x = 0$  and apply Lemma 7 for an estimate on the remaining part of the interval.

**c.** If a turning point is located at the boundary point  $a$ , the boundary condition  $u(a) = 0$  eliminates the approximate eigenfunctions which have an even index and hence it also eliminates the associated eigenvalues.

**d.** If the interval  $(a, b)$  contains several turning points, i.e. if we study the problem

$$(9.3) \quad -\varepsilon u'' + \tilde{p}u' + xqu = ru, \quad u(a) = A, \quad u(b) = B,$$

where  $\tilde{p}$  has several distinct zeros in  $[a, b]$ , we can do exactly the same as before. Each turning point gives rise to a denumerable set of eigenvalues, which satisfy Theorem 1 (or the analogous result for problem (9.1)) and the spectrum is the union of these sets. In order to generalize the proof of Theorem 1 to this case we have only to perform a transformation analogous to (2.1) and to construct a complete set of approximate eigenfunctions for each turning point. The construction (and proof) of asymptotic approximations of the solutions is analogous to the cases sketched above.

**e.** If the interval contains a turning point of higher order or if two (or more) simple turning points coalesce in the limit for  $\varepsilon \rightarrow +0$ , i.e. if  $p(x, 0)$  has a multiple zero, then the spacing between the eigenvalues tends to zero for  $\varepsilon \rightarrow +0$  and the set of eigenvalues tends to a dense subset of the positive real axis. In order to prove such a result we impose on the coefficients  $\tilde{p}$  of (9.3) the more general condition

$$\tilde{p}(x, 0) = x|x|^{p-1}(1 + \mathcal{O}(\varepsilon)) \quad \text{or} \quad \tilde{p}(x, 0) = |x|^p(1 + \mathcal{O}(\varepsilon)).$$

To this problem we apply the analogue of the symmetrizing transformation (2.1), which results in the equation

$$(9.4) \quad -\varepsilon v'' + \tilde{p}^2 v / 4\varepsilon - \frac{1}{2} p' v + x q v = \lambda v, \quad v(a) = v(b) = 0.$$

If  $0 \leq \nu < 1$ , its Rayleigh quotient is bounded from below by an arbitrarily large constant if  $\varepsilon$  is small enough, such that all eigenvalues vanish at infinity in the limit for  $\varepsilon \rightarrow +0$ . If  $\nu > 1$ , we substitute  $x = \varepsilon^{1/(1+\nu)} \xi$  and we multiply the equation (9.4) by  $\varepsilon^{(\nu-1)/(\nu+1)}$ . Comparing the Rayleigh quotient of the resulting equation to the Rayleigh quotient of Hermite's operator (cf. § 5) we can show that all eigenvalues of (9.4) tend to zero with the order  $\mathcal{O}(\varepsilon^{(\nu-1)/(\nu+1)})$  and that their spacing diminishes with the same factor. For more details see [7].

f. By analogous methods we can attack the elliptic singularly perturbed boundary value problem on bounded domain  $G \in \mathbb{R}^n$ .

$$\varepsilon L u + \sum_{i=1}^n p_i \partial u / \partial x_i + q u = 0, \quad u|_{\partial G} \text{ prescribed,}$$

where  $L$  is a uniformly elliptic operator and where the vector  $\mathbf{p}$  has an isolated zero with a nonzero Jacobian; cf. [6, chaps. 4, 5, 6] and [14].

#### REFERENCES

- [1] L. R. ABRAHAMSSON, *A priori estimates for solutions of singular perturbations with turning points*, report 56, Department of Computer Sciences, Uppsala University, Sweden, 1975.
- [2] R. C. ACKERBERG AND R. E. O'MALLEY, *Boundary layer problems exhibiting resonance*, *Studies Appl. Math.*, 49 (1970), pp. 277–295.
- [3] L. PAMELA COOK AND W. ECKHAUS, *Resonance in a boundary value problem of singular perturbation type*, *Ibid.*, 52 (1973), pp. 129–139.
- [4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Interscience, New York, 1953.
- [5] P. P. N. DE GROEN, *Spectral properties of second order singularly perturbed boundary value problems with turning points*, *J. Math. Anal. Appl.*, 57 (1977), pp. 119–149.
- [6] ———, *Singularly Perturbed Differential Operators of Second Order*, Mathematical Centre Tract 68, ISBN 90 6196 120 3, Mathematisch Centrum, Amsterdam, 1976.
- [7] ———, *A singular perturbation problem of turning point type*, *New Developments in Differential Equations*, Proceedings of the second Scheveningen conference on differential equations, W. Eckhaus, ed., North-Holland Mathematics Studies 21, North-Holland, Amsterdam, 1976, pp. 117–124.
- [8] B. J. MATKOVSKY, *On boundary layer problems exhibiting resonance*, *SIAM Rev.*, 17 (1975), pp. 82–100.
- [9] F. W. J. OLVER, *Uniform asymptotic expansions and singular perturbations*, *Asymptotic Methods and Singular Perturbations*, SIAM-AMS proceedings vol. 10 (1976), pp. 105–117.
- [10] ———, *Sufficient conditions for Ackerberg–O'Malley resonance*, *SIAM J. Math. Anal.*, 9 (1978), pp. 328–355.
- [11] R. E. O'MALLEY, *Topics in singular perturbations*, *Advances in Math.*, 2 (1968), pp. 365–470.
- [12] W. ECKHAUS, *Matched Asymptotic Expansions and Singular Perturbations*, North-Holland Mathematics Studies 6, North-Holland, Amsterdam, 1973.
- [13] P. P. N. DE GROEN, *Elliptic singular perturbations of first-order operators with critical points*, *Proc. Roy. Soc. Edinburgh Sect. A*, 74A (1974/75), pp. 91–113.
- [14] JU. I. KIFER, *On the spectral stability of invariant tori under small random perturbations of dynamical systems*, *Dokl. Akad. Nauk SSSR*, 235 (1977), pp. 512–515 = *Soviet Math. Dokl.*, 18 (1977), pp. 952–956.

## CHARACTERIZATION OF CONTINUOUS SELECTIONS FOR THE METRIC PROJECTION FOR GENERALIZED SPLINES\*

MANFRED SOMMER†

**Abstract.** In this paper we give a characterization of those generalized spline spaces which admit continuous selections for the metric projection. We denote by generalized splines those weak Chebyshev spaces which can be decomposed in Chebyshev spaces by finitely many knots. This characterization is a partial solution of a problem raised by Lazar–Morris–Wulbert and generalizes a result of Nürnberger–Sommer established for polynomial splines. For constructing a continuous selection we show some properties of generalized splines. We prove an interpolation theorem and give a characterization of the existence of best approximations. These results generalize in a certain sense results of Karlin, Rice and Schumaker established for polynomial splines.

**Introduction.** If  $G$  is a nonempty subset of a normed linear space  $E$ , then for each  $f$  in  $E$  we define  $P_G(f) := \{g_0 \in G \mid \|f - g_0\| = \inf \{\|f - g\| \mid g \in G\}\}$  which is called the *set of best approximations* of  $f$  from  $G$ .  $P_G$  defines a set-valued mapping of  $E$  into  $2^G$  which in the literature is called the *metric projection* onto  $G$ . A continuous mapping  $s$  of  $E$  into  $G$  is called a *continuous selection for the metric projection*  $P_G$  (or, more briefly, continuous selection) if  $s(f)$  is in  $P_G(f)$  for each  $f$  in  $E$ .

In this paper we treat the problem of the existence of continuous selections for  $n$  dimensional subspaces  $G$  of  $C[a, b]$ , with  $C[a, b]$  as usual the Banach space of real-valued continuous functions on  $[a, b]$  under the uniform norm.

Lazar, Morris and Wulbert [7] have been the first to characterize those one dimensional subspaces  $G$  of  $C(X)$ ,  $X$  compact, which admit a continuous selection. They have raised the problem of characterizing the corresponding  $n$ -dimensional subspaces.

With new methods and in the setting of weak Chebyshev subspaces Nürnberger and Sommer [9] have established the existence of continuous selections for a subclass of those weak Chebyshev subspaces of  $C[a, b]$  whose nonzero elements have no zero intervals. From this, there follows a result of Brown [2] for five dimensional subspaces of  $C[-1, 1]$ . Combining the result in [9] with recent results of Sommer [14] and Sommer and Strauss [16] we get a characterization of the spaces which have continuous selections from among the  $n$ -dimensional weak Chebyshev subspaces  $G$  of  $C[a, b]$  whose nonzero elements have no zero intervals:

There exists a continuous selection for  $G$  if and only if each  $g$  in  $G$ ,  $g \neq 0$ , has at most  $n - 1$  distinct zeros on  $[a, b] \setminus \{x_0\}$  where  $x_0$  only depends on  $G$ .

Recently, Nürnberger [8] has shown that the weak Chebyshev property is necessary for the existence of continuous selections for subspaces  $G$  of  $C[a, b]$ . Thus the above formulated problem of Lazar–Morris–Wulbert is solved for all  $n$ -dimensional subspaces  $G$  of  $C[a, b]$  except for the following case:  $G$  is weak Chebyshev and there exists at least one nonzero  $g$  in  $G$  vanishing on intervals. We denote this subclass of the class of the  $n$ -dimensional weak Chebyshev spaces by  $\mathcal{X}_n$ .

For special elements of  $\mathcal{X}_n$  the problem of Lazar–Morris–Wulbert has been treated by Nürnberger and Sommer [10] and Sommer [13]. Nürnberger and Sommer [10] have given a characterization of those spline spaces which admit continuous selections and Sommer [13] has given a characterization of those 1-Chebyshev spaces which also

\* Received by the editors May 25, 1978, and in revised form December 4, 1978.

† Institut für Angewandte Mathematik der Universität Erlangen-Nürnberg, Erlangen, West Germany.

admit continuous selections. Spline spaces and also special 1-Chebyshev spaces are elements of  $\mathcal{L}_n$ .

In this paper we examine the problem of Lazar–Morris–Wulbert for the elements of  $\mathcal{L}_n$ . We define a great subclass  $\mathcal{V}_n$  of  $\mathcal{L}_n$  consisting of all  $n$ -dimensional weak Chebyshev spaces  $G$  which can be decomposed in Chebyshev spaces by a finite set of knots (see Sommer [15]). Therefore, we may denote these spaces by generalized splines. Since the spline spaces with dimension  $n$  are elements of  $\mathcal{V}_n$  and the splines are the prototypes of the weak Chebyshev spaces [6], the class  $\mathcal{V}_n$  seems to be the most important subclass of the class of the weak Chebyshev spaces. But there are also many elements of  $\mathcal{L}_n$  which are not elements of  $\mathcal{V}_n$ . We give a complete characterization of those spaces  $G$  in  $\mathcal{V}_n$  which admit continuous selections. We show that a continuous selection for  $G$  exists if and only if the following conditions are satisfied:

- (\*) No nonzero  $g$  in  $G$  has more than one zero interval in  $[a, b]$  and the number of the boundary zeros of  $g$  is bounded in a certain sense.

In order to prove the characterization we first prove an interpolation property and a characterization theorem for best approximations for any  $G$  in  $\mathcal{V}_n$  satisfying condition (\*). From these theorems there follow for a special class of spline spaces results of Karlin [5, p. 503], Rice [11, p. 152] and Schumaker [12] established for all spline spaces. By using our results we are able to construct continuous selections provided that condition (\*) is satisfied.

The construction of the selection is highly local and based on local alternation elements whose local uniqueness is guaranteed by condition (\*).

If for any  $G$  in  $\mathcal{V}_n$  condition (\*) is not satisfied, then we are able to show the nonexistence of a continuous selection applying a fundamental lemma of Lazar–Morris–Wulbert.

Our construction of a continuous selection is based on the construction of a continuous selection for splines established by Nürnberger and Sommer [10]. While in that paper the authors have been able to use well known results from spline theory, we have at first in this paper to establish some results about the elements of  $\mathcal{V}_n$ .

Finally we show that from our characterization it follows the characterization of continuous selections for splines established in [10]: There exists a continuous selection for splines of degree  $m$  with  $k$  fixed knots ( $m + k + 1 = n$ ) if and only if  $k \leq m + 1$ . We also apply our characterization theorem to other special elements of  $\mathcal{V}_n$ , namely to the continuously composed Chebyshev spaces and get a characterization of the existence of continuous selections for those spaces.

We also show by examples that not all  $G$  in  $\mathcal{V}_n$  have the same behavior as the spline functions. Therefore, the class  $\mathcal{V}_n$  does not only consist of those weak Chebyshev spaces which have the same properties as the splines. This statement is also verified by a result of Sommer [15] having shown that there are elements of  $\mathcal{V}_n$  which are not uniqueness spaces in the  $L_1$ -norm, while  $L_1$ -uniqueness for spline spaces is always satisfied.

**1. Preliminaries.** In the following let  $G$  be an  $n$ -dimensional subspace of  $C[a, b]$ .

**DEFINITION 1.1.**  $G$  is called *Chebyshev* if each  $g$  in  $G$  has at most  $n - 1$  zeros on  $[a, b]$ .

$G$  is called *weak Chebyshev* if each  $g$  in  $G$  has at most  $n - 1$  changes of sign, i.e. there do not exist points  $a \leq x_0 < x_1 < \dots < x_n \leq b$  such that  $g(x_i) \cdot g(x_{i+1}) < 0$  for  $i = 0, \dots, n - 1$ .

We denote the class of all  $n$ -dimensional weak Chebyshev subspaces of  $C[a, b]$  by  $\mathcal{W}_n$ .

Jones and Karlovitz [4] have characterized the elements of  $\mathcal{W}_n$ . For this characterization we need the following definition:



DEFINITION 1.2. If  $f$  is in  $C[a, b]$ , then  $g$  in  $P_G(f)$  is called an *alternation element* (AE) of  $f$  if there exist  $n + 1$  distinct points  $a \leq x_0 < x_1 < \dots < x_n \leq b$  such that  $\varepsilon(-1)^i(f - g)(x_i) = \|f - g\|$ ,  $i = 0, \dots, n$ ,  $\varepsilon = \pm 1$ . The points  $x_0, \dots, x_n$  are called *alternating extreme points* of  $f - g$ .

Jones and Karlovitz [4] have proved the following theorem:

THEOREM 1.3. *The following statements are equivalent:*

- (i)  $G$  is in  $\mathcal{W}_n$ .
- (ii) For each  $f$  in  $C[a, b]$  there exists at least one AE in  $P_G(f)$ .
- (iii) Given  $a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$  there exists a  $g$  in  $G$ ,  $g \neq 0$ , such that  $(-1)^{i+1}g(x) \geq 0$ ,  $x_{i-1} < x < x_i$ ,  $i = 1, \dots, n$ .
- (iv) If  $g_1, g_2, \dots, g_n$  is a basis of  $G$ , then  $a \leq t_1 < t_2 < \dots < t_n \leq b$ ,  $a \leq s_1 < s_2 < \dots < s_n \leq b$  imply

$$\det |g_i(t_j)| \det |g_i(s_j)| \geq 0.$$

Applying this theorem and Definition 1.1 it is easy to show:

LEMMA 1.4. (i) *If  $G$  is weak Chebyshev, then there exists a  $g$  in  $G$  with exactly  $n - 1$  changes of sign on  $(a, b)$ ;*

(ii) *If  $G$  is not weak Chebyshev, then there exists a  $g$  in  $G$  with at least  $n$  changes of sign on  $(a, b)$ .*

Furthermore we need the following standard definition:

DEFINITION 1.5. A zero  $x_0$  of  $f$  in  $C[a, b]$  is said to be an *isolated zero* if there is a neighborhood of  $x_0$  such that  $f(x) \neq 0$  on  $U \setminus \{x_0\}$ .

A zero  $x_0$  of  $f$  in  $C[a, b]$  is said to be a *double zero* if  $x_0$  is an isolated zero on  $(a, b)$  and  $f$  does not change sign at  $x_0$ .

A zero  $x_0$  of  $f$  in  $C[a, b]$  is said to be a *simple zero* if  $x_0$  is not a double zero of  $f$  or if  $x_0 = a$  or  $x_0 = b$ .

Two zeros  $x_1, x_2$  of  $f$  in  $C[a, b]$  are said to be *separated* if there is a  $x_0$ ,  $x_1 < x_0 < x_2$ , with  $f(x_0) \neq 0$ .

Let  $Z(f)$  be the set of all distinct zeros of  $f$  and  $Z_d(f)$  the set of all double zeros of  $f$ . Furthermore, let  $\text{bd } Z(f)$  be the set of the boundary points of  $Z(f)$ .

We denote by  $|Z(f)|$  and  $|Z^*(f)|$  the number of the distinct zeros of  $f$  or the number of the zeros of  $f$  counting simple zeros as one zero and double zeros as two zeros, respectively.

In [15] we have shown that weak Chebyshev spaces, under appropriate hypotheses, can be decomposed in Chebyshev spaces by a finite set of knots. For this we need the following definition:

DEFINITION 1.6. A zero  $x_0$  of  $f$  in  $C[a, b]$  is said to be a *nonvanishing zero* with respect to  $G$  if there is a  $g$  in  $G$  with  $g(x_0) \neq 0$ .

In the following the term "with respect to  $G$ " will be omitted.

We have proved in [15]:

THEOREM 1.7. *Let  $G$  be in  $\mathcal{W}_n$  and each  $x$  in  $[a, b]$  be a nonvanishing zero. Let at least one nonzero  $g$  be in  $G$  having zero intervals. Assume also that there exists a  $\delta > 0$  such that if  $g$  in  $G$  and  $g \equiv 0$  on  $[c, d] \subset [a, b]$  where  $c, d \in \{x \in [a, b] \mid g(x) \neq 0\} \cup \{a, b\}$ , then  $d - c \geq \delta$ . Then there exists a minimal set of knots  $a = x_0 < x_1 < \dots < x_s = b$  such that the spaces  $G^i = G|_{[x_{i-1}, x_i]}$  are Chebyshev with dimension  $n_i$  for  $i = 1, \dots, s$ .*

We define:

$$\mathcal{V}_n = \{G \in \mathcal{W}_n \mid G \text{ fulfills the hypotheses of Theorem 1.7}\}.$$

By Theorem 1.7  $\mathcal{V}_n$  contains exactly those spaces  $G$  in  $\mathcal{W}_n$  which we can decompose by finitely many knots in Chebyshev spaces. In § 3 we will give a characterization of those

elements of  $\mathcal{V}_n$  which admit continuous selections. In order to establish this characterization we have to define the following: Let  $G$  be in  $\mathcal{V}_n$  and  $a = x_0 < x_1 < \dots < x_s = b$  be knots for  $G$  according to Theorem 1.7. Then we define for any  $i, j \in \{0, 1, \dots, s\}$ ,  $i < j$ :

$$\bar{G}_{ij} = \{g \in G \mid g \equiv 0 \text{ on } [x_i, x_j]\}, \quad \dim \bar{G}_{ij} = m_{ij}.$$

In general the spaces  $\bar{G}_{ij}$  are not weak Chebyshev. But we now define two subclasses of  $\mathcal{V}_n$  for which all  $\bar{G}_{ij}$  are weak Chebyshev:

$$\begin{aligned} \tilde{\mathcal{V}}_n &= \{G \in \mathcal{V}_n \mid |\text{bd } Z(g)| \leq m_{ij} \text{ for each } g \in \bar{G}_{ij} \text{ and each } \bar{G}_{ij}\} \\ \check{\mathcal{V}}_n &= \{G \in \tilde{\mathcal{V}}_n \mid \text{No nonzero } g \in G \text{ has two separated zero intervals}\} \end{aligned}$$

Furthermore we define for any  $G$  in  $\mathcal{V}_n$  and any  $k, l \in \{0, 1, \dots, s\}$ ,  $k < l$ :

$$G^{kl} = G|_{[x_k, x_l]}, \quad \dim G^{kl} = n_{kl}$$

and for any subinterval  $[x_i, x_j] \subset [x_k, x_l]$ :

$$\bar{G}_{ij}^{kl} = \bar{G}_{ij}|_{[x_k, x_l]}, \quad \dim \bar{G}_{ij}^{kl} = m_{ij}^{kl}.$$

In [15] we have proved that  $G^{kl}$  is weak Chebyshev for any  $k, l \in \{0, 1, \dots, s\}$ .

**2. Properties of  $\tilde{\mathcal{V}}_n$  and  $\check{\mathcal{V}}_n$ .** In order to construct a continuous selection for  $G$  provided that  $G$  is in  $\check{\mathcal{V}}_n$  we need some properties of  $\tilde{\mathcal{V}}_n$ .

In [15] we have shown that any  $G$  in  $\tilde{\mathcal{V}}_n$  is a uniqueness space for approximation in the  $L_1$ -norm. For this we have proved a fundamental lemma which is also very important for our characterization of the existence of continuous selections:

LEMMA 2.1. *Let  $G$  be in  $\tilde{\mathcal{V}}_n$ . Then for any  $i, j \in \{0, 1, \dots, s\}$ ,  $i < j$ , the following are true:*

- (i)  $\bar{G}_{ij}$  is weak Chebyshev with dimension  $m_{ij}$ .
- (ii) For any function  $g_1$  in  $\bar{G}_{ij}|_{[x_j, b]}$  there is a  $\tilde{g}_1$  in  $G$  such that  $\tilde{g}_1 = g_1$  on  $[x_j, b]$  and  $\tilde{g}_1 \equiv 0$  on  $[a, x_j]$ .
- (iii) For any function  $g_2$  in  $\bar{G}_{ij}|_{[a, x_i]}$  there is a  $\tilde{g}_2$  in  $G$  such that  $\tilde{g}_2 = g_2$  on  $[a, x_i]$  and  $\tilde{g}_2 \equiv 0$  on  $[x_i, b]$ .

Next we show an interpolation property for any  $G$  in  $\check{\mathcal{V}}_n$  which we need for a characterization theorem for best approximations from  $G$ . For the proof of this interpolation theorem we first need a lemma on the number of separated zeros of functions in weak Chebyshev spaces.

LEMMA 2.2. (Stockenberg [17]). *Let  $G$  be in  $\mathcal{W}_n$ . Then the following assertions hold:*

- (i) *If there is a  $g$  in  $G$  with  $n$  separated, nonvanishing zeros  $x_1 < x_2 < \dots < x_n$ , then  $g(x) = 0$  for all  $x$  with  $x \leq x_1$  or  $x \geq x_n$ .*
- (ii) *No  $g$  in  $G$  has more than  $n$  separated, nonvanishing zeros.*

LEMMA 2.3. *Let  $G$  be in  $\tilde{\mathcal{V}}_n$ . Let  $n$  points  $a \leq y_1 < y_2 < \dots < y_n \leq b$  be given satisfying*

$$y_{n-n_{is}} < x_i < y_{n_{oi}+1}, \quad i = 1, \dots, s-1,$$

(for  $n - n_{is} = 0$  and  $n_{oi} + 1 = n + 1$  the first or the second inequality is omitted, respectively). *Then for any  $n$  real numbers  $\{z_i\}_{i=1}^n$  there exists exactly one  $g_0$  in  $G$  with  $g_0(y_i) = z_i$  for  $i = 1, \dots, n$ .*

*Proof.* We first remember that  $n_{ij} = \dim G^{ij} = \dim G|_{[x_i, x_j]}$ . Let any function  $g_0 \in G$  with  $g_0(y_i) = 0$  for  $i = 1, \dots, n$ . Then the lemma is proved if we can show that  $g_0 \equiv 0$ .

We now assume that  $g_0 \not\equiv 0$  and distinguish two cases: *First:* Let  $g_0$  have no zero intervals. Then from Lemma 2.2 it follows that  $g_0$  has exactly  $n$  distinct zeros on  $[a, b]$  such that  $y_1 = a$  and  $y_n = b$ . Since there is at least one nonzero element  $g \in G$  having

zero intervals, by Lemma 2.1 there exists a nonzero function  $\tilde{g} \in G$  with exactly one zero interval  $[a, x_i]$  or  $[x_j, b]$ , respectively. Therefore  $\dim \bar{G}_{0i} = m_{0i} \geq 1$  or  $\dim \bar{G}_{js} = m_{js} \geq 1$ , respectively.

Without loss of generality let  $m_{0i} \geq 1$ . Since  $\bar{G}_{0i}$  is weak Chebyshev by Lemma 2.1 and  $G \in \mathcal{V}_n$ , by Lemma 1.4 there is a  $\bar{g} \in \bar{G}_{0i}$  such that  $\bar{g}$  has exactly one maximal zero interval  $[a, x_i]$  with  $x_j \geq x_i$  and  $\bar{g}$  has exactly  $m_{0i} - 1$  changes of sign on  $(x_j, b)$ . Because of  $|\text{bd } Z(\bar{g})| \leq m_{0i}$  the function  $\bar{g}$  has exactly  $m_{0i}$  zeros on  $[x_j, b]$  where  $m_{0i} - 1$  zeros are zeros with changes of sign. In particular  $\bar{g}(b) \neq 0$ , since  $\bar{g}(x_j) = 0$ .

Let  $r_1$  be the number of the common zeros of  $g_0$  and  $\bar{g}$  on  $[a, b]$ . We classify the other  $n - r_1 - 1$  distinct zeros of  $g_0$  on  $(x_j, b)$  as follows:

Let  $r_2$  be the number of the double zeros having the property that for each of these zeros there exists a neighborhood  $U$  such that  $g_0 \cdot \bar{g} \geq 0$  on  $U$ .

Let  $r_3$  be the number of the double zeros having the property that for each of these zeros there exists a neighborhood  $U$  such that  $g_0 \cdot \bar{g} \leq 0$  on  $U$ .

Let  $r_4$  be the number of changes of sign.

Then  $n = r_1 + r_2 + r_3 + r_4 + 1$  and for sufficiently small  $c > 0$  either the function  $g_0 - c\bar{g}$  or the function  $g_0 + c\bar{g}$  has at least  $n$  nonvanishing separated zeros on  $[a, b]$ . But by Lemma 2.2 this is not possible.

*Second:* Let  $[x_i, x_j]$  be the maximal zero interval of  $g_0$  ( $i < j$ ). Therefore, the function  $g_0$  has no zero interval in  $[a, x_i] \cup [x_j, b]$ . Without loss of generality we may assume that  $x_j < b$ . By hypothesis, the function  $g_0$  has at least  $n - n_{0j}$  separated zeros on  $(x_j, b]$  and because of  $g_0(x_j) = 0$  even  $n - n_{0j} + 1$  separated zeros on  $[x_j, b]$ . Since  $g_0 \in \bar{G}_{ij}$ , we get  $\dim \bar{G}_{ij} \geq 1$ . Since  $g_0 \neq 0$  on  $[x_j, b]$ , it follows from Lemma 2.1 that  $m_{0j} = \dim \bar{G}_{0j} \geq 1$ . From the definition of  $G^{0j}$  and  $\bar{G}_{0j}$  it follows immediately that  $n = n_{0j} + m_{0j}$ .

Since  $g_0 \in \bar{G}_{ij}$ , by Lemma 2.1 there exists a  $\tilde{g} \in \bar{G}_{0j}$  such that  $\tilde{g} = g_0$  on  $[x_j, b]$ . Therefore  $|\text{bd } Z(\tilde{g})| \geq n - n_{0j} + 1 = m_{0j} + 1$ . But this is a contradiction of the hypothesis that  $|\text{bd } Z(g)| \leq m_{0j}$  for all  $g \in \bar{G}_{0j}$ .

Furthermore we need a lemma dealing with the spaces  $\bar{G}_{ij}^{kl}$ . We show that for any  $k, l \in \{0, 1, \dots, s\}$ ,  $k < l$ , these spaces also satisfy the conditions made for the elements of  $\mathcal{V}_n$ .

LEMMA 2.4. *Let  $G$  be in  $\mathcal{V}_n^{\tilde{x}}$ . Then for any  $k, l \in \{0, 1, \dots, s\}$ ,  $k < l$ , and any subinterval  $[x_i, x_j]$  of  $[x_k, x_l]$  the following is true:*

$$|\text{bd } Z(g)| \leq m_{ij}^{kl} \quad \text{for each } g \text{ in } \bar{G}_{ij}^{kl}.$$

*Proof.* At first we treat the case that  $[x_k, x_l]$  is a boundary interval of  $[a, b]$ .

*First:*  $a = x_0 < x_1 < b$  (the case  $a < x_k < x_s = b$  follows analogously). We assume that there is a subinterval  $[x_i, x_j] \subset [a, x_l]$  and a function  $g_0 \in \bar{G}_{ij}$  such that  $|\text{bd } Z(g_0)| \geq m_{ij}^{0l} + 1$  on  $[a, x_l]$ . Since  $|\text{bd } Z(g_0)| \leq m_{ij}$  on  $[a, b]$ , we get  $m_{ij} > m_{ij}^{0l}$ . Therefore, there exist exactly  $m_{ij} - m_{ij}^{0l}$  linearly independent functions  $g_1, g_2, \dots, g_{m_{ij} - m_{ij}^{0l}}$  in  $\bar{G}_{ij}$  vanishing identically on  $[a, x_l]$ . Then we get  $\bar{G}_{0l} = \langle g_1, g_2, \dots, g_{m_{ij} - m_{ij}^{0l}} \rangle$ , since  $\bar{G}_{0l} \subset \bar{G}_{ij}$ . Hence  $m_{0l} = m_{ij} - m_{ij}^{0l}$ .

We now show that  $g_0, g_1, \dots, g_{m_{0l}}$  are linearly independent of  $[x_i, b]$ . If there is a  $\tilde{g} \in \bar{G}_{0l}$  such that  $\tilde{g} = g_0$  on  $[x_i, b]$ , then  $g_0 \equiv 0$  on  $[x_i, x_l]$ , since otherwise  $g_0 - \tilde{g}$  has two separated zero intervals  $[x_i, x_j]$  and  $[x_i, b]$ . This would be a contradiction of the hypothesis that  $G \in \mathcal{V}_n$ . Then  $|\text{bd } Z(g_0)| \geq m_{ij}^{0l} + 1$  on  $[a, x_l]$ . By Lemma 1.4 there exists a  $\bar{g} \in \bar{G}_{0l}$  with exactly  $m_{0l} - 1$  changes of sign on  $(x_i, b)$ . Hence  $|\text{bd } Z(\bar{g})| = m_{0l}$  on  $[x_i, b]$ .

Then for sufficiently small  $c > 0$  the function  $\bar{g} - cg_0$  has exactly  $m_{ij}^{0l} + 1$  separated zeros on  $[a, x_l]$  and at least  $m_{0l}$  separated zeros on  $[x_i, b]$ . But this is a contradiction,

because  $\bar{g} - cg_0 \in \bar{G}_{ij}$ . Thus we have proved that  $g_0, g_1, \dots, g_{m_{0l}}$  are linearly independent on  $[x_i, b]$ . Then by Lemma 1.4 there exists a function

$$\hat{g} = a_0 g_0 + \sum_{i=1}^{m_{0l}} a_i g_i$$

with at least  $m_{0l}$  changes of sign on  $(x_i, b)$ . Since by Lemma 2.1  $\bar{G}_{0l}$  is weak Chebyshev with dimension  $m_{0l}$ , we get  $a_0 \neq 0$ . Then  $|\text{bd } Z(\hat{g})| \geq m_{ij}^{0l} + 1$  on  $[a, x_i]$  and  $|\text{bd } Z(\hat{g})| \geq m_{0l}$  on  $(x_i, b)$ . Therefore  $|\text{bd } Z(\hat{g})| \geq m_{ij} + 1$ . Because of  $\hat{g} \in \bar{G}_{ij}$  we get a contradiction again.

*Second:*  $a < x_k < x_l < b$ . At first we consider again the boundary interval  $[a, x_l] \subset [a, b]$ . By the first case it follows:

$$|\text{bd } Z(g)| \leq m_{ij}^{0l} \quad \text{for each } g \in \bar{G}_{ij}^{0l}$$

where  $[x_i, x_j]$  is an arbitrary subinterval of  $[a, x_l]$ . From a remark in § 1 it follows that  $G^{0l}$  is weak Chebyshev with dimension  $n_{0l}$ . Therefore, the space  $G^{0l} \subset C[a, x_l]$  satisfies the same hypotheses as the space  $G \in \mathcal{V}_n$  if we replace  $[a, b]$  by  $[a, x_l]$  and the dimension  $n$  by  $n_{0l}$ . Since  $[x_k, x_l]$  is a boundary interval of  $[a, x_l]$ , we may conclude as in the first case and get the desired statement.

Now we are able to show that all  $g$  in  $P_G(f)$  coincide on a knot interval for any  $f$  in  $C[a, b]$ .

**LEMMA 2.5.** *Let  $G$  be in  $\mathcal{V}_n$ . Then there exists an interval  $[x_i, x_j]$  such that  $g = \tilde{g}$  on  $[x_i, x_j]$  for all  $g, \tilde{g}$  in  $P_G(f)$ . Furthermore, for each  $g$  in  $P_G(f)$  the error  $f - g$  has at least  $n_{ij} + 1$  alternating extreme points on  $[x_i, x_j]$ .*

*Proof.* Let  $f \in C[a, b]$ . By Theorem 1.3 there exists at least one AE  $g_0 \in P_G(f)$ . Without loss of generality let  $g_0 \equiv 0$ .

If there is an interval  $[x_i, x_{i+1}]$  such that  $f - 0$  has at least  $n_{i+1} + 1$  alternating extreme points on  $[x_i, x_{i+1}]$ , then by the well-known characterization theorem of Chebyshev spaces all  $g \in P_G(f)$  coincide on  $[x_i, x_{i+1}]$ , since  $G^{i+1}$  is Chebyshev.

But if there does not exist such an interval, there will exist an interval  $[x_i, x_j]$  such that  $f - 0$  has  $n_{ij} + 1$  alternating extreme points  $x_i \leq t_1 < t_2 < \dots < t_{n_{ij}+1} \leq x_j$ , but  $f - 0$  has no  $n_{kl} + 1$  alternating extreme points on any subinterval  $[x_k, x_l] \subset [x_i, x_j]$ . Then

$$t_{n_{ij}+1-n_p} < x_p < t_{n_p+1}, \quad p = i+1, \dots, j-1$$

(for  $n_{ip} = n_{ij}$  the second inequality is omitted).

Now let an arbitrary  $g \in P_G(f)$  be given. Then because of  $\|f\| = \varepsilon(-1)^p f(t_p) \geq \varepsilon(-1)^p (f-g)(t_p)$ ,  $p = 1, \dots, n_{ij} + 1$ ,  $\varepsilon = \pm 1$ , we get  $\varepsilon(-1)^p g(t_p) \geq 0$  for  $p = 1, \dots, n_{ij} + 1$ . Therefore, the function  $g$  has at least one zero on each interval  $[t_p, t_{p+1}]$ , but in general not  $n_{ij}$  distinct zeros on  $[t_1, t_{n_{ij}+1}]$ , since it is possible that the zeros on  $[t_{p-1}, t_p]$  and  $[t_p, t_{p+1}]$  coincide at  $t_p$ . But we show that in this case the function  $g$  has an isolated double zero at  $t_p$ .

We now choose exactly one zero of  $g$  on each interval  $[t_p, t_{p+1}]$  for  $p = 1, \dots, n_{ij}$  as follows:

If  $g$  has a zero on  $[t_1, t_2]$ , then we define  $z_1$  to be an arbitrary zero on  $[t_1, t_2]$ . If  $g$  has no zero on  $[t_1, t_2]$ , then  $g(t_2) = 0$  and we define  $z_1 = t_2$ . Let now  $p-1$  zeros  $z_r \in [t_r, t_{r+1})$ ,  $r = 1, \dots, p-1$ , of  $g$  be defined. We define a zero  $z_p \in [t_p, t_{p+1})$  as follows:

If  $g$  has a zero on  $(t_p, t_{p+1})$ , then we define  $z_p$  to be an arbitrary zero on  $(t_p, t_{p+1})$ . If  $g$  has no zero on  $(t_p, t_{p+1})$ , then  $g(t_p) = 0$  or  $g(t_{p+1}) = 0$ . We distinguish:

- (i) If  $g(t_p) = g(t_{p+1}) = 0$  and  $z_{p-1} < t_p$ , then we define  $z_p = t_p$ . If  $g(t_p) = g(t_{p+1}) = 0$  and  $z_{p-1} = t_p$  and  $g$  has a double zero at  $t_p$ , then we define  $z_p = t_p$ . If

$g(t_p) = g(t_{p+1}) = 0$  and  $z_{p-1} = t_p$  and  $g$  has a change of sign at  $t_p$ , then we define

$$z_p = t_{p+1}.$$

(ii) If  $g(t_{p+1}) \neq 0$ , then  $g(t_p) = 0$  and we define  $z_p = t_p$ .

(iii) If  $g(t_p) \neq 0$ , then  $g(t_{p+1}) = 0$  and we define  $z_p = t_{p+1}$ .

Thus we have defined  $n_{ij}$  zeros where at most two of these zeros coincide. We next show that from  $z_{p-1} = z_p$  for some  $p \in \{2, \dots, n_{ij}\}$  it follows that  $g$  has a double isolated zero at  $z_{p-1} = z_p$ : Let  $z_{p-1} = z_p$  for some  $p \in \{2, \dots, n_{ij}\}$ . Because  $z_{p-1} \in [t_{p-1}, t_p]$  and  $z_p \in [t_p, t_{p+1}]$  we get  $z_{p-1} = z_p = t_p$  and  $g(t_p) = 0$ . Moreover,  $g$  has no zero on  $(t_{p-1}, t_p) \cup (t_p, t_{p+1})$ . If also  $g(t_{p+1}) = 0$ , then by selection of  $z_p$   $g$  has a double zero at  $z_p$  and we are ready.

Therefore, there remains only the case that  $g(t_{p+1}) \neq 0$ . Then  $\varepsilon(-1)^{p+1}g(t_{p+1}) > 0$ . We distinguish:

(i) If  $g(t_{p-1}) \neq 0$ , then  $\varepsilon(-1)^{p-1}g(t_{p-1}) > 0$  and thus we get that  $t_p$  is a double zero of  $g$ .

(ii) If  $g(t_{p-1}) = 0$ , then by selection of  $z_{p-1}$   $g$  has a change of sign at  $t_{p-1}$  and, moreover,  $z_{p-2} = t_{p-1}$ . If also  $g(t_{p-2}) = g(t_{p-3}) = \dots = g(t_1) = 0$ , then we would get by definition:

$$z_1 = t_1, \quad z_2 = t_2, \dots, z_{p-1} = t_{p-1} < t_p.$$

But because of  $z_{p-1} = z_p = t_p$  this case is not possible. Therefore, there is a  $t_j$  with  $g(t_j) \neq 0$ . Let  $t_{p-s}$  be the greatest point less than  $t_{p-1}$  such that  $g(t_{p-s}) \neq 0$ . Then  $g(t_{p-s+1}) = \dots = g(t_{p-1}) = g(t_p) = 0$ . Then  $g$  has zeros with changes of sign at  $t_{p-s+1}, \dots, t_{p-1}$  and because  $z_{p-2} = t_{p-1}, z_{p-3} = t_{p-2}, \dots, z_{p-s} = t_{p-s+1}$  no further zero on  $[t_{p-s}, t_p]$  by definition. Because  $\varepsilon(-1)^{p-s}g(t_{p-s}) > 0$  we get  $\varepsilon(-1)^{p-1}g(x) > 0$  for all  $x \in (t_{p-1}, t_p)$ . Then because  $\varepsilon(-1)^{p+1}g(x) > 0$  for all  $x \in (t_p, t_{p+1})$  the function  $g$  has a double zero at  $t_p$ .

Thus we have shown that if  $z_{p-1} = z_p = t_p$  for some  $p \in \{2, \dots, n_{ij}\}$ ,  $g$  has a double zero at  $t_p$ . Because  $t_{n_{ij}+1-n_{pi}} < x_p < t_{n_{ip}+1}$  for  $p = i+1, \dots, j-1$ , and because  $z_p \in [t_p, t_{p+1}]$  for  $p = 1, \dots, n_{ij}$  we get:

$$z_{n_{ij}-n_{pi}} < x_p < z_{n_{ip}+1}, \quad p = i+1, \dots, j-1.$$

If all  $z_p$  are distinct, then by applying Lemma 2.3 and Lemma 2.4 to the space  $G^{ij}$  we conclude that  $g \equiv 0$  on  $[x_i, x_j]$  and we are ready.

But if only  $r < n_{ij}$  of the zeros  $\{z_p\}_{p=1}^{n_{ij}}$  are distinct, then the function  $g$  has  $s = n_{ij} - r$  isolated double zeros  $z_{p_1} = z_{p_1+1}, z_{p_2} = z_{p_2+1}, \dots, z_{p_s} = z_{p_s+1}$ . We denote the  $r$  distinct points from  $\{z_p\}_{p=1}^{n_{ij}}$  by  $v_1, \dots, v_r$ , arranged ascendingly. We choose  $\varepsilon > 0$  such that

$$\begin{aligned} \text{(i)} \quad & z_{n_{ij}-n_{pi}} + \varepsilon < x_p < z_{n_{ip}+1} - \varepsilon, \quad p = i+1, \dots, j-1, \\ \text{(ii)} \quad & \varepsilon < \frac{1}{4} \min_{p=0, \dots, r} (v_{p+1} - v_p), \quad v_0 = x_i, \quad v_{r+1} = x_j, \end{aligned}$$

is satisfied. In case  $v_1 = x_i$  or  $v_r = x_j$  or  $v_1 = x_i$  and  $v_r = x_j$  we only determine the minimum for  $p = 1, \dots, r$  or for  $p = 0, \dots, r-1$  or for  $p = 1, \dots, r-1$ , respectively.

We now add to each double zero  $z_{p_1}, z_{p_2}, \dots, z_{p_s}$  one further point  $z_{p_1} + \varepsilon, z_{p_2} + \varepsilon, \dots, z_{p_s} + \varepsilon$  and get a new set  $\{v_1, \dots, v_r, z_{p_1} + \varepsilon, \dots, z_{p_s} + \varepsilon\}$  consisting of  $n_{ij}$  points. We denote the elements of this set by  $\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_{n_{ij}}$ , arranged ascendingly. Then we get:

$$\tilde{v}_{n_{ij}-n_{pi}} < x_p < \tilde{v}_{n_{ip}+1}, \quad p = i+1, \dots, j-1.$$

Let  $U(z_{p_k})$  be a sufficiently small neighborhood of  $z_{p_k}$  on which the function  $g$  has only the zero  $z_{p_k}$  (remember that  $z_{p_k}$  is an isolated double zero of  $g$ ). Then by applying

Lemma 2.3 and Lemma 2.4 to the space  $G^{ij}$  there exists exactly one  $\tilde{g} \in G^{ij}$  such that

- (i)  $\tilde{g}(z_{p_k} + \varepsilon) = 0, k = 1, \dots, s;$
- (ii)  $\tilde{g}(z_p) = 0, p \notin \{p_1, p_1 + 1, p_2, p_2 + 1, \dots, p_s, p_s + 1\};$
- (iii)  $\tilde{g}(z_{p_k}) = \text{sgn } g(x), x \in U(z_{p_k}) \setminus \{z_{p_k}\}, k = 1, \dots, s.$

Then the functions  $g$  and  $\tilde{g}$  have at least  $n_{ij} - 2s$  common zeros and for sufficiently small  $c > 0$  the function  $g - c\tilde{g}$  has at least two distinct zeros on  $[z_{p_k} - \varepsilon, z_{p_k} + \varepsilon]$  for  $k = 1, \dots, s$ . Thus  $g - c\tilde{g}$  has at least  $n_{ij}$  distinct zeros, denoted by  $w_1, \dots, w_{n_{ij}}$ , arranged ascendingly. Then because of the choice of  $\varepsilon$  we get:

$$w_{n_{ij}-n_{pj}} < x_p < w_{n_{ip}+1}, \quad p = i + 1, \dots, j - 1.$$

This is true, since  $w_p = z_p$  for  $p \notin \{p_1, p_1 + 1, p_2, p_2 + 1, \dots, p_s, p_s + 1\}$  and  $\{w_{p_k}, w_{p_k+1}\} \subset [z_{p_k} - \varepsilon, z_{p_k} + \varepsilon] = [z_{p_{k+1}} - \varepsilon, z_{p_{k+1}} + \varepsilon]$  for  $k = 1, \dots, s$ .

Now we may apply Lemma 2.3 and Lemma 2.4 to  $G^{ij}$  and we get that  $g - c\tilde{g} \equiv 0$  on  $[x_i, x_j]$ . But this is a contradiction, since  $(g - c\tilde{g})(z_{p_k}) \neq 0$  for  $k = 1, \dots, s$ .

Thus we have proved that all of the zeros  $\{z_p\}_{p=1}^{n_{ij}}$  are distinct and we get that  $g \equiv 0$  on  $[x_i, x_j]$  as shown above. Since  $g \in P_G(f)$  has been chosen arbitrarily, all best approximations of  $f$  from  $G$  vanish identically on  $[x_i, x_j]$ . Therefore,  $f - g$  has  $n_{ij} + 1$  alternating extreme points on  $[x_i, x_j]$  for any  $g \in P_G(f)$ .

By applying Lemma 2.5 we now are able to prove a characterization theorem for a best approximation from  $G$ , in case  $G$  is in  $\tilde{\mathcal{V}}_n$ .

**THEOREM 2.6.** *Let  $G$  be in  $\tilde{\mathcal{V}}_n$ . A function  $g$  in  $G$  is a best approximation for  $f$  in  $C[a, b]$  from  $G$  if and only if there exists a knot interval  $[x_i, x_j]$  such that  $f - g$  has at least  $n_{ij} + 1$  alternating extreme points on  $[x_i, x_j]$ .*

*Proof.* Let  $f \in C[a, b]$  and  $g \in P_G(f)$ . Then by Lemma 2.5 there exists a knot interval  $[x_i, x_j]$  such that  $f - g$  has at least  $n_{ij} + 1$  alternating extreme points on  $[x_i, x_j]$ .

Conversely let  $f \in C[a, b]$  and  $g \in G$  and  $[x_i, x_j] \subset [a, b]$  be a knot interval such that  $f - g$  has at least  $n_{ij} + 1$  alternating extreme points on  $[x_i, x_j]$  i.e. there exist  $n_{ij} + 1$  points  $x_i \leq t_1 < t_2 < \dots < t_{n_{ij}+1} \leq x_j$  such that  $\varepsilon(-1)^p(f - g)(t_p) = \|f - g\|$  for  $p = 1, \dots, n_{ij} + 1$ ,  $\varepsilon = \pm 1$ . If  $g \notin P_G(f)$ , then there exists a  $\tilde{g} \in G$  such that  $\varepsilon(-1)^p(f - \tilde{g})(t_p) < \varepsilon(-1)^p(f - g)(t_p)$  for  $p = 1, \dots, n_{ij} + 1$  and, therefore,

$$\varepsilon(-1)^p(\tilde{g} - g)(t_p) > 0.$$

Hence  $\tilde{g} - g$  has  $n_{ij}$  changes of sign on  $(x_i, x_j)$ . Since  $G^{ij}$  is weak Chebyshev by Theorem 1.4 in [15], this is a contradiction.

**3. The characterization theorem.** Now we are able to give a characterization of the existence of continuous selections for all  $G$  in  $\mathcal{V}_n$ . We will prove the following statement:

**THEOREM 3.1.** *Let  $G$  be in  $\mathcal{V}_n$ . Then there exists a continuous selection for  $G$  if and only if  $G$  is in  $\tilde{\mathcal{V}}_n$ .*

At first we will show the nonexistence of a continuous selection for  $G$ , in case  $G$  is in  $\mathcal{V}_n$ , but not in  $\tilde{\mathcal{V}}_n$ . For proving this we need the following fundamental lemma established by Lazar, Morris and Wulbert [7].

**LEMMA 3.2.** *If  $s$  is a continuous selection of  $C[a, b]$  into  $G$  and  $f$  is in  $C[a, b]$ ,  $\|f\| = 1$  and  $0$  is in  $P_G(f)$ , then there is a  $g_0$  in  $P_G(f)$  such that*

- (i) *for every  $x$  in  $\text{bd } Z(P_G(f)) \cap f^{-1}(1)$  and every  $g$  in  $P_G(f)$  there is a neighborhood  $U$  of  $x$  for which  $g_0 \geq g$  on  $U$  and*
- (ii) *for every  $x$  in  $\text{bd } Z(P_G(f)) \cap f^{-1}(-1)$  and every  $g$  in  $P_G(f)$  there is a neighborhood  $V$  of  $x$  for which  $g_0 \leq g$  on  $V$ .*

Here  $Z(P_G(f)) = \{x \in [a, b] | g(x) = 0 \text{ for all } g \in P_G(f)\}$ .

We will now show by two lemmas that in case  $G$  is in  $\mathcal{V}_n$  but not in  $\check{\mathcal{V}}_n$  there does not exist any continuous selection for  $G$ . This proves the one part of Theorem 3.1.

**LEMMA 3.3.** *Let  $G$  be in  $\mathcal{V}_n$ . Let no  $g$  in  $G$  have two separated zero intervals. If there exist some  $i, j \in \{0, 1, \dots, s\}$  and a function  $g_0$  in  $\bar{G}_{ij}$  with  $|\text{bd } Z(g_0)| > m_{ij}$ , then there does not exist any continuous selection for  $G$ .*

*Proof.* Since by Lemma 2.2 no  $g \in G$  has more than  $n$  separated, nonvanishing zeros and all  $x \in [a, b]$  are nonvanishing, there exists an integer  $p$  such that  $|\text{bd } Z(g)| \leq p$  for all  $g \in G$ . Now let  $\bar{G}_{ij}$  be such a subspace of  $G$  having an element  $g_0$  with  $|\text{bd } Z(g_0)| = m > m_{ij}$ . Without loss of generality we may assume that

$$m = |\text{bd } Z(g_0)| \geq |\text{bd } Z(g)| \quad \text{for all } g \in \bar{G}_{ij}.$$

This is possible because  $|\text{bd } Z(g)| \leq p$  for all  $g \in \bar{G}_{ij}$ . Furthermore we set  $\|g_0\| = 1$ .

Let  $I = [x_k, x_l] \supset [x_i, x_j]$  be the maximal zero interval of  $g_0$ . We now only treat the case  $a < x_k < x_l < b$ , since the cases  $a = x_k$  and  $x_l = b$  follow analogously. Furthermore, we only treat the case that  $x_k$  is in the closure of  $\{x \in [a, b] \mid g_0(x) < 0\}$  and  $x_l$  is in the closure of  $\{x \in [a, b] \mid g_0(x) > 0\}$ . That means that  $g_0(x) < 0$  on  $[x_k - \delta, x_k)$  and  $g_0(x) > 0$  on  $(x_l, x_l + \delta]$  for  $\delta > 0$  sufficiently small. By hypothesis, the function  $g_0$  has only the zero interval  $I$  and, therefore, exactly  $m - 2$  distinct zeros

$$a = z_0 \leq z_1 < z_2 < \dots < z_r < x_k < x_l < z_{r+1} < \dots < z_{m-2} \leq z_{m-1} = b$$

on  $[a, b] \setminus I$ .

We define  $m$  points  $\{t_p\}_{p=0}^{m-1}$  by

$$\begin{aligned} t_p &= (z_p + z_{p+1})/2, \quad \text{for } p = 0, \dots, r-1, \\ t_r &= (z_r + x_k)/2, \\ t_{r+1} &= (x_l + z_{r+1})/2, \\ t_p &= (z_{p-1} + z_p)/2, \quad \text{for } p = r+2, \dots, m-1. \end{aligned}$$

We choose  $\varepsilon > 0$  such that

$$\{z_1, z_2, \dots, z_{m-2}, x_k, x_l\} \cap [t_p - \varepsilon, t_p + \varepsilon] = \emptyset \quad \text{for } p = 1, \dots, m-2.$$

We now construct a function  $f \in C[a, b]$  as follows:

- (a) Let  $f$  have  $n_{p+1} + 1$  alternating extreme points on  $(x_p, x_{p+1})$  for  $p = k, \dots, l-1$  with  $|f| = 1$  on these points.
- (b)  $f(x_k) = 1$  and  $f(x) = 1$  for all  $x \in [x_l, t_{r+1}]$ .
- (c) If  $z_1 > a$ , then we set  $f(x) = \text{sgn } g_0(t_0)$  for all  $x \in [a, t_0]$ . If  $z_1 = a$ , then we set  $f(a) = -\text{sgn } g_0(t_1)$  and  $f(x) = \text{sgn } g_0(t_1)$  for all  $x \in [t_1 - \varepsilon, t_1 + \varepsilon]$ .
- (d) If  $z_{m-2} < b$ , then we set  $f(x) = \text{sgn } g_0(t_{m-1})$  for all  $x \in [t_{m-1}, b]$ . If  $z_{m-2} = b$ , then we set  $f(b) = -\text{sgn } g_0(t_{m-2})$ .
- (e) If for some  $p \in \{1, \dots, m-2\}$   $z_p$  is a zero with a change of sign on  $(a, b)$ , then we set

$$f(x) = \begin{cases} \text{sgn } g_0(t_p) & \text{for all } x \in [z_p, t_p] & \text{if } p \leq r, \\ \text{sgn } g_0(t_{p+1}) & \text{for all } x \in [z_p, t_{p+1}] & \text{if } p \geq r+1. \end{cases}$$

If for some  $p \in \{1, \dots, m-2\}$   $z_p$  is a double zero on  $(a, b)$ , then we set

$$f(z_p) = \begin{cases} -\operatorname{sgn} g_0(t_p) & \text{if } p \leq r, \\ -\operatorname{sgn} g_0(t_{p+1}) & \text{if } p \geq r+1, \end{cases}$$

$$f(x) = \begin{cases} \operatorname{sgn} g_0(t_p) & \text{for all } x \in [t_p - \varepsilon, t_p + \varepsilon] & \text{if } p \leq r, \\ \operatorname{sgn} g_0(t_{p+1}) & \text{for all } x \in [t_{p+1} - \varepsilon, t_{p+1} + \varepsilon] & \text{if } p \geq r+1. \end{cases}$$

(f)  $\max\{-1 + g_0(x), -1\} \leq f(x) \leq \min\{1 + g_0(x), 1\}$  for all  $x \in [a, b]$ .

Because of (a) every  $g \in P_G(f)$  vanishes identically on  $[x_k, x_l]$ . From  $\|f\| = 1$  on  $[a, b]$  it follows immediately that  $0 \in P_G(f)$  and, because of  $\|f - g_0\| \leq 1$ ,  $g_0 \in P_G(f)$ , too. Therefore,  $x_k, x_l \in \operatorname{bd} Z(P_G(f))$ .

At first we show that  $Z_d(g_0) \subset Z(P_G(f))$ , in case  $Z_d(g_0) \neq \emptyset$ . We assume that there is a function  $\tilde{g} \in P_G(f)$  and an  $\tilde{x} \in Z_d(g_0)$  such that  $\tilde{g}(\tilde{x}) \neq 0$ .

Since  $\tilde{g} \in P_G(f)$ , it follows from the definition of  $f$ :

$$\begin{aligned} \tilde{g} \cdot g_0 &\geq 0 && \text{on } [t_p - \varepsilon, t_p] \text{ for } p = 0, \dots, m-1 && \text{if } a < z_1, z_{m-2} < b, \\ \tilde{g} \cdot g_0 &\geq 0 && \text{on } [t_p - \varepsilon, t_p] \text{ for } p = 1, \dots, m-1 && \text{and} \\ \tilde{g}(a) &= 0 && \text{or } \operatorname{sgn} \tilde{g}(a) = -\operatorname{sgn} g_0(t_1) && \text{if } a = z_1, z_{m-2} < b, \\ (*) \quad \tilde{g} \cdot g_0 &\geq 0 && \text{on } [t_p - \varepsilon, t_p] \text{ for } p = 0, \dots, m-2 && \text{and} \\ \tilde{g}(b) &= 0 && \text{or } \operatorname{sgn} \tilde{g}(b) = -\operatorname{sgn} g_0(t_{m-2}) && \text{if } a < z_1, z_{m-2} = b, \\ \tilde{g} \cdot g_0 &\geq 0 && \text{on } [t_p - \varepsilon, t_p] \text{ for } p = 1, \dots, m-2 && \text{and} \\ \tilde{g}(a) &= 0 && \text{or } \operatorname{sgn} \tilde{g}(a) = -\operatorname{sgn} g_0(t_1) && \text{and} \\ \tilde{g}(b) &= 0 && \text{or } \operatorname{sgn} \tilde{g}(b) = -\operatorname{sgn} g_0(t_{m-2}) && \text{if } a = z_1, z_{m-2} = b. \end{aligned}$$

Now we consider for sufficiently small  $c > 0$  the function

$$\bar{g} = g_0 + c\tilde{g}.$$

Then  $\bar{g} \equiv 0$  on  $[x_k, x_l]$  and it follows immediately from (\*) that  $\bar{g}$  has at least as many changes of sign on  $(a, x_k)$  and on  $(x_l, b)$  as the function  $g_0$ . Additionally the points  $x_k$  and  $x_l$  are boundary points of  $Z(\bar{g})$  and, if  $g_0(a) = 0$  or  $g_0(b) = 0$ ,  $Z(\bar{g})$  has further boundary points on neighborhoods of  $a$  or  $b$ .

Since  $\tilde{x} \in Z_d(g_0)$ , there exists a  $p \in \{1, \dots, m-2\}$  such that  $\tilde{x} = z_p$ . Without loss of generality let  $z_p > x_l$ . Then  $t_p < z_p < t_{p+1}$  and because of  $\tilde{g} \cdot g_0 \geq 0$  on  $[t_p - \varepsilon, t_p] \cup [t_{p+1} - \varepsilon, t_{p+1}]$  and  $\operatorname{sgn} \tilde{g}(z_p) = -\operatorname{sgn} g_0(t_{p+1})$  the function  $\bar{g}$  has two changes of sign on a neighborhood of  $z_p$ . But this is also true for all double zeros of  $g_0$  which are no zeros of  $\tilde{g}$ . In this way we get

$$|\operatorname{bd} Z(\bar{g})| > |\operatorname{bd} Z(g_0)| = m.$$

But this is a contradiction, because  $\bar{g} \in \bar{G}_{ij}$  and  $|\operatorname{bd} Z(g)| \leq m$  for all  $g \in \bar{G}_{ij}$ . Thus we have shown that  $Z_d(g_0) \subset Z(P_G(f))$ . Since  $Z_d(g_0) \subset \operatorname{bd} Z(g_0)$ , we even get  $Z_d(g_0) \subset \operatorname{bd} Z(P_G(f)) \cap (f^{-1}(1) \cup f^{-1}(-1))$ .

Now we distinguish the following cases: *First:*  $Z_d(g_0) \neq \emptyset$ . Without loss of generality we may assume that there is an  $\tilde{x} \in Z_d(g_0)$  such that  $\tilde{x} > x_l$  and  $f(\tilde{x}) = 1$ . We now apply Lemma 3.2: If there exists a continuous selection, then there exists a  $g_1 \in P_G(f)$  such that

- (i) for  $x_l$  and  $g_0$  there is a neighborhood  $U$  of  $x_l$  for which  $g_1 \geq g_0$  on  $U$  and
- (ii) for  $\tilde{x}$  and 0 there is a neighborhood  $V$  of  $\tilde{x}$  for which  $g_1 \geq 0$  on  $V$ .



Since  $g_1 \cong g_0 > 0$  on  $(x_i, x_i + \delta)$  for a sufficiently small  $\delta > 0$ , by hypothesis,  $g_1$  has no zero interval in  $[x_i, b]$ . Therefore,  $\tilde{x}$  is an isolated zero of  $g_1$  and because of (ii)  $\tilde{x} \in Z_d(g_1)$ . Then it is easy to verify that for all sufficiently great positive numbers  $d$  the function  $g_0 + dg_1$  satisfies

$$|\text{bd } Z(g_0 + dg_1)| \cong |\text{bd } Z(g_0)|.$$

Since  $\tilde{x} \in Z_d(g_0)$  and  $\tilde{x} > x_i$ , there is a  $p \in \{r + 1, \dots, m - 2\}$  with  $\tilde{x} = z_p$ . Then it follows from the definition of  $f$  that  $g_1 \cong 0$  on  $[t_p - \varepsilon, t_p]$  and on  $[t_{p+1} - \varepsilon, t_{p+1}]$ . Since  $g_0 + dg_1$  has no zero interval in  $[x_i, b]$  for any  $d > 0$ , the function  $g_0 + dg_1$  has for some  $d > 0$  at least two changes of sign on  $(t_p, t_{p+1})$  and, therefore, we get

$$|\text{bd } Z(g_0 + dg_1)| \cong 2 + |\text{bd } Z(g_0)|.$$

But this is a contradiction, because  $g_0 + dg_1 \in \bar{G}_{ij}$ . Thus there does not exist any continuous selection in this case.

*Second:*  $Z_d(g_0) = \emptyset$  and  $\{a, b\} \cap Z(g_0) = \emptyset$ . Then  $g_0$  has exactly  $m - 2$  zeros with changes of sign on  $(a, b) \setminus I$  and, by hypothesis, one further change of sign on  $(x_k - \delta, x_l + \delta)$  for sufficiently small  $\delta > 0$ . Therefore,  $g_0$  has exactly  $m - 1$  changes of sign on  $(a, b)$ . Since  $\dim \bar{G}_{ij} = m_{ij}$ , there exist  $n - m_{ij}$  functions in  $G$  linearly independent on  $[x_i, x_j]$  and by Lemma 1.4, therefore, a function  $h \in G$  with at least  $n - m_{ij} - 1$  changes of sign on  $(x_i, x_j)$ .

We now distinguish: If  $n - m_{ij} - 1$  is an odd number, then for sufficiently small  $c > 0$  either the function  $g_0 + ch$  or the function  $g_0 - ch$  has at least  $n - m_{ij} - 1$  changes of sign on  $(x_i, x_j)$  and further  $m$  changes of sign on  $(a, b)$  (at least one on a neighborhood of  $x_k$  and another one on a neighborhood of  $x_l$ ) and, therefore, at least  $n - m_{ij} - 1 + m \cong n - m_{ij} - 1 + m_{ij} + 1 = n$  changes of sign on  $(a, b)$ . But this is a contradiction of the hypothesis that  $G$  is weak Chebyshev.

Therefore, there remains only the case that  $n - m_{ij} - 1$  is an even number. In this case we first apply Lemma 3.2 again: If there exists a continuous selection, then there exists a  $g_1 \in P_G(f)$  such that

- (i) for  $x_k$  and 0 there is a neighborhood  $U$  of  $x_k$  for which  $g_1 \cong 0$  on  $U$  and
- (ii) for  $x_l$  and  $g_0$  there is a neighborhood  $V$  of  $x_l$  for which  $g_1 \cong g_0$  on  $V$ .

Since  $g_1 \cong g_0 > 0$  on  $(x_i, x_i + \delta)$ , by hypothesis,  $g_1$  has no zero interval in  $[x_i, b]$ . Let  $[x_h, x_l]$  be the maximal zero interval of  $g_1$ . We denote all distinct zeros of  $g_1$  on  $[a, b] \setminus [x_h, x_l]$  by  $a = y_0 \cong y_1 < \dots < y_u = x_h < x_l = y_{u+1} < y_{u+2} < \dots < y_{v-1} \cong y_v = b$ . We define:

$$\tilde{d} = \frac{1}{2} \min_{\substack{p=0, \dots, v-1 \\ p \neq u}} \|g_1|_{[y_p, y_{p+1}]}\|$$

where  $[y_0, y_1]$  or  $[y_{v-1}, y_v]$  or  $[y_0, y_1]$  and  $[y_{v-1}, y_v]$  are omitted, if  $y_1 = a$  or  $y_{v-1} = b$  or  $y_1 = a$  and  $y_{v-1} = b$ , respectively. Then  $\tilde{d} > 0$ . We set  $d = \min(\frac{1}{2}, \tilde{d})$  and define the function  $g_2$  by  $g_2 = g_1 - dg_0$  (remember that  $\|g_0\| = 1$ ). Then from the definition of  $f$  it follows that

$$\begin{aligned} g_1 \cdot g_0 \cong 0 & \quad \text{on } [z_p, t_p] \text{ for } p = 0, \dots, r, \quad \text{and} \\ & \quad \text{on } [z_p, t_{p+1}] \text{ for } p = r + 1, \dots, m - 2 \quad \text{and} \\ & \quad \text{on } [x_i, t_{r+1}]. \end{aligned}$$

Therefore,  $Z(g_2)$  has at least one boundary point on each interval  $(z_p, z_{p+1})$  for

$p = 0, \dots, r-1$  and on  $(x_b, z_{r+1}]$  and on  $(z_p, z_{p+1}]$  for  $p = r+1, \dots, m-3$ . These are at least  $m-2$  points. Moreover, we get two further boundary points on neighborhoods of  $x_k$  and  $x_l$ . Therefore,  $|\text{bd } Z(g_2)| \geq |\text{bd } Z(g_0)|$  and from  $|\text{bd } Z(g_2)| \leq m$  it finally follows that  $|\text{bd } Z(g_2)| = m$ .

We now distinguish once more: If  $Z_d(g_2) \neq \emptyset$ , then we conclude as in the first case by replacing  $g_0$  by  $g_2$  and get in this way that there does not exist any continuous selection for  $G$ . If  $Z_d(g_2) = \emptyset$ , then  $g_2$  has exactly  $m-2$  changes of sign on  $(a, b) \setminus I$ . Since  $g_1 \geq 0$  on the neighborhood  $U$  of  $x_k$ , we get  $g_2 > 0$  on  $[x_k - \delta, x_k]$  for sufficiently small  $\delta > 0$ . Since  $g_1 > dg_0$  on  $(x_b, x_l + \varepsilon]$  for sufficiently small  $\varepsilon > 0$ , we also get  $g_2 > 0$  on  $(x_b, x_l + \varepsilon]$ . By hypothesis,  $n - m_{ij} - 1$  is an even number. But if we replace  $g_0$  by  $g_2$ , then we may conclude in the same way as in the case that  $n - m_{ij} - 1$  is an odd number. Therefore, for sufficiently small  $c > 0$  either the function  $g_2 - ch$  or the function  $g_2 + ch$  has at least  $n - m_{ij} - 1 + m - 2 + 2 \geq n - m_{ij} - 1 + m_{ij} + 1 = n$  changes of sign on  $(a, b)$  (at least one on a neighborhood of  $x_k$  and another one on a neighborhood of  $x_l$ ). But this is a contradiction of the hypothesis that  $G$  is weak Chebyshev.

*Third:*  $Z_d(g_0) = \emptyset$  and  $a \in Z(g_0)$ ,  $b \notin Z(g_0)$  or  $a \notin Z(g_0)$ ,  $b \in Z(g_0)$ . Without loss of generality we may assume that  $a \notin Z(g_0)$ ,  $b \in Z(g_0)$ . We distinguish two cases:

(i)  $g(b) = 0$  for all  $g \in P_G(f)$ . Then  $b \in \text{bd } Z(P_G(f)) \cap (f^{-1}(1) \cup f^{-1}(-1))$ . Without loss of generality let  $f(b) = 1$ . Then  $g_0 \leq 0$  on a neighborhood of  $b$  by hypothesis. If there is a continuous selection, then by Lemma 3.2 there exists a function  $g_1 \in P_G(f)$  and a neighborhood  $U$  of  $b$  for which  $g_1 \geq 0$  on  $U$  and a neighborhood  $V$  of  $x_l$  for which  $g_1 \geq g_0$  on  $V$ . Therefore,  $g_1$  has no zero interval in  $[x_b, b]$ . Then it is easy to verify that for sufficiently great  $d > 0$  the set  $Z(g_0 + dg_1)$  has at least  $m+1$  boundary points. But this is a contradiction, because  $g_0 + dg_1 \in \tilde{G}_{ij}$  and, therefore,  $|\text{bd } Z(g_0 + dg_1)| \leq m$ .

(ii) There exists a  $\tilde{g} \in P_G(f)$  with  $\tilde{g}(b) \neq 0$ . Then for some constant  $c > 0$  the set  $Z(g_0 + c\tilde{g})$  has at least  $m$  boundary points and  $\{a, b\} \cap Z(g_0 + c\tilde{g}) = \emptyset$ . Then we conclude as in the first or the second case by replacing  $g_0$  by  $g_0 + c\tilde{g}$ .

*Fourth:*  $Z_d(g_0) = \emptyset$  and  $\{a, b\} \subset Z(g_0)$ . If  $g(b) = 0$  for all  $g \in P_G(f)$ , then we can conclude as in the third case.

Otherwise, there is a  $\tilde{g} \in P_G(f)$  with  $\tilde{g}(b) \neq 0$ . Then for some constant  $c > 0$  the function  $g_0 + c\tilde{g}$  satisfies  $|\text{bd } Z(g_0 + c\tilde{g})| \geq m$  and  $b \notin Z(g_0 + c\tilde{g})$ . Then we may conclude as in the first three cases by replacing  $g_0$  by  $g_0 + c\tilde{g}$ .

LEMMA 3.4. *Let  $G$  be in  $\mathcal{V}_n$ . Let  $\tilde{g}$  be a function in  $G$  having two separated zero intervals. Then there does not exist any continuous selection for  $G$ .*

*Proof.* Let  $\tilde{g} \in G$  having two separated knot intervals  $[x_{i-1}, x_i]$  and  $[x_j, x_{j+1}]$  with  $i < j$ . Without loss of generality we may assume that there does not exist any  $g \in G$  such that  $g \equiv 0$  on  $[x_{i-1}, x_i] \cup [x_j, x_{j+1}]$ ,  $g \neq 0$  on  $[x_i, x_j]$  and  $g$  has a zero interval in  $[x_i, x_j]$ . Such a choice of the knots  $x_i, x_j$  is always possible. Then in particular  $\tilde{g}$  has no zero interval in  $[x_i, x_j]$ : We define:  $\tilde{G} = \tilde{G}_{i-1, i} \cap \tilde{G}_{j, j+1}$ . Then  $\dim \tilde{G} \geq 1$ , since  $\tilde{g} \in \tilde{G}$  and it follows immediately that no  $g \in \tilde{G}$ ,  $g \neq 0$  on  $[x_i, x_j]$ , has a zero interval in  $[x_i, x_j]$ . Since  $\tilde{G} \subset G$  and  $G$  is weak Chebyshev, there exists a nonnegative number  $r \leq n-1$  and a function  $g_0 \in \tilde{G}$  such that  $g_0$  has exactly  $r$  changes of sign  $x_i < z_1 < z_2 < \dots < z_r < x_j$  and, furthermore, no  $g \in \tilde{G}$  has more than  $r$  changes of sign on  $(x_i, x_j)$ . We may assume that  $g_0 \geq 0$  on a neighborhood of  $x_i$  and  $\|g_0\| \leq 1$ . Then we choose  $r$  distinct points  $\{v_p\}_{p=1}^r$  satisfying  $x_i < z_1 < v_1 < z_2 < v_2 < \dots < z_r < v_r < x_j$  and we choose  $\varepsilon > 0$  such that  $\{z_1, z_2, \dots, z_r, x_j\} \cap [v_p - \varepsilon, v_p + \varepsilon] = \emptyset$ ,  $p = 1, \dots, r$ , and  $x_i + \varepsilon < z_1$ .

We now construct a function  $f \in C[a, b]$  as follows:

(a)  $f(x) = g_0(x)$  for all  $x \in [a, x_{i-1}] \cup [x_{j+1}, b]$ .

(b) Let  $f$  have  $n_i + 1$  alternating extreme points on  $(x_{i-1}, x_i)$  and  $n_{i+1} + 1$  alternating extreme points on  $(x_j, x_{j+1})$  with  $|f| = 1$  on these points.

- (c)  $f(x) = 1$  for all  $x \in [x_i, x_i + \varepsilon]$   
 $f(x) = (-1)^p$  for all  $x \in [v_p - \varepsilon, v_p + \varepsilon]$  for  $p = 1, \dots, r$   
 $f(x_j) = (-1)^{r+1}$ .

- (d)  $\max\{-1 + g_0(x), -1\} \leq f(x) \leq \min\{1 + g_0(x), 1\}$  for all  $x \in [x_{i-1}, x_{j+1}]$ .

Then  $\|f - 0\| = \|f - g_0\| = 1$ . Since  $f - 0$  has  $n_i + 1$  alternating extreme points on  $[x_{i-1}, x_i]$  and  $G^{i-1}$  is Chebyshev with dimension  $n_i$ , it follows from the well-known alternation theorem for the Chebyshev spaces, that all  $g \in P_G(f)$  vanish identically on  $[x_{i-1}, x_i]$ . Therefore,  $0 \in P_G(f)$  and  $g_0 \in P_G(f)$ , too. Now let  $g \in P_G(f)$ ,  $g \neq 0$  on  $[x_i, x_j]$ . Then  $g \equiv 0$  on  $[x_{i-1}, x_i] \cup [x_j, x_{j+1}]$  and, therefore,  $g \in \tilde{G}$ . By hypothesis,  $g$  has no zero interval in  $[x_i, x_j]$ . Then it follows from the definition of  $f$  that  $g$  has at least  $r$  changes of sign on  $(x_i, x_j)$  and by hypothesis, therefore, exactly  $r$  changes of sign on  $(x_i, x_j)$ . Thus  $(-1)^r g \geq 0$  on a neighborhood of  $x_j$ .

Since  $g_0$  has no zero interval in  $[x_i, x_j]$ , we get  $x_i, x_j \in \text{bd } Z(P_G(f))$ . We now apply Lemma 3.2: If there exists a continuous selection for  $G$ , then there exists a  $\bar{g} \in P_G(f)$  such that for  $x_i$  and  $g_0$  there is a neighborhood  $U$  of  $x_i$  for which  $\bar{g} \geq g_0$  on  $U$  and for  $x_j$  and  $0 \in P_G(f)$  there is a neighborhood  $V$  of  $x_j$  for which  $(-1)^{r+1} \bar{g} \geq 0$  on  $V$ .

Since  $\bar{g} \geq g_0$  on  $U$ , we get  $\bar{g} \neq 0$  and, therefore,  $\bar{g}$  has no zero interval in  $[x_i, x_j]$ . As shown above there is a neighborhood  $W$  of  $x_j$  for which  $(-1)^r \bar{g} \geq 0$  on  $W$ . Thus there is a  $\bar{x} \in W \cap V$  with  $(-1)^r \bar{g}(\bar{x}) > 0$  and hence we get a contradiction to Lemma 3.2. Therefore, there does not exist any continuous selection for  $G$ .

If  $G$  is in  $\mathcal{V}_n$  but not in  $\tilde{\mathcal{V}}_n$ , then it follows from Lemma 3.3 and Lemma 3.4 that there does not exist any continuous selection for  $G$ . Therefore, we have only to treat the case that  $G$  is in  $\tilde{\mathcal{V}}_n$ . In this case we are able to show the existence of a continuous selection. For constructing such a selection we need the following two lemmas:

LEMMA 3.5. (Nürnbergger and Sommer [9]). *Let  $G$  be in  $\mathcal{W}_n$  and  $f$  be in  $C[a, b]$ . If  $g_1, g_2$  in  $P_G(f)$  are two AE's for  $f$ , then at least one of the following is true:*

- (i)  $g_1 - g_2$  has at least  $n + 1$  distinct zeros on  $[a, b]$ ;  
(ii)  $g_1 - g_2$  has at least  $n + 2$  zeros on  $[a, b]$  counting multiplicities.

LEMMA 3.6. *Let  $G$  be in  $\tilde{\mathcal{V}}_n$ . Then for any  $j \in \{0, 1, \dots, s\}$  and any  $g$  in  $\tilde{G}_{0j}$  with  $g \neq 0$  on  $[x_j, x_{j+1}]$  the following is true:*

$$|Z^*(g)| \leq m_{0j} + 1.$$

*Proof.* We assume that there is a  $j \in \{0, 1, \dots, s\}$  and a  $g_0 \in \tilde{G}_{0j}$ ,  $g_0 \neq 0$  on  $[x_j, x_{j+1}]$ , satisfying  $|Z^*(g_0)| \geq m_{0j} + 2$ . Since  $G \in \tilde{\mathcal{V}}_n$  and  $g_0 \equiv 0$  on  $[a, x_j]$ ,  $g_0$  has no zero interval in  $[x_j, b]$ . Therefore, we may assume that  $g_0$  has exactly  $p$  distinct zeros on  $[x_j, b]$  and  $g_0(b) = 0$ . The case  $g_0(b) \neq 0$  follows analogously. Let  $\bar{x} = \max\{x \in [x_j, b] \mid g_0(x) = 0\}$ . Let  $x_j < y_1 < y_2 < \dots < y_r < b$  be all the zeros with changes of sign and  $x_j < z_1 < z_2 < \dots < z_t < b$  be all double zeros of  $g_0$ . Therefore,  $p = r + t + 2$  and because of  $|Z^*(g_0)| \geq m_{0j} + 2$  we get  $r + 2t + 2 \geq m_{0j} + 2$ .

We choose  $m_{0j} - r - 1$  points

$$\max(\bar{x}, x_{s-1}) < y_{r+1} < \dots < y_{m_{0j}-1} < b.$$

Since  $\tilde{G}_{0j}$  is weak Chebyshev by Lemma 2.1, there is by Theorem 1.3 a nonzero  $\bar{g} \in \tilde{G}_{0j}$  such that

$$\begin{aligned} (-1)^i \bar{g}(x) &\geq 0, & y_{i-1} < x < y_i, & i = 1, \dots, m_{0j} \\ y_0 &= a, & y_{m_{0j}} &= b. \end{aligned}$$

Without loss of generality let  $\bar{g} \cdot g_0 \geq 0$  on  $[a, y_{r+1}]$ . Since  $G \in \tilde{\mathcal{V}}_n$ ,  $\bar{g}$  has no two separated zero intervals. Therefore,  $\bar{g} \neq 0$  on  $[x_{s-1}, b]$ , because  $\bar{g} \equiv 0$  on  $[a, x_u]$  with  $x_u \geq x_j$ .

We distinguish: *First*:  $\bar{g}(z_i) \neq 0$  for  $i = 1, \dots, t$ . Then for sufficiently small  $c > 0$  the function  $g_0 - c\bar{g}$  has no zero interval in  $[x_j, b]$  and at least  $1 + r + 2t \geq m_{0j} + 1$  distinct zeros on  $[x_j, b]$ . Since  $g_0 - c\bar{g} \in \bar{G}_{0j}$ , this is a contradiction of the hypothesis that  $|\text{bd } Z(g_0 - c\bar{g})| \leq m_{0j}$ .

*Second*: There is some  $i_0 \in \{1, \dots, t\}$  such that  $\bar{g}(z_{i_0}) = 0$ . Then for sufficiently small  $c > 0$  the function  $\bar{g} - cg_0$  has no zero interval in  $[x_j, b]$  and at least one zero at  $x_j$ ,  $r$  zeros at  $y_1, y_2, \dots, y_r$ , one zero on a neighborhood of  $y_p$  for  $p = r + 1, \dots, m_{0j} - 1$  and one zero at  $z_{i_0}$ . These are at least  $m_{0j} + 1$  distinct zeros on  $[x_j, b]$  and, therefore, we get a contradiction again.

Now we are able to construct a continuous selection for  $G$ , in case  $G$  is in  $\check{\mathcal{V}}_n$ . For this we need local AE's. These are local best approximations of the following form: If  $f$  is in  $C[a, b]$  and  $G$  is in  $\mathcal{W}_n$  and if we approximate  $f$  by  $\tilde{G} = G|_{[c, d]}$  ( $\dim \tilde{G} = m$ ) for any subinterval  $[c, d]$  of  $[a, b]$ , then  $g_0$  in  $G$  is said to be a *local AE* for  $f$ , if  $\|f - g_0\|_{[c, d]} \leq \|f - g\|_{[c, d]}$  for all  $g$  in  $G$  and if there are  $m + 1$  points  $c \leq y_0 < y_1 < \dots < y_m \leq d$  such that

$$\varepsilon(-1)^i(f - g_0)(y_i) = \|f - g_0\|_{[c, d]} \quad \text{for } i = 0, \dots, m, \varepsilon = \pm 1.$$

In general  $g_0$  is not an AE for  $f$  from  $G$  and, therefore,  $g_0$  is not in  $P_G(f)$  in general.

The construction of a continuous selection for any  $G$  in  $\check{\mathcal{V}}_n$  is based on the construction of a continuous selection for splines established by Nürnberger and Sommer [10]. But while in that paper the authors have been able to use well-known results from spline theory, we now use some of those results about the elements of  $\check{\mathcal{V}}_n$  which we have shown in § 2.

LEMMA 3.7. *Let  $G$  be in  $\check{\mathcal{V}}_n$ . Then there exists a continuous selection for  $G$ .*

*Proof.* Let  $f \in C[a, b]$  and  $g_0 \in P_G(f)$  arbitrarily. Then by Lemma 2.5 there exists an interval  $[x_p, x_{p+1}]$  such that  $g = g_0$  on  $[x_p, x_{p+1}]$  for all  $g \in P_G(f)$ . We construct a continuous selection step by step:

(i) *Local approximation.* If  $\dim \bar{G}_{0, p+1} \geq 1$ , then we approximate  $f - g_0$  in  $[x_{p+1}, b]$  by  $\bar{G}_{0, p+1}$ . Since  $\bar{G}_{0, p+1}$  is weak Chebyshev by Lemma 2.1, Theorem 1.3 guarantees the existence of a local AE  $g_1 \in P_{\bar{G}_{0, p+1}}(f - g_0)$  for which

$$\begin{aligned} \|f - g_0 - g_1\|_{[a, x_{p+1}]} &= \|f - g_0\|_{[a, x_{p+1}]} \leq \|f - g_0\|, \\ \|f - g_0 - g_1\|_{[x_{p+1}, b]} &\leq \|f - g_0 - 0\|_{[x_{p+1}, b]} \leq \|f - g_0\|. \end{aligned}$$

Therefore  $g_0 + g_1 \in P_G(f)$ . If  $\bar{G}_{0, p+1} = \langle 0 \rangle$ , then  $g_0 = \tilde{g}_0$  on  $[x_{p+1}, b]$  for all  $g_0, \tilde{g}_0 \in P_G(f)$ , because otherwise by Lemma 2.1 there exists a nonzero  $\bar{g} \in \bar{G}_{0, p+1}$ . In this case we define the function  $g_1$  by  $g_1 \equiv 0$ .

(ii) *Uniqueness of local AE's on  $[x_{p+1}, x_{p+2}]$ .* We will now show for approximation in  $[x_{p+1}, b]$  that any two AE's  $g_1, \bar{g}_1 \in P_{\bar{G}_{0, p+1}}(f - g_0)$  are the same on  $[x_{p+1}, x_{p+2}]$ , i.e.  $g_1 = \bar{g}_1$  on  $[x_{p+1}, x_{p+2}]$ . We assume to the contrary that  $g_1 \neq \bar{g}_1$  on  $[x_{p+1}, x_{p+2}]$ . Since  $g_1 = \bar{g}_1$  on  $[a, x_{p+1}]$  and  $G \in \check{\mathcal{V}}_n$ , the function  $g_1 - \bar{g}_1$  has no zero interval in  $[x_{p+1}, b]$ . Then by Lemma 3.6 we get  $|Z^*(g_1 - \bar{g}_1)| \leq m_{0, p+1} + 1$  on  $[x_{p+1}, b]$  and, since  $G \in \check{\mathcal{V}}_n$ ,  $|Z(g_1 - \bar{g}_1)| \leq m_{0, p+1}$  on  $[x_{p+1}, b]$ . But because of Lemma 3.5 this is not possible and, therefore, we get that  $g_1 = \bar{g}_1$  on  $[x_{p+1}, x_{p+2}]$ .

(iii) We show: If  $\tilde{g}_0 \in P_G(f)$ ,  $\tilde{g}_0 \neq g_0$ , and  $\tilde{g}_1 \in P_{\bar{G}_{0, p+1}}(f - \tilde{g}_0)$  is a local AE for approximation in  $[x_{p+1}, b]$ , then  $g_0 + g_1 = \tilde{g}_0 + \tilde{g}_1$  on  $[x_p, x_{p+2}]$ . Since  $g_0 = \tilde{g}_0$  on  $[x_p, x_{p+1}]$ ,

the function  $g_0 - \tilde{g}_0 \in \bar{G}_{p,p+1}$ . Then by Lemma 2.1 the function  $\bar{g}_0$ , defined by

$$\bar{g}_0 = \begin{cases} g_0 - \tilde{g}_0 & \text{on } [x_{p+1}, b], \\ 0 & \text{on } [a, x_{p+1}] \end{cases}$$

is an element of  $\bar{G}_{0,p+1}$ . The functions  $f - g_0 - g_1$  and  $f - \tilde{g}_0 - \tilde{g}_1 = f - g_0 - (-g_0 + \tilde{g}_0 + \tilde{g}_1)$  have  $m_{0,p+1} + 1$  local alternating extreme points on  $[x_{p+1}, b]$ . Since  $\bar{g}_0 \in \bar{G}_{0,p+1}$ , the function  $\bar{g}_1 = \tilde{g}_1 - \bar{g}_0 \in \bar{G}_{0,p+1}$  and this function is a local AE for  $f - g_0$  by approximation in  $[x_{p+1}, b]$ . Since according to (ii) all of these local AE's coincide on  $[x_{p+1}, x_{p+2}]$ , we must have  $g_1 = \tilde{g}_1 - g_0 + \tilde{g}_0$  on  $[x_{p+1}, x_{p+2}]$ . Because  $g_1 = \tilde{g}_1 \equiv 0$  on  $[x_p, x_{p+1}]$  we get finally

$$g_0 + g_1 = \tilde{g}_0 + \tilde{g}_1 \quad \text{on } [x_p, x_{p+2}].$$

(iv) This method will be continued in  $[x_{p+2}, b]$  in the following way: If  $\dim \bar{G}_{0,p+2} \geq 1$ , then we approximate  $f - g_0 - g_1$  in  $[x_{p+2}, b]$  by  $\bar{G}_{0,p+2}$  and by Theorem 1.3 we get a local AE  $g_2 \in P_{\bar{G}_{0,p+2}}(f - g_0 - g_1)$ . As in (ii) we see that all of these AE's coincide on  $[x_{p+2}, x_{p+3}]$  and as in (iii) we see that  $g_0 + g_1 + g_2 = \tilde{g}_0 + \tilde{g}_1 + \tilde{g}_2$  on  $[x_p, x_{p+3}]$  for any choice of  $g_0, \tilde{g}_0, g_0 + g_1, \tilde{g}_0 + \tilde{g}_1$ . We also see that  $g_0 + g_1 + g_2 \in P_G(f)$ . If  $\bar{G}_{0,p+2} = \langle 0 \rangle$ , then we define  $g_2$  by  $g_2 \equiv 0$ .

(v) We continue this method up to the last interval  $[x_{s-1}, b]$  and get a function  $\hat{g} = g_0 + g_1 + \dots + g_{s-1-p} \in P_G(f)$  such that  $\hat{g} = \tilde{g}_0 + \tilde{g}_1 + \dots + \tilde{g}_{s-1-p}$  on  $[x_p, b]$  for any choice of  $g_0, \tilde{g}_0, g_0 + g_1, \tilde{g}_0 + \tilde{g}_1, \dots, g_0 + \dots + g_{s-2-p}, \tilde{g}_0 + \dots + \tilde{g}_{s-2-p}$ .

(vi) Using the same kind of arguments as in (i) to (v) for the interval  $[a, x_p]$  we get a function  $\hat{g} = g_{-p} + g_{-p+1} + \dots + g_{-1} + g_0 \in P_G(f)$  where for each  $i \in \{1, 2, \dots, p\}$ , in case  $\dim \bar{G}_{p+1-i,s} \geq 1$ ,  $g_{-i}$  is a local AE in  $P_{\bar{G}_{p+1-i,s}}(f - g_0 - g_{-1} - \dots - g_{-i+1})$  by approximation in  $[a, x_{p+1-i}]$  and, in case  $\bar{G}_{p+1-i,s} = \langle 0 \rangle$ ,  $g_{-i}$  is defined by  $g_{-i} \equiv 0$ . As before,  $\hat{g} = \tilde{g}_{-p} + \tilde{g}_{-p+1} + \dots + \tilde{g}_{-1} + \tilde{g}_0$  on  $[a, x_{p+1}]$  for any choice of  $g_0, \tilde{g}_0, g_{-1} + g_0, \tilde{g}_{-1} + \tilde{g}_0, \dots, g_{-p+1} + \dots + g_0, \tilde{g}_{-p+1} + \dots + \tilde{g}_0$ .

Now we define:  $s(f) = g_{-p} + g_{-p+1} + \dots + g_{-1} + g_0 + g_1 + \dots + g_{s-1-p}$  which is an element of  $P_G(f)$ .

The continuity of this selection follows exactly in the same way as in the case of the spline functions established in [10]. Therefore, we will omit the proof of the continuity of this selection.

Thus by applying of Lemma 3.3, Lemma 3.4 and Lemma 3.7, Theorem 3.1 is completely proved and so we have given a complete characterization of the existence of continuous selections for  $\mathcal{V}_n$ .

**4. Examples.** In this section we will define some important subclasses of  $\mathcal{V}_n$  and will apply the results of § 3 to those classes.

At first we will show that the spline spaces and the generalized spline spaces in  $\mathcal{V}_n$  have not the same behavior in general, even if we consider generalized spline spaces in  $\check{\mathcal{V}}_n$ . In order to show this we first define the spline spaces: Let  $m, k \in \mathbb{N}$  with  $m + k + 1 = n$ . Let  $a = x_0 < x_1 < \dots < x_{k+1} = b$  be a partition of  $[a, b]$ . Then the space  $S_{m,k}$  of spline functions of degree  $m$  with the  $k$  fixed knots  $x_1, x_2, \dots, x_k$  is spanned by the functions  $1, x, \dots, x^m, (x - x_1)_+^m, (x - x_2)_+^m, \dots, (x - x_k)_+^m$ . In [15] we have shown that  $S_{m,k} \in \check{\mathcal{V}}_n$ . From results of Curry and Schoenberg [3] it follows that  $S_{m,k} \in \check{\mathcal{V}}_n$  if and only if  $k \leq m + 1$ . Therefore, Lemma 2.3, Lemma 2.5 and Theorem 2.6 are valid for all spline spaces  $S_{m,k}$  with  $k \leq m + 1$ . But from results of Karlin [5, p. 503], Rice [11, p. 152] and Schumaker [12] it follows that the statements of Lemma 2.3, Lemma 2.5 and Theorem 2.6 are also valid for  $k > m + 1$  and, therefore, for any spline space  $S_{m,k}$ . But this is not true for all elements of  $\mathcal{V}_n$  as we will show by the following example. Therefore, there are elements of  $\mathcal{V}_n$  having not the same behavior as spline spaces.

*Example.* We define four functions in  $C[-2, 2]$  by  $g_0(x) = x$ ,

$$g_1(x) = (x - 1)_+, \quad g_2(x) = (-1 - x)_+,$$

$$g_3(x) = \begin{cases} 0, & x \in [-2, -1], \\ 1 - x^2, & x \in [-1, 1], \\ 0, & x \in [1, 2], \end{cases}$$

where

$$(x - t)_+ = \begin{cases} x - t & \text{if } x \geq t, \\ 0 & \text{if } x < t. \end{cases}$$

Then  $G = \langle g_0, g_1, g_2, g_3 \rangle \in C[-2, 2]$  is weak Chebyshev with dimension 4 and can be decomposed in Chebyshev spaces by the knots  $x_1 = -1, x_2 = 0, x_3 = 1$ . Hence  $\bar{G}_{01} = \langle g_1, g_3 \rangle, \bar{G}_{02} = \langle g_1 \rangle, \bar{G}_{03} = \langle g_1 \rangle, \bar{G}_{12} = \langle g_1, g_2 \rangle, \bar{G}_{13} = \langle g_1, g_2 \rangle, \bar{G}_{14} = \langle g_2 \rangle, \bar{G}_{23} = \langle g_1, g_2 \rangle, \bar{G}_{24} = \langle g_2 \rangle, \bar{G}_{34} = \langle g_2, g_3 \rangle$ . Therefore,  $G \in \mathcal{V}_n$ . Since  $g_3$  has two separated zero intervals, we get  $G \notin \mathcal{V}_n$ . Since  $n_{01} = 2, n_{02} = 3, n_{03} = 3, n_{12} = 2, n_{13} = 2, n_{14} = 3, n_{23} = 2, n_{24} = 3, n_{34} = 2$ , the points  $y_1 = -2, y_2 = -1, y_3 = 1, y_4 = 2$  satisfy the condition

$$y_{4-n_{i4}} < x_i < y_{n_{0i}+1} \quad \text{for } i = 1, 2, 3.$$

Then Lemma 2.3 is not fulfilled, because  $g_3(y_i) = 0$  for  $i = 1, 2, 3, 4$  and  $g_3 \not\equiv 0$ .

We now define a function  $f$  in  $C[-2, 2]$  by

$$f(x) = \begin{cases} -3 - 2x, & x \in [-2, -1], \\ 1 - 2x^2, & x \in [-1, 1], \\ -3 + 2x, & x \in [1, 2]. \end{cases}$$

Then  $f$  has exactly five alternating extreme points  $-2, -1, 0, 1, 2$  and, therefore, 0 is in  $P_G(f)$ . Then it is easy to verify that also  $g_1 + g_2 + g_3$  is in  $P_G(f)$ . But there is no knot interval  $[x_i, x_j]$  on which  $f - (g_1 + g_2 + g_3)$  has at least  $n_{ij} + 1$  alternating extreme points. Thus Lemma 2.5 is not satisfied. But because of 0,  $g_1 + g_2 + g_3$  in  $P_G(f)$  Theorem 2.6 is also not satisfied, since  $g_1 + g_2 + g_3$  has no zero interval in  $[-2, 2]$ .

*Remark.* As shown before all spline spaces  $S_{m,k}$  satisfying  $k \leq m + 1$  are elements of  $\mathcal{V}_n$ . For this special class of spline spaces there follow from Lemma 2.3, Lemma 2.5 and Theorem 2.6 results of Karlin [5], Rice [11] and Schumaker [12] established for all spline spaces.

Now we will apply Theorem 3.1 to the spaces  $S_{m,k}$ . We have shown before that  $S_{m,k}$  ( $m + k + 1 = n$ ) is an element of  $\mathcal{V}_n$  if and only if  $k \leq m + 1$ . Thus from Theorem 3.1 we get immediately a characterization of those spline spaces which admit a continuous selection.

**THEOREM 4.1.** *Let  $S_{m,k}$  be a spline space. Then there exists a continuous selection for  $S_{m,k}$  in and only if  $k \leq m + 1$ .*

In this way we have obtained a result established by Nürnberger and Sommer [10].

Since the statements of § 2 are also valid for spline spaces having knots with multiplicity less than or equal to  $m$ , Theorem 4.1 is also true in this case.

Finally we will derive from Theorem 3.1 a characterization of the existence of continuous selections for another very important subclass of  $\mathcal{V}_n$ , that is for the continuously composed Chebyshev spaces (CC spaces). For defining this class let  $a = x_0 < x_1 < \dots < x_s = b$  be a partition of  $[a, b]$  and for  $i = 1, \dots, s$  let  $G^i$  be Chebyshev spaces with dimension  $n_i, n_i \geq 1$ , on  $[x_{i-1}, x_i]$ . Bartelt [1] has proved that

each CC space  $G$ , defined by

$$G = \{g \in C[a, b] \mid g|_{[x_{i-1}, x_i]} \in G^i, \quad i = 1, \dots, s\}$$

is weak Chebyshev with dimension  $\sum_{i=1}^s n_i - (s - 1)$ . Therefore, if  $\sum_{i=1}^s n_i - (s - 1) = n$ ,  $G$  is an element of  $\mathcal{W}_n$  and, since each  $x \in [a, b]$  is a nonvanishing zero,  $G$  is an element of  $\mathcal{V}_n$ , too.

It should be observed that the class of the CC spaces is a small subclass of  $\mathcal{V}_n$ , since by definition the elements of those spaces satisfy at the knots no stronger condition than continuity in general. But for many elements of  $\mathcal{V}_n$  stronger conditions are valid at the knots, e.g. any function of the spline space  $S_{m,s-1}$  with the knots  $x_1, x_2, \dots, x_{s-1}$  is  $(m - 1)$ -times continuously differentiable. We will make this clearer by the following example.

*Example.* Let  $a = x_0 < x_1 < \dots < x_s = b$  be a partition of  $[a, b]$ . Then the spline space  $S_{m,s-1}$  with the knots  $x_1, \dots, x_{s-1}$  is spanned by the functions  $1, x, \dots, x^m, (x - x_1)_+^m, (x - x_2)_+^m, \dots, (x - x_{s-1})_+^m$  and, therefore,  $\dim S_{m,s-1} = m + 1 + s - 1 = m + s$ . As said before  $S_{m,s-1} \in \mathcal{V}_n$  if  $n = m + s$ .

From the definition of  $S_{m,s-1}$  it follows immediately that

$$S_{m,s-1} = \{g \in C^{m-1}[a, b] \mid g|_{[x_{i-1}, x_i]} \in \mathbb{P}_m, \quad i = 1, \dots, s\}$$

where  $\mathbb{P}_m$  is the space of all polynomials of degree  $\leq m$ . Now we can see that  $S_{m,s-1}$  is no CC space, because the CC space  $G_0$  belonging to the given knots  $x_1, \dots, x_{s-1}$  and to the Chebyshev spaces  $G^i = \mathbb{P}_m$  for  $i = 1, \dots, s$  is defined by

$$G_0 = \{g \in C[a, b] \mid g|_{[x_{i-1}, x_i]} \in \mathbb{P}_m, \quad i = 1, \dots, s\}.$$

Hence  $\dim G_0 = \sum_{i=1}^s (m + 1) - (s - 1) = sm + 1 > n = m + s = \dim S_{m,s-1}$  for  $m > 1$ . For example, the function  $g_0$ , defined by

$$g_0(x) = \begin{cases} 0, & x \in [a, x_{s-1}], \\ x - x_{s-1}, & x \in [x_{s-1}, b] \end{cases}$$

is for  $m > 1$  an element of  $G_0$  but not of  $S_{m,s-1}$ .

In [15] we have proved that any CC space  $G$  is an element of  $\check{\mathcal{V}}_n$ . Thus it is only to examine in which case  $G$  is in  $\check{\mathcal{V}}_n$ . We get:

LEMMA 4.2. *Let  $G$  in  $\mathcal{W}_n$  be a CC space. Then  $G$  is in  $\check{\mathcal{V}}_n$  if and only if  $n_{ij} \leq 2$  for any  $i, j \in \{1, \dots, s - 1\}, i < j$ .*

*Proof. Necessity:* We assume that  $n_{ij} \geq 3$  for some tuple  $(i, j)$  with  $i, j \in \{1, \dots, s - 1\}$ . Since  $G^{ij}$  is also a CC space, we get  $n_{ij} = \dim G^{ij} = \sum_{p=i+1}^j n_p - (j - i - 1)$ . Since  $n_p \geq 1$  for  $p = i + 1, \dots, j$ , there is, therefore, either some  $p \in \{i + 1, \dots, j\}$  with  $n_p \geq 3$  or there are at least two integers  $r, t \in \{i + 1, \dots, j\}$  such that  $n_r = n_t = 2$  and  $n_p \leq 2$  otherwise.

In the first case we can construct a function  $g_0 \in G$  satisfying  $g_0 \equiv 0$  on  $[a, x_{p-1}] \cup [x_p, b]$  and  $g_0((x_{p-1} + x_p)/2) = 1$ . Since  $x_i > a$  and  $x_j < b$ ,  $g_0$  has two separated zero intervals in  $[a, b]$ . But this is a contradiction of the hypothesis that  $G \in \check{\mathcal{V}}_n$ .

In the second case we can choose two knots  $x_r, x_t$  with  $x_i \leq x_r < x_{r+1} < x_t \leq x_j$  such that  $n_{r+1} = n_t = 2$  and  $n_p = 1$  for all  $p \in \{r + 2, \dots, t - 1\}$ . Then  $n_{rt} = \dim G^{rt} = \sum_{p=r+1}^t n_p - (t - r - 1) = 3$  and  $n_{r+1,t} = n_{r,t-1} = 2$ . Applying Lemma 2.3 to the weak Chebyshev space  $G^{rt}$  we can construct a function  $g_0 \in G$  satisfying  $g_0 \equiv 0$  on  $[a, x_r] \cup [x_t, b]$  and  $g_0 \neq 0$  on  $[x_r, x_t]$ . But this is a contradiction again.

*Sufficiency:* We assume that  $G$  is not in  $\check{\mathcal{V}}_n$ . Therefore, there exists a function  $g_0 \in G$  having two separated zero intervals  $[x_h, x_i], [x_j, x_k], x_i < x_j$ , such that  $g_0 \neq 0$  on

$[x_i, x_{i+1}]$  and on  $[x_{j-1}, x_j]$ . If  $x_{i+1} = x_j$ , then  $g_0 \neq 0$  on  $[x_i, x_{i+1}]$ . Since  $g_0$  has at least the two zeros  $x_i$  and  $x_{i+1}$  on  $[x_i, x_{i+1}]$ , it follows from Definition 1.1 that  $n_{i+1} \geq 3$ . But this is a contradiction of the hypothesis that  $n_{ij} \leq 2$  for any  $i, j \in \{1, \dots, s-1\}$ .

If  $x_{i+1} < x_j$ , then because of  $g_0(x_i) = 0$ ,  $g_0(x_j) = 0$  and  $g_0 \neq 0$  on  $[x_i, x_{i+1}]$  and on  $[x_{j-1}, x_j]$  we get  $n_{i+1} \geq 2$  and  $n_j \geq 2$ . Here we have to consider again that  $G^{i+1}$  and  $G^j$  are Chebyshev. Then  $n_{ij} = \dim G^{ij} = \sum_{p=i+1}^j n_p - (j-i-1) \geq 2 + \sum_{p=i+2}^{j-1} n_p + 2 - (j-i-1) \geq 2 + j - i - 2 + 2 - (j-i-1) = 3$ . But this is a contradiction again.

Thus it follows from Theorem 3.1:

**THEOREM 4.3.** *Let  $G$  in  $\mathcal{W}_n$  be a CC space. Then there exists a continuous selection for  $G$  if and only if  $n_{ij} \leq 2$  for any  $i, j \in \{1, \dots, s-1\}$ ,  $i < j$ .*

*Remark.* The dimensions  $n_{01} = n_1$  and  $n_{s-1,s} = n_s$  of the Chebyshev space  $G^1$  or  $G^s$ , respectively, are not considered in the above characterization. Therefore, the existence or nonexistence of a continuous selection depends only on the "inner" dimensions  $n_p$ ,  $p = 2, \dots, s-1$ .

Thus, if we choose only one knot  $a = x_0 < x_1 < x_2 = b$ , then we always get a continuous selection for any choice of the dimensions  $n_1$  and  $n_2$ .

**Acknowledgment.** I thank the referees for many helpful comments about the rewriting of this paper.

#### REFERENCES

- [1] M. W. BARTELT, *Weak Chebyshev sets and splines*, J. Approximation Theory, 14 (1975), pp. 30–37.
- [2] A. L. BROWN, *On continuous selections for the metric projection in spaces of continuous functions*, J. Functional Analysis, 8 (1971), pp. 431–449.
- [3] H. B. CURRY AND I. J. SCHOENBERG, *On Polya frequency functions IV: the fundamental spline functions and their limits*, J. Analyse Math., 17 (1966), pp. 71–107.
- [4] R. C. JONES AND L. A. KARLOVITZ, *Equioscillation under nonuniqueness in the approximation of continuous functions*, J. Approx. Theory, 3 (1970), pp. 138–145.
- [5] S. KARLIN, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.
- [6] S. KARLIN AND W. J. STUDDEN, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York, 1966.
- [7] A. J. LAZAR, P. D. MORRIS AND D. E. WULBERT, *Continuous selections for metric projections*, J. Functional Analysis, 3 (1969), pp. 193–216.
- [8] G. NÜRNBERGER, *Nonexistence of continuous selections for the metric projection*, preprint.
- [9] G. NÜRNBERGER AND M. SOMMER, *Weak Chebyshev subspaces and continuous selections for the metric projection*, Trans. Amer. Math. Soc., 238 (1978), pp. 129–138.
- [10] ———, *Characterization of continuous selections of the metric projection for spline functions*, J. Approximation Theory, 22 (1978), pp. 320–330.
- [11] R. J. RICE, *The Approximation of Functions*, Vol. II, Addison-Wesley, Reading, MA, 1969.
- [12] L. L. SCHUMAKER, *Uniform approximation by Tchebycheffian spline functions*, J. Math. Mech., 18 (1968), pp. 369–378.
- [13] M. SOMMER, *Continuous selections of the metric projection for 1-Chebyshev spaces*, J. Approximation Theory, to appear.
- [14] ———, *Nonexistence of continuous selections of the metric projection for a class of weak Chebyshev spaces*, Trans. Amer. Math. Soc., to appear.
- [15] ———, *Weak Chebyshev spaces and best  $L_1$ -approximation*, preprint.
- [16] M. SOMMER AND H. STRAUSS, *Eigenschaften von schwach tchebyscheffischen Räumen*, J. Approximation Theory, 21 (1977), pp. 257–268.
- [17] B. STOCKENBERG, *On the number of zeros of functions in a weak Tchebyshev-space*, Math. Z., 156 (1977), pp. 49–57.



## ON THE BOUNDARY VALUE PROBLEM FOR SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS WITH A SINGULARITY OF THE SECOND KIND\*

FRANK R. DE HOOG† AND RICHARD WEISS‡

**Abstract.** A Fredholm theory for linear boundary value problems  $t^\alpha y' = G(t)y + g(t)$ ,  $0 < t \leq 1$ ,  $y \in C[0, 1] \cap C^1(0, 1]$ ,  $\alpha \geq 1$ ;  $B_0 y(0) + B_1 y(1) = \gamma$  is established, together with existence and regularity results for continuous solutions of nonlinear systems of ordinary differential equations  $t^\alpha y' = f(t, y)$ . This theory is applied to boundary value problems on infinite intervals and is illustrated by two examples. Finally, some fundamental properties of the generalized linear eigenvalue problem  $t^\alpha y' - (G(t) - \lambda H(t))y = 0$ ,  $0 < t \leq 1$ ,  $y \in C[0, 1] \cap C^1(0, 1]$ ;  $B_0 y(0) + B_1 y(1) = 0$ , are derived.

**1. Introduction.** Boundary value problems for singular systems of ordinary differential equations

$$(1.1) \quad \begin{aligned} t^\alpha y' &= f(t, y), & 0 < t \leq 1, & \quad y \in C[0, 1] \cap C^1(0, 1], \\ b(y(0), y(1)) &= 0 \end{aligned}$$

where  $\alpha \geq 1$ ,  $y$  is an  $n$  vector and  $f, b$  are continuous nonlinear mappings on appropriate domains, and linear eigenvalue problems

$$(1.2) \quad \begin{aligned} t^\alpha y' - (A(t) + \lambda C(t))y &= 0, & 0 < t \leq 1, & \quad y \in C[0, 1] \cap C(0, 1], \\ B_0 y(0) + B_1 y(1) &= 0 \end{aligned}$$

where  $A, C \in C[0, 1]$  and  $B_0, B_1$  are matrices, often occur in applied mathematics. When  $\alpha = 1$ , (1.1), (1.2) are said to have a singularity of the first kind, while the singularity is of the second kind when  $\alpha > 1$ .

The case  $\alpha = 1$  is obtained, for instance, when partial differential equations are reduced to ordinary differential equations in the presence of symmetry. A variety of examples can be found in Rentrop [10], [11]. Certain analytic aspects of problems with a singularity of the first kind and their numerical solution by difference schemes have recently been studied in de Hoog and Weiss [3], [4], [5].

A singularity of the second kind arises when a differential equation on an infinite interval is transformed to one on a finite interval. There is a large variety of sources of differential equations on infinite intervals, ranging from exterior problems for elliptic equations in separating coordinates to similarity solutions of the equations of boundary layer theory, see Schlichting [13]. Eigenvalue problems (1.2) with a singularity of the second kind are common in quantum physics.

The present paper provides a study of basic analytic properties of (1.1), (1.2) for  $\alpha > 1$ . We establish a Fredholm theory for the case when (1.1) is linear and provide existence and smoothness results for nonlinear problems. This theory is applied to two examples. Finally we investigate some fundamental properties of invariant subspaces associated with isolated eigenvalues of (1.2).

The results developed here also provide the analytic background for the derivation and analysis of approximate methods for the case  $\alpha > 1$ , which is examined in de Hoog and Weiss [6].

**2. The scalar case.** Here, we examine the equation

$$(2.1) \quad t^\alpha y' - \lambda y = t^{\alpha-\rho} g(t), \quad 0 < t \leq 1, \quad \text{Re } \lambda \neq 0,$$

\* Received by the editors April 12, 1977, and in revised form October 11, 1978.

† Computer Centre, The Australian National University, Canberra, Australia.

‡ Institut für Numerische Mathematik, Technische Universität, Vienna, Austria.

where  $\alpha, \rho$  are real,  $\alpha > 1$  and  $g \in C[0, 1]$ . The study of such scalar equations is the first step in the analysis of vector systems. The general solution of (2.1) is

$$(2.2) \quad y(t) = Y(t)y(\delta) + Y(t) \int_{\delta}^t Y^{-1}(s)s^{-\rho}g(s) ds$$

where

$$Y(t) = \exp [\lambda(\delta^{1-\alpha} - t^{1-\alpha})/(\alpha - 1)]$$

and  $0 < \delta \leq 1$ . For the analysis of continuous solutions of (2.1) it is convenient to examine the operator  $\mathcal{B}_{\rho}$  defined by

$$(\mathcal{B}_{\rho}g)(t) = \begin{cases} t^{\rho-\alpha}Y(t) \int_0^t Y^{-1}(s)s^{-\rho}g(s) ds; & 0 < t \leq 1, \operatorname{Re} \lambda < 0, \\ t^{\rho-\alpha}Y(t) \int_{\delta}^t Y^{-1}(s)s^{-\rho}g(s) ds; & 0 < t \leq 1, \operatorname{Re} \lambda > 0, \\ -g(0)/\lambda; & t = 0. \end{cases}$$

We shall now establish various properties of  $\mathcal{B}_{\rho}$ . As the proofs of these results are very similar for  $\sigma = \operatorname{Re} \lambda < 0$  and  $\sigma > 0$ , they will be given only for the case  $\sigma < 0$ . However, all arguments carry over to  $\sigma > 0$  without difficulty.

The first result we require is

LEMMA 2.1. *Let  $g \in C^1[0, 1]$ . Then*

(i) *if  $\operatorname{Re} \lambda < 0$ ,*

$$(\mathcal{B}_{\rho}g)(t) = \{(\alpha - \rho)t^{\alpha-1}(\mathcal{B}_{\rho+1-\alpha}g)(t) + t^{\alpha}(\mathcal{B}_{\rho-\alpha}g')(t) - g(t)\}/\lambda$$

and

(ii) *if  $\operatorname{Re} \lambda > 0$ ,*

$$(\mathcal{B}_{\rho}g)(t) = \{(\alpha - \rho)t^{\alpha-1}(\mathcal{B}_{\rho+1-\alpha}g)(t) + t^{\alpha}(\mathcal{B}_{\rho-\alpha}g')(t) - [g(t) - (t/\delta)^{\rho-\alpha}Y(t)g(\delta)]\}/\lambda.$$

*Proof.* If  $\operatorname{Re} \lambda < 0$ ,

$$(\mathcal{B}_{\rho}g)(t) = t^{\rho-\alpha}Y(t) \int_0^t [s^{-\alpha}Y^{-1}(s)]s^{\alpha-\rho}g(s) ds.$$

Note that  $Y^{-1}$  satisfies the adjoint equation

$$\frac{d}{dt}Y^{-1}(t) = -\lambda t^{-\alpha}Y^{-1}(t)$$

and that

$$\lim_{t \rightarrow 0_+} t^{\alpha-\rho}Y^{-1}(t) = 0.$$

Integration by parts therefore yields the result for  $\operatorname{Re} \lambda < 0$ . A similar argument establishes the result when  $\operatorname{Re} \lambda > 0$ .  $\square$

LEMMA 2.2. *There exists a constant  $C$  independent of  $\delta$  such that*

$$\|\mathcal{B}_{\rho}g\|_{\delta} \leq C\|g\|_{\delta}$$

where  $\|\cdot\|_{\delta} = \sup_{t \in (0, \delta)} |\cdot|$ .

*Proof.* If  $\sigma = \operatorname{Re} \lambda < 0$ ,

$$|(\mathcal{B}_{\rho}g)(t)| \leq t^{\rho-\alpha} e^{-\sigma t^{1-\alpha}/(\alpha-1)} \int_0^t e^{\sigma s^{1-\alpha}/(\alpha-1)} s^{-\rho} ds \|g\|_{\delta}, \quad 0 \leq t \leq \delta.$$

Let

$$\delta_1 = \begin{cases} 1, & \alpha - \rho \geq 0, \\ \min \{1, [\sigma/2(\alpha - \rho)]^{1/(\alpha-1)}\}, & \alpha - \rho < 0. \end{cases}$$

Then, it is not difficult to verify that

$$s^{\alpha-\rho} e^{\sigma s^{1-\alpha}/(2(\alpha-1))} \leq \begin{cases} t^{\alpha-\rho} e^{\sigma t^{1-\alpha}/(2(\alpha-1))}, & 0 \leq s \leq t \leq \delta_1 \\ \delta_1^{\alpha-\rho} e^{\sigma \delta_1^{1-\alpha}/(2(\alpha-1))}, & s > \delta_1. \end{cases}$$

Combining these estimates, we obtain that

$$s^{\alpha-\rho} e^{\sigma s^{1-\alpha}/(2(\alpha-1))} \leq C_1 t^{\alpha-\rho} e^{\sigma t^{1-\alpha}/(2(\alpha-1))}, \quad 0 \leq s \leq t \leq 1,$$

where

$$C_1 = \max \{1, \delta_1^{\alpha-\rho}\}$$

and hence

$$\begin{aligned} |(\mathcal{B}_\rho g)(t)| &\leq C_1 e^{-\sigma t^{1-\alpha}/(2(\alpha-1))} \int_0^t e^{-\sigma s^{1-\alpha}/(2(\alpha-1))} s^{-\alpha} ds \|g\|_\delta \\ &= -2C_1 \|g\|_\delta / \sigma, \quad 0 \leq t \leq \delta. \end{aligned}$$

This establishes the result for  $\operatorname{Re} \lambda < 0$ . A similar argument can be used when  $\operatorname{Re} \lambda > 0$ .  $\square$

LEMMA 2.3. *If  $g \in C[0, 1]$  then  $\mathcal{B}_\rho g \in C[0, 1] \cap C^1(0, 1)$ .*

*Proof.* Clearly,  $\mathcal{B}_\rho g \in C(0, 1] \cap C^1(0, 1]$ , and it only remains to show that

$$\lim_{t \rightarrow 0^+} (\mathcal{B}_\rho g)(t) = -g(0)/\lambda \equiv (\mathcal{B}_\rho g)(0).$$

For  $\operatorname{Re} \lambda < 0$ , Lemma 2.1 yields

$$-g(0)/\lambda = (\mathcal{B}_\rho g(0))(t) + (\rho - \alpha)t^{\alpha-1}(\mathcal{B}_{\rho+1-\alpha} 1)(t)g(0)/\lambda.$$

Hence, from Lemma 2.2,

$$\begin{aligned} |(\mathcal{B}_\rho g)(t) - (\mathcal{B}_\rho g)(0)| &= |(\mathcal{B}_\rho [g - g(0)])(t) - (\rho - \alpha)g(0)t^{\alpha-1}(\mathcal{B}_{\rho+1-\alpha} 1)(t)/\lambda| \\ &\leq \text{const.} \left\{ \sup_{s \in (0, t)} |g(s) - g(0)| + t^{\alpha-1}|g(0)| \right\}. \end{aligned}$$

Since  $g$  is continuous and  $\alpha - 1 > 0$  it follows that the term on the right hand side can be made arbitrarily small. This establishes the result for  $\operatorname{Re} \lambda < 0$  and a similar argument can be used if  $\operatorname{Re} \lambda > 0$ .  $\square$

Further smoothness can be established when  $\alpha$  is an integer. In particular, we have

LEMMA 2.4. *Let  $\alpha$  be an integer greater than one and  $g \in C^m[0, 1]$ . Then,  $\mathcal{B}_\rho g \in C^m[0, 1] \cap C^{m+1}(0, 1]$  and*

$$|(\mathcal{B}_\rho g)^{(m)}(0)| \leq \text{const.} \sum_{k=0}^m |g^{(k)}(0)|.$$

*Proof.* Differentiation yields

$$(2.3) \quad (\mathcal{B}_\rho g)'(t) = (\rho - \alpha)(\mathcal{B}_\rho g)(t)/t + \lambda(\mathcal{B}_\rho g)(t)/t^\alpha + g(t)/t^\alpha.$$

If  $\operatorname{Re} \lambda < 0$ , two applications of Lemma 2.1 to (2.3) give

$$(2.4) \quad \begin{aligned} (\mathcal{B}_\rho g)'(t) &= (\rho - \alpha)[(\mathcal{B}_\rho g)(t) - (\mathcal{B}_{\rho-\alpha+1}g)(t)]/t + (\mathcal{B}_{\rho-\alpha}g')(t) \\ &= (\rho - \alpha)[(\alpha - \rho)t^{\alpha-2}(\mathcal{B}_{\rho-\alpha+1}g)(t) \\ &\quad + t^{\alpha-1}(\mathcal{B}_{\rho-\alpha}g')(t) - (2\alpha - \rho - 1)t^{\alpha-2}(\mathcal{B}_{\rho+2-2\alpha}g)(t) \\ &\quad - t^{\alpha-1}(\mathcal{B}_{\rho+1-2\alpha}g')(t)]/\lambda + (\mathcal{B}_{\rho-\alpha}g')(t). \end{aligned}$$

The result for  $m = 1$  now follows from Lemma 2.3. A simple inductive argument based on (2.4) completes the proof for  $\operatorname{Re} \lambda < 0$ .

For  $\operatorname{Re} \lambda > 0$ , two applications of Lemma 2.1 to (2.3) yield

$$(2.5) \quad \begin{aligned} (\mathcal{B}_\rho g)'(t) &= (\rho - \alpha)[(\alpha - \rho)t^{\alpha-2}(\mathcal{B}_{\rho-\alpha+1}g)(t) + t^{\alpha-1}(\mathcal{B}_{\rho-\alpha}g')(t) \\ &\quad - (2\alpha - \rho - 1)t^{\alpha-2}(\mathcal{B}_{\rho+2-2\alpha}g)(t) \\ &\quad - t^{\alpha-1}(\mathcal{B}_{\rho+1-2\alpha}g')(t)]/\lambda + (\mathcal{B}_{\rho-\alpha}g')(t) \\ &\quad + g(\delta)Y(t)\left(\frac{t}{\delta}\right)^{\rho-\alpha} \left[ (\rho - \alpha)t^{-1}\lambda^{-1} + t^{-\alpha} + (\alpha - \rho)\left(\frac{t}{\delta}\right)^{1-\alpha} t^{-1}\lambda^{-1} \right] \end{aligned}$$

and the result follows as previously.  $\square$

We now return to (2.1) and examine continuous solutions.

LEMMA 2.5. *If  $\operatorname{Re} \lambda < 0$ ,  $\alpha > 1$  and  $\rho \leq \alpha$ , then for every  $g \in C[0, 1]$  there is a unique  $y \in C[0, 1]$  which satisfies (2.1).*

*Proof.* From (2.2) every continuous solution of (2.1) satisfies

$$\begin{aligned} y(t) &= Y(t) \int_0^t Y^{-1}(s)s^{-\rho}g(s) ds + Y(t)[y(\delta) - \int_0^\delta Y^{-1}(s)s^{-\rho}g(s) ds] \\ &= t^{\alpha-\rho}(\mathcal{B}_\rho g)(t) + Y(t) \left[ y(\delta) - \int_0^\delta Y^{-1}(s)s^{-\rho}g(s) ds \right]. \end{aligned}$$

Clearly,

$$\lim_{t \rightarrow 0^+} Y(t) = \infty$$

and from Lemma 2.3,  $\mathcal{B}_\rho g \in C[0, 1]$ . It follows that  $y \in C[0, 1]$  iff

$$y(\delta) - \int_0^\delta Y^{-1}(s)s^{-\rho}g(s) ds = 0,$$

and hence the unique  $y \in C[0, 1]$  is

$$y(t) = t^{\alpha-\rho}(\mathcal{B}_\rho g)(t). \quad \square$$

LEMMA 2.6. *If  $\operatorname{Re} \lambda > 0$ ,  $\alpha > 1$  and  $\rho \leq \alpha$ , then for every  $g \in C[0, 1]$  and scalar  $\eta$ , there is a unique  $y \in C[0, 1]$  satisfying (2.1) and  $y(\delta) = \eta$ .*

*Proof.* As in Lemma 2.5 the solution has the form

$$y(t) = Y(t)\eta + t^{\alpha-\rho}(\mathcal{B}_\rho g)(t).$$

Since  $Y \in C^\infty[0, 1]$  the result follows from Lemma 2.3.  $\square$

Combining the results of Lemmas 2.5 and 2.6 we obtain

THEOREM 2.1. *Let  $\alpha > 1$ ,  $\rho \leq \alpha$  and  $g \in C[0, 1]$ . Then every continuous solution of (2.1) has the form*

$$y(t) = PY(t)y(\delta) + t^{\alpha-\rho}(\mathcal{B}_\rho g)(t),$$

where  $P = 0$  if  $\operatorname{Re} \lambda < 0$  and  $P = 1$  when  $\operatorname{Re} \lambda > 0$ .

We now consider (2.1) when  $\alpha = 1$ . Define

$$(2.6) \quad (\mathcal{B}_\rho g)(t) = \begin{cases} \int_0^1 \bar{s}^{(\lambda+\rho)} g(ts) ds, & 0 \leq t \leq 1, \operatorname{Re} \lambda < 0; \rho \leq 1, \\ t^\lambda \int_\delta^t \bar{s}^{(\lambda+1)} g(s) ds, & 0 < t \leq 1, \operatorname{Re} \lambda > 0; \rho = 1, \\ -\frac{g(0)}{(\lambda + \rho - 1)}, & t = 0. \end{cases}$$

**THEOREM 2.2.** Let  $\alpha = 1$ ,  $\rho$  be as in (2.6) and  $g \in C[0, 1]$ . Then

(i)  $\|\mathcal{B}_\rho g\|_\delta \leq \text{const.} \|g\|_\delta$ ,

where the constant is independent of  $\delta$ ,

(ii)  $\mathcal{B}_\rho g \in C[0, 1] \cap C^1(0, 1)$ ,

(iii) for  $\operatorname{Re} \lambda < 0$  and  $g \in C^1[0, 1]$  we have  $\mathcal{B}_\rho g \in C^1[0, 1]$  and

$$(\mathcal{B}_\rho g)' = \mathcal{B}_{\rho-1} g'$$

(iv) every solution of (2.1) which is in  $C[0, 1] \cap C^1(0, 1)$  has the form

$$y(t) = P\left(\frac{t}{\delta}\right)^\lambda y(\delta) + t^{1-\rho} (\mathcal{B}_\rho g)(t)$$

where  $P = 0$  if  $\operatorname{Re} \lambda < 0$  and  $P = 1$  if  $\operatorname{Re} \lambda > 0$ .

*Proof.* (i) This is clear for  $\operatorname{Re} \lambda < 0$ , while for  $\operatorname{Re} \lambda > 0$  it follows immediately from Lemma 3.4 in de Hoog and Weiss [3].

(ii) Again, the result is obvious for  $\operatorname{Re} \lambda < 0$ ; for  $\operatorname{Re} \lambda > 0$  see Lemma 3.4, de Hoog and Weiss [3].

(iii) The proof is obvious.

(iv) See pp. 778–779 in de Hoog and Weiss [3].  $\square$

*Remark.* When  $\alpha > 1$ ,  $\rho \leq \alpha$  and  $\alpha, \rho$  are integers, it follows from Lemma 2.4 that the solution  $y$  of Theorem 2.1 is in  $C^m[0, 1]$  provided that  $g \in C^m[0, 1]$ . For  $\alpha = 1$  this is true when  $\operatorname{Re} \lambda < 0$ , but does not hold for  $\operatorname{Re} \lambda > 0$ . In this case the smoothness properties of  $y$  depend on the size of  $\operatorname{Re} \lambda$  as well as  $g$ ; as can be seen from Theorem 2.2, (iv).

### 3. Linear systems. Initially we examine

$$(3.1) \quad t^\alpha y' - My = t^{\alpha-\rho} g(t), \quad 0 < t \leq 1, \quad \alpha \geq 1,$$

where  $M$  is an  $n \times n$  matrix whose eigenvalues  $\lambda_j$  satisfy  $\operatorname{Re} \lambda_j \neq 0$ ,  $j = 1, \dots, n$ . The general solution is

$$(3.2) \quad y(t) = Y(t)y(\delta) + Y(t) \int_\delta^t Y^{-1}(s) s^{-\rho} g(s) ds$$

where

$$Y(t) = \begin{cases} \exp [M(\delta^{\alpha-1} - t^{\alpha-1})/(\alpha - 1)], & \alpha \neq 1, \\ [t/\delta]^M = \exp [\log (t/\delta)M], & \alpha = 1 \end{cases}$$

is the fundamental solution satisfying

$$t^\alpha Y' - MY = 0, \quad 0 < t \leq 1, \quad Y(\delta) = I,$$

and  $0 < \delta \leq 1$ . Let

$$(3.3) \quad \begin{aligned} Q &= \frac{1}{2\pi i} \int_{\Gamma_-} (\lambda I - M)^{-1} d\lambda \\ P &= \frac{1}{2\pi i} \int_{\Gamma_+} (\lambda I - M)^{-1} d\lambda \end{aligned}$$

where  $\Gamma_-$  and  $\Gamma_+$  are closed contours in the left- and right-hand side of the complex plane respectively such that each eigenvalue of  $M$  is enclosed by either  $\Gamma_-$  or  $\Gamma_+$ . Clearly,  $Q$  and  $P$  are projections onto the invariant subspaces of  $M$  associated with the eigenvalues having negative and positive real part respectively.

Via the Jordan decomposition of  $M$  it is straightforward to obtain an explicit representation of  $Y(t)$ . This representation immediately yields

LEMMA 3.1. *Let  $\alpha \geq 1$ . Then for an arbitrary vector  $\eta$ ,  $Y\eta \in C[0, 1]$  iff  $Q\eta = 0$ .*

For  $\alpha > 1$  we define

$$(\mathcal{B}_\rho g)(t) = \begin{cases} t^{\rho-\alpha} Y(t) \int_0^t QY^{-1}(s)s^{-\rho}g(s) ds \\ \quad + t^{\rho-\alpha} Y(t) \int_\delta^t PY^{-1}(s)s^{-\rho}g(s) ds, & 0 < t \leq 1, \\ -M^{-1}g(0), & t = 0. \end{cases}$$

The above operator is the analogue of  $\mathcal{B}_\rho$  defined in § 2. This becomes apparent on noting that (for  $\alpha > 1$ )

$$\begin{aligned} Y(t)QY^{-1}(s) &= \frac{1}{2\pi i} \int_{\Gamma_-} e^{\lambda(s^{1-\alpha}-t^{1-\alpha})/(\alpha-1)} (\lambda I - M)^{-1} d\lambda, \\ Y(t)PY^{-1}(s) &= \frac{1}{2\pi i} \int_{\Gamma_+} e^{\lambda(s^{1-\alpha}-t^{1-\alpha})/(\alpha-1)} (\lambda I - M)^{-1} d\lambda, \end{aligned}$$

which yields

$$(3.4) \quad \begin{aligned} (\mathcal{B}_\rho g)(t) &= \frac{1}{2\pi i} \int_{\Gamma_-} t^{\rho-\alpha} e^{-\lambda t^{1-\alpha}/(\alpha-1)} \int_0^t e^{\lambda s^{1-\alpha}/(\alpha-1)} s^{-\rho} (\lambda I - M)^{-1} g(s) ds d\lambda \\ &\quad + \frac{1}{2\pi i} \int_{\Gamma_+} t^{\rho-\alpha} e^{-\lambda t^{1-\alpha}/(\alpha-1)} \int_\delta^t e^{\lambda s^{1-\alpha}/(\alpha-1)} s^{-\rho} (\lambda I - M)^{-1} g(s) ds d\lambda. \end{aligned}$$

Since  $\Gamma_-$  and  $\Gamma_+$  are contours in the left- and right-hand sides of the complex plane respectively, the results of § 2 immediately yield

LEMMA 3.2. *Let  $\alpha > 1$  and  $g \in C[0, 1]$ . Then*

- (i)  $\|\mathcal{B}_\rho g\|_\delta \leq \text{const.} \|g\|_\delta$

where the constant is independent of  $\delta$

- (ii)  $\mathcal{B}_\rho g \in C[0, 1] \cap C^1(0, 1)$ .

Regarding continuous solutions of (3.1) we find

LEMMA 3.3. *Let  $\alpha > 1$ ,  $\rho \leq \alpha$  and  $g \in C[0, 1]$ . Then every  $y \in C[0, 1] \cap C^1(0, 1)$  which satisfies (3.1), has the form*

$$(3.5) \quad y(t) = Y(t)Py(\delta) + t^{\alpha-\rho}(\mathcal{B}_\rho g)(t).$$

*Proof.* Equation (3.2) may be rewritten as

$$y(t) = Y(t)Py(\delta) + t^{\alpha-\rho}(\mathcal{B}_\rho g)(t) + Y(t)Q[y(\delta) - \int_0^\delta QY^{-1}(s)s^{-\rho}g(s) ds].$$

From (3.3),  $QPy(\delta) = 0$  and, from Lemma 3.2,  $\mathcal{B}_\rho g \in C[0, 1]$ . Lemma 3.1 now yields the result.  $\square$

For the case  $\alpha = 1$ , Theorem 2.2 yields

LEMMA 3.4. Let  $\alpha = 1$ ,  $g \in C[0, 1]$ ,  $\rho \leq 1$  when  $P = 0$ ,  $\rho = 1$  when  $P \neq 0$ , and

$$(\mathcal{B}_\rho g)(t) = \begin{cases} \int_0^1 Qs^{-(M+\rho I)} g(ts) ds \\ + t^M \int_\delta^t Ps^{-(M+I)} g(s) ds, & 0 < t \leq 1, \\ -(M + (\rho - 1)I)^{-1} g(0), & t = 0. \end{cases}$$

Then the results of Lemmas 3.2 and 3.3 are valid.

Finally we consider the case when in (3.1)  $\alpha = 0$ ,  $\rho = 0$  and  $M = 0$ . On defining

$$(\mathcal{B}_0 g)(t) = \int_\delta^t g(s) ds$$

we immediately obtain

LEMMA 3.5. Let  $\alpha = \rho = 0$ ,  $M = 0$  and  $g \in C[0, 1]$ , Then

(i)  $\|\mathcal{B}_0 g\|_\delta \leq \delta \|g\|_\delta$ ;

(ii) every  $y \in C[0, 1] \cap C^1(0, 1]$  satisfying (3.1) has the form

$$y(t) = y(\delta) + (\mathcal{B}_0 g)(t).$$

We now examine the system

$$(3.6) \quad T(t)y' - My(t) = g(t), \quad 0 < t \leq 1,$$

where  $y$ ,  $g$  are  $n$ -vectors,  $g \in C[0, 1]$ , and

$$(i) \quad M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1r} \\ 0 & M_{22} & \cdots & M_{2r} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & M_{rr} \end{bmatrix};$$

(ii)  $T(t) = \text{diag}(t^{\alpha_1} I_1, t^{\alpha_2} I_2, \dots, t^{\alpha_r} I_r)$ , where the  $I_k$  are unit matrices;

(iii) either  $\alpha_k \geq 1$ ,  $k = 1, \dots, r$  or  $\alpha_k \geq 1$ ,  $k = 1, \dots, r-1$  and  $\alpha_r = 0$ ;

(iv) each  $M_{kk}$  is a square matrix of the same size as  $I_k$  which is nonsingular when  $\alpha_k \neq 0$  and has no eigenvalues that are purely imaginary. When  $\alpha_r = 0$ ,  $M_{rr} = 0$ .

In the sequel we shall say that

Condition 3.1 holds if all  $\alpha_k$  are integers, and  $M_{kk}$  has no eigenvalues with positive real part whenever  $\alpha_k = 1$ .

Let

$$M = D + U, \quad D = \text{diag}(M_{11}, M_{22}, \dots, M_{rr}),$$

$$Y(t) = \text{diag}(Y_1(t), \dots, Y_r(t)),$$

$$P = \text{diag}(P_1, \dots, P_r),$$

$$Q = \text{diag}(Q_1, \dots, Q_r)$$

and

$$R = I - P - Q = \begin{cases} \text{diag}(0, \dots, 0, I_r), & \alpha_r = 0, \\ 0, & \alpha_r \neq 0, \end{cases}$$

where

$$Y_k(t) = \begin{cases} \exp[(\delta^{1-\alpha} - t^{1-\alpha})M_{kk}/(\alpha_k - 1)], & \alpha_k \neq 1, \\ \exp[\log(t/\delta)M_{kk}] & , \quad \alpha_k = 1, \end{cases}$$

and  $P_k, Q_k$  are defined by (3.3) with  $M$  replaced by  $M_{kk}$  if  $\alpha_k \neq 0$  and  $P_k = Q_k = 0$  if  $\alpha_k = 0$ . In addition, define

$$\begin{aligned} (\mathcal{B}g)(t) &= Y(t) \int_0^t QY^{-1}(s)T^{-1}(s)g(s) ds \\ &\quad + Y(t) \int_\delta^t PY^{-1}(s)T^{-1}(s)g(s) ds + R \int_\delta^t g(s) ds. \end{aligned}$$

From Lemmas 3.2, 3.4 and 3.5 we obtain

LEMMA 3.6. *Let  $g \in C[0, 1]$ . Then*

$$(i) \quad \|\mathcal{B}g\|_\delta \leq \text{const.} \|g\|_\delta, \quad \|\mathcal{B}Rg\|_\delta \leq \delta \|Rg\|_\delta$$

where the constant is independent of  $\delta$ ,

$$(ii) \quad \mathcal{B}g \in C[0, 1] \cap C^1(0, 1];$$

$$(iii) \quad ((I - R)\mathcal{B}g)(0) = (\mathcal{B}(I - R)g)(0) = -(D + R)^{-1}(I - R)g(0).$$

With the aid of Lemmas 3.3, 3.4 and 3.5, it is easy to verify that any continuous solution of (3.6) must satisfy

$$\begin{aligned} y(t) &= Y(t)[Py(\delta) + Ry(\delta)] + (\mathcal{B}[Uy + g])(t) \\ &= Y(t)(P + R)\eta + (\mathcal{B}[Uy + g])(t) \end{aligned}$$

where  $\eta = (P + R)y(\delta)$ . Now consider the iteration

$$y_{\nu+1}(t) = Y(t)(P + R)\eta + (\mathcal{B}[Uy_\nu + g])(t), \quad \nu = 0, 1, \dots; \quad y_0 = 0.$$

Since  $P, Q, R, Y(t)$  and  $T(t)$  are block diagonal and  $U$  is strictly upper triangular it follows that

$$(\mathcal{B}U)^k \equiv 0, \quad k \geq n.$$

Hence

$$(3.7) \quad y(t) = y_k(t) = \Phi(t)\eta + (\mathcal{H}g)(t), \quad k \geq n,$$

where

$$\Phi(t) = \sum_{k=0}^n [(\mathcal{B}U)^k Y(P + R)](t)$$

and

$$(3.8) \quad \mathcal{H} = \sum_{k=0}^n (\mathcal{B}U)^k \mathcal{B}.$$

Some basic properties of  $\mathcal{H}$  and  $\Phi(t)$  which are immediate consequences of Lemma 3.6 and the structure of  $Y(t), \Phi(t)$  and  $\mathcal{H}$  are listed in

LEMMA 3.7. *Let  $g \in C[0, 1]$ . Then*

$$(i) \quad \|\mathcal{H}g\|_\delta \leq \text{const.} \|g\|_\delta, \quad \|\mathcal{H}Rg\|_\delta \leq \text{const.} \delta \|Rg\|_\delta,$$

where the constants are independent of  $\delta$ ,

$$(ii) \quad \mathcal{H}g \in C[0, 1] \cap C^1(0, 1],$$

$$(iii) \quad (\mathcal{H}g)(0) = -(M + R)^{-1}(I - R)g(0) + (M + R)^{-1}(\mathcal{B}Rg)(0),$$

$$(iv) \quad \Phi \in C[0, 1] \cap C^1(0, 1] \text{ and } \Phi(0) = (M + R)^{-1}R.$$



Furthermore, if Condition 3.1 holds, then  $\Phi \in C^\infty[0, 1]$ , and if in addition  $\alpha_r \neq 0$ , then  $\Phi(t)$  and all its derivatives vanish at  $t = 0$ .

The most general linear equation that we shall consider is

$$(3.9) \quad T(t)y' - (M + A(t))y(t) = g(t), \quad 0 < t \leq 1, \quad y \in C[0, 1] \cap C^1(0, 1],$$

where  $T, M$  are defined as previously,  $A, g \in C[0, 1]$  and

$$(3.10) \quad (I - R)A(0) = 0.$$

It follows from (3.7) that every solution of (3.9) must satisfy

$$(3.11) \quad y(t) = \Phi(t)\eta + (\mathcal{H}g)(t) + (\mathcal{H}Ay)(t)$$

where  $\eta = (P + R)y(\delta)$ . Now for some  $\eta \in X_n$  consider the iteration

$$(3.12) \quad y_{\nu+1}(t) = \Phi(t)[P + R]\eta + (\mathcal{H}g)(t) + (\mathcal{H}Ay_\nu)(t), \quad \nu = 0, 1, 2, \dots; \quad y_0 \in C[0, 1].$$

From Lemma 3.7,  $y_\nu \in C[0, 1]$ ,  $\|\mathcal{H}g\|_\delta \leq \text{const.} \|g\|_\delta$  and  $\|\mathcal{H}A\|_\delta \leq \text{const.} (\|I - R\|_\delta + \delta\|RA\|_\delta)$ . Hence, by (3.10), the iteration is contracting for  $t \in [0, \delta]$  if  $\delta$  is taken sufficiently small. This establishes the existence and uniqueness of a continuous solution of (3.11) on  $[0, \delta]$  for any  $\eta$  when  $\delta$  is sufficiently small. A standard contraction and translation argument on the rest of the interval now establishes the existence of a unique solution there.

Hence, every solution of (3.9) satisfies

$$(3.13) \quad y(t) = Z(t)\eta + (\mathcal{G}g)(t)$$

where

$$\mathcal{G} = (I - \mathcal{H}A)^{-1}\mathcal{H}, \quad Z = (I - \mathcal{H}A)^{-1}\Phi,$$

and

$$\eta = (P + R)y(\delta)$$

when  $\delta$  is sufficiently small. Note that  $Z$  is the unique solution of

$$(3.14)$$

$$T(t)Z'(t) - (M + A(t))Z(t) = 0; \quad PZ(\delta) = P, \quad RZ(\delta) = R, \quad Z \in C[0, 1] \cap C^1(0, 1]$$

and that  $\tilde{y} = \mathcal{G}g$  is the unique particular solution satisfying

$$\begin{aligned} T(t)\tilde{y}'(t) - (M + A(t))\tilde{y}(t) &= g(t); & P\tilde{y}(\delta) &= 0, \\ R\tilde{y}(\delta) &= 0, & \tilde{y} &\in C[0, 1] \cap C^1(0, 1]. \end{aligned}$$

Let  $p = \text{rank } [P + R]$ ,  $W$  be an  $n \times p$  matrix consisting of linearly independent columns of  $P + R$ , and define

$$(3.15) \quad X(t) = Z(t)W.$$

Then, by (3.13), we have

**THEOREM 3.1.** *Any solution of (3.9) has the form*

$$(3.16) \quad y(t) = X(t)\beta + \tilde{y}(t)$$

with a unique  $\beta \in X_p$ .

We now consider (3.9) subject to the linear boundary conditions

$$(3.17) \quad B_0y(0) + B_1y(1) = \gamma.$$

Our aim is to establish conditions on  $B_0$  and  $B_1$  which lead to a Fredholm alternative for (3.9), (3.17). To do this, it is convenient to introduce the differential expression

$$l(y) = Ty' - (M + A)y$$

and associate with it the operator defined by

$$\mathcal{L}y = l(y)$$

for  $y \in \mathcal{D} = \{y \in C[0, 1] \mid Ty' \in C[0, 1], B_0y(0) + B_1y(1) = 0\}$ . Then we have

**THEOREM 3.2.** *If*

$$(3.18) \quad \text{rank } [B_0, B_1] = k$$

*then  $\mathcal{L}$  is Fredholm with index  $p - k$ . Furthermore, if  $\mathcal{L}^{-1}$  exists, it is bounded.*

*Proof.* From (3.16),  $y$  and  $l(y) \in C[0, 1]$  iff

$$(3.19) \quad y(t) = (\mathcal{G}g)(t) + X(t)\beta = \tilde{y}(t) + X(t)\beta$$

for some  $g \in C[0, 1]$  and  $\beta$ . Hence,  $y \in \mathcal{D}$  iff (3.19) holds and

$$[B_0X(0) + B_1X(1)]\beta = -[B_0\tilde{y}(0) + B_1\tilde{y}(1)].$$

Thus  $g \in C[0, 1]$  is in the range of  $\mathcal{L}$  iff

$$(3.20) \quad B_0\tilde{y}(0) + B_1\tilde{y}(1) \in \text{range } [B_0X(0) + B_1X(1)].$$

To examine this condition, we need only examine the  $k$  linearly independent rows of  $[B_0, B_1]$ . Hence we may assume that  $B_0$  and  $B_1$  have  $k$  rows.

Let  $\text{rank } [B_0X(0) + B_1X(1)] = q$  and  $v_l, l = 1, \dots, k - q$  be a basis for the nullspace of  $[B_0X(0) + B_1X(1)]^*$ . Then (3.20) is satisfied iff

$$(3.21) \quad v_l^* [B_0\tilde{y}(0) + B_1\tilde{y}(1)] = 0, \quad l = 1, \dots, k - q.$$

We now show that the  $k - q$  linear functionals on  $C[0, 1]$  defined by (3.21) are linearly independent. If this were not so, then there would exist a nonzero vector  $w$  in the nullspace of  $[B_0X(0) + B_1X(1)]^*$ , i.e.

$$(3.22) \quad w^* [B_0X(0) + B_1X(1)] = 0$$

such that

$$(3.23) \quad w^* [B_0\tilde{y}(0) + B_1\tilde{y}(1)] \equiv 0.$$

By (3.16), (3.22) and (3.23),

$$w^* [B_0y(0) + B_1y(1)] = 0$$

whenever  $y$  satisfies (3.9) for some  $g$ . As the set of such  $y$  contains  $C^1[0, 1]$ , it follows that

$$w^* [B_0, B_1] = 0,$$

which contradicts (3.18).

Thus, the range of  $\mathcal{L}$  is the intersection of the nullspaces of  $k - q$  linearly independent bounded linear functionals on  $C[0, 1]$ . So the range is closed and its codimension is  $k - q$ . Clearly the nullspace of  $\mathcal{L}$  is  $\{X(t)\beta \mid [B_0X(0) + B_1X(1)]\beta = 0\}$  and so has dimension  $p - q$ .

To establish that  $\mathcal{L}$  is Fredholm with index  $p - k$  it remains to show that  $\mathcal{L}$  is closed.

Let  $y_\nu$  be a sequence in  $\mathcal{D}$  such that  $y_\nu \rightarrow y$  and  $v_\nu = \mathcal{L}y_\nu \rightarrow v$  as  $\nu \rightarrow \infty$ . Since the range of

$\mathcal{L}$  is closed it follows that  $v \in \text{range}(\mathcal{L}) \subseteq C[0, 1]$  and

$$y(t) = Z(t)[Py(\delta) + Ry(\delta)] + (\mathcal{G}v)(t) \in \mathcal{D}.$$

Hence  $\mathcal{L}$  is closed.

As  $\mathcal{L}$  is closed,  $\mathcal{D}$  with the norm

$$\|y\|_{\mathcal{L}} = \|y\|_{\infty} + \|\mathcal{L}y\|_{\infty}$$

is a Banach space. Hence if  $\mathcal{L}^{-1}$  exists it is bounded by the bounded inverse theorem.  $\square$

Observing (3.10) and noting that  $B$  and  $R$  commute, we obtain from (3.11) and Lemma 3.7, (iii), (iv),

$$(I - R)(M + R)y(0) = -(I - R)g(0).$$

Since  $(I - R)(M + R) = M$ , this yields

$$(3.24) \quad \begin{aligned} & (M + R)y(0) = (R - I)g(0) + Ry(0), \\ \text{or} & \\ & y(0) = (M + R)^{-1}((R - I)g(0) + Ry(0)). \end{aligned}$$

Hence the boundary conditions (3.17) are equivalent to

$$(3.25) \quad B_0Ry(0) + B_1y(1) = \tilde{\gamma}$$

where

$$\tilde{\gamma} = \gamma + B_0(M + R)^{-1}(I - R)g(0).$$

It turns out that (3.25) is advantageous for some numerical schemes applied to the boundary value problem in question.

In applications we are primarily interested in the case when  $\mathcal{L}$  is Fredholm with index zero. We therefore assume that  $B_0, B_1$  are  $p \times n$  matrices,  $\gamma$  is a  $p$  vector and that (3.18) holds with  $k = p$ . On substitution of (3.16) into (3.17) we find

**THEOREM 3.3.** *The problem (3.9), (3.17) has a unique solution for all  $g \in C[0, 1]$  and  $\gamma$  iff the  $p \times p$  matrix  $[B_0X(0) + B_1X(1)]$  is nonsingular.*

*Remarks.* The restriction that the solution be continuous at  $t = 0$  is unsatisfactory when constructing numerical schemes. What is desired in this case is an algebraic restriction on the solution. It turns out that the relation obtained from (3.24),

$$(3.26) \quad Qy(0) = Q(M + R)^{-1}((R - I)g(0) + Ry(0))$$

is satisfactory. Equations (3.25) and (3.26) are the  $n$  linearly independent boundary conditions which must be employed when (3.9), (3.17) is discretized by a difference scheme. This and related questions are discussed in de Hoog and Weiss [6].

Clearly, certain extensions to the above theory are possible. For example, the structure of  $M$  can be generalized to  $M = D + U$  where  $D$  is the block diagonal matrix defined previously and  $U$  satisfies  $S_1US_2U \cdots S_nU \equiv 0$  for any set of block diagonal matrices  $S_1, \cdots, S_n$ . Then Lemma 3.7 and all subsequent results can be established in a straightforward manner. Also  $M$  can be replaced by  $M + E$  when  $\|E\|$  is small. In this case, the iteration corresponding to (3.12) will still converge and all subsequent theory is easily extended. Such an analysis can be used to examine the perturbation in the solution due to perturbations in the boundary conditions and in the coefficients of the differential equations.

Another possible extension is to allow that the coefficients of  $RA$  be in  $C(0, 1] \cap L_1(0, 1)$  rather than in  $C[0, 1]$ . The iteration (3.12) still converges when  $\delta$  is sufficiently

small and hence Theorems 3.1, 3.2 and 3.3 remain valid. It is for this reason that we have not treated explicitly the case where some of the  $\alpha_k$  are in  $(0, 1)$ .

**4. Nonlinear problems.** Here we examine systems of the form

$$(4.1) \quad T(t)y' - f(t, y) = 0, \quad y \in C[0, \delta] \cap C^1(0, \delta],$$

where  $T$  is as in (3.6), and  $f$  is a nonlinear mapping from a subset of  $[0, 1] \times X_n$  to  $X_n$ . Under appropriate hypotheses on  $f$  we shall derive an existence result for (4.1) with  $\delta$  sufficiently small.

We now list these hypotheses.

(i) There is vector  $\zeta \in X_n$ , with  $(0, \zeta)$  in the domain of  $f$ , such that

$$(I - R)f(0, \zeta) = 0.$$

With  $\zeta$  and some  $\rho_0 > 0$  we associate the set

$$S_{\rho_0} = \{z \in X_n \mid |z - \zeta| \leq \rho_0\}.$$

(ii)  $f(t, z)$  and  $\partial f(t, z)/\partial z$  are continuous on  $[0, 1] \times S_{\rho_0}$ ,

(iii) The matrix

$$M = (I - R)M = (I - R)\frac{\partial}{\partial z}f(0, \zeta)$$

is block upper triangular as in (3.6).

**THEOREM 4.1.** *Assume that the above conditions hold. Then there are positive constants  $\gamma$ ,  $\delta$  and  $\rho \leq \rho_0$  such that (4.1) subject to the conditions*

$$(4.2) \quad (P + R)y(\delta) = (P + R)\eta + (P + R)\zeta$$

has a unique solution on

$$S_{\rho, \delta} = \{x \in C[0, \delta] \mid \|x - \zeta\|_{\delta} \leq \rho\}$$

provided that

$$|(P + R)\eta| \leq \gamma.$$

*Proof.* By hypothesis (ii) there are constants  $F$  and  $L$  and nondecreasing scalar functions  $a, b \in C[0, 1]$ ,  $c, d \in C[0, \delta_0]$  with  $a(0) = b(0) = c(0) = d(0) = 0$ , such that for all  $t \in [0, 1]$  and  $z \in S_{\rho_0}$ ,

$$(4.3a) \quad |f(t, z)| \leq F,$$

$$(4.3b) \quad \left| \frac{\partial f}{\partial z}(t, z) \right| \leq L,$$

$$(4.3c) \quad |(I - R)(f(t, z) - M(z - \zeta))| \leq a(t) + c(|z - \zeta|)|z - \zeta|,$$

$$(4.3d) \quad \left| (I - R)\left(\frac{\partial f}{\partial z}(t, z) - M\right) \right| \leq b(t) + d(|z - \zeta|).$$

We now rewrite (4.1) as

$$T(t)(y - \zeta)' - M(y - \zeta) - (I - R)(f(t, y) - M(y - \zeta)) - Rf(t, y) = 0.$$

With the new dependent variable  $x = y - \zeta$ , the problem (4.1), (4.2) becomes

$$T(t)x' - Mx - (I - R)(f(t, \zeta + x) - Mx) - Rf(t, \zeta + x) = 0,$$

$$(P + R)x(\delta) = (P + R)\eta.$$

By (3.7), this can be written as

$$(4.4) \quad x = \Psi(x),$$

where

$$(4.5) \quad \begin{aligned} & (\Psi(z))(t) \\ &= \Phi(t)(P+R)\eta + (\mathcal{H}(I-R)(f(\cdot, \zeta+z) - Mz))(t) + (\mathcal{H}Rf(\cdot, \zeta+z))(t). \end{aligned}$$

We now show that  $\Psi$  is a contraction on

$$T_{\rho,\delta} = \{x \in C[0, \delta] \mid \|x\|_\delta \leq \rho\}$$

for sufficiently small  $\delta$ ,  $\rho$  and  $|(P+R)\eta|$ . First, note that although  $\Phi(t)$  and  $H$  depend on  $\delta$ , we have the estimates

$$(4.6) \quad \|\Phi\|_\delta \leq \varphi, \quad \|\mathcal{H}z\|_\delta \leq h(\delta\|Rz\|_\delta + \|(I-R)z\|_\delta)$$

where the constants  $\varphi$  and  $h$  are independent of  $\delta$ . The first estimate follows from the definition of  $\Phi$  and Lemma 3.6(i), while the second estimate is just Lemma 3.7(i). When  $z \in T_{\rho,\delta}$  it follows from (4.5), (4.6) and (4.3a, c) that

$$\|\Psi(z)\| \leq \varphi|(P+R)\eta| + h(\delta F + a(\delta) + c(\rho)\rho).$$

Hence if  $\rho$ ,  $\delta$  and  $\gamma$  are taken so small that

$$(4.7) \quad hc(\rho) \leq \frac{1}{3}, \quad h(\delta F + a(\delta)) \leq \rho/3, \quad \gamma\varphi \leq \rho/3,$$

then  $\Psi$  maps  $T_{\rho,\delta}$  into itself. The mean value theorem for operators yields

$$\begin{aligned} \Psi(z_1) - \Psi(z_2) &= \mathcal{H}(I-R) \int_0^1 \left( \frac{\partial f}{\partial z}(\cdot, z_2 + s(z_1 - z_2)) - M \right) ds (z_1 - z_2) \\ &\quad + \mathcal{H}R(f(\cdot, z_1) - f(\cdot, z_2)), \quad z_1, z_2 \in T_{\rho,\delta}, \end{aligned}$$

and using (4.5), (4.3b, d),

$$\|\Psi(z_1) - \Psi(z_2)\| \leq h(b(\delta) + d(\rho) + \delta L)\|z_1 - z_2\|.$$

So, if in addition to (4.7),  $\rho$  and  $\delta$  are such that

$$(4.8) \quad h(b(\delta) + d(\rho) + \delta L) < 1,$$

then  $\Psi$  is a contraction on  $T_{\rho,\delta}$ .  $\square$

Let  $W$  be the  $n \times p$  matrix introduced in § 3, whose columns span  $\text{range}(P+R)$ . Then for each  $\eta \in X_n$  there is a unique vector  $\beta \in X_p$  such that  $(R+P)\eta = W\beta$ . Hence Theorem 4.1 ensures the existence of a  $p$  parameter family of solutions to (4.1), parameterized by  $\beta$ , with  $|\beta|$  sufficiently small, and  $p$  additional conditions are needed to extract a particular solution out of the family.

Theorem 4.1 extends a result of Russell [12] who considers the case where the matrix  $M$  consists of one block and has only eigenvalues with negative real part.

**5. Smoothness results.** Here we examine the smoothness of solutions of (4.1). Use will be made of the following lemma.

LEMMA 5.1. *Let Condition 3.1 be satisfied and  $g \in C^m[0, 1]$ . Then*

- (i)  $\mathcal{H}g \in C^m[0, 1]$ , and there are linear operators  $\mathcal{D}_i^k: C[0, \delta] \rightarrow C[0, \delta]$ ,  $k = 0, \dots, l$ ;  $l = 1, 2, \dots$ , with

$$\|\mathcal{D}_i^k\|_\delta \leq C_i^k,$$

where the constant  $C_i^k$  is independent of  $\delta$ , and matrix-valued functions  $d_i^k \in$

$C^\infty[0, \delta]$ ,  $k = 0, \dots, l-1$ , with  $d_l^k(0) = 0$ , such that

$$(\mathcal{H}g)^{(m)}(t) = \sum_{k=0}^m (\mathcal{D}_m^k g^{(k)})(t) + \sum_{k=0}^{m-1} d_m^k(t) g^{(k)}(\delta), \quad 0 \leq t \leq \delta;$$

(ii) when  $R = 0$ , there are matrices  $E_l^k$  such that for  $h \in C[0, \delta]$ ,

$$(\mathcal{D}_l^k h)(0) = E_l^k h(0).$$

*Proof.* (i) For the scalar case, the simplest cases of all, the result for  $\alpha \neq 0$  follows immediately from Eqns. (2.4), (2.5) in Lemma 2.4 and Lemma 2.6(iii), respectively, using induction in  $l$ . For  $\alpha = 0$ , the result is obvious.

When  $M$  consists of just one block, the result follows at once from (3.4) for the case  $\alpha > 1$ , from the analogous representation of  $\mathcal{B}_\rho$  for  $\alpha = 1$ , and the result is obvious for  $\alpha = 0$ . This immediately yields the result for the case when  $M$  is block diagonal, i.e.  $\mathcal{H} = \mathcal{B}$ .

The general result now follows from the representation (3.8).

(ii) when  $\alpha_r \neq 0$ , (i.e.  $R = 0$ ), and  $\partial^l f / \partial t^l(0, \zeta) = 0$ ,  $l = 0, \dots, m$ , then  $y^{(l)}(0) = 0$ , the general situation in the way outlined in (i).  $\square$

The main result of this section is

**THEOREM 5.1.** Assume

- (i)  $f$  satisfies the hypotheses of § 4;
- (ii) Condition 3.1 is satisfied;
- (iii)  $f \in C^m([0, 1] \times S_{\rho_0})$ .

Then

- (i)  $y \in C^m[0, \delta] \cap C^{m+1}(0, \delta]$ ;
- (ii) when  $\alpha_r \neq 0$ , (i.e.  $R = 0$ ), and  $\partial^l f / \partial t^l(0, \zeta) = 0$ ,  $l = 0, \dots, m$ , then  $y^{(l)}(0) = 0$ ,  $l = 0, \dots, m$ .

*Proof.* (i) We first assume that  $R = 0$ , and start with (4.4) which we write as

$$x(t) = \Phi(t)Px(\delta) + (\mathcal{H}\theta(\cdot, x))(t)$$

where

$$\theta(t, z) = f(t, \zeta + z) - Mz.$$

Note that by Lemma 3.7(iii), (iv)

$$(5.1) \quad x(0) = -M^{-1}\theta(0, x(0)).$$

Since  $(\mathcal{H}g)(0) = -M^{-1}g(0)$ , it follows that

$$|M^{-1}| \leq h,$$

where  $h$  is defined in (4.6). So (4.3d) and (4.8) applied to (5.1) yield

$$(5.2) \quad x(0) = 0.$$

The argument in the proof of Theorem 4.1 yields the (uniform) convergence of the sequence

$$(5.3) \quad \begin{aligned} x_0(t) &= 0, \\ x_{i+1}(t) &= \Phi(t)Px(\delta) + (\mathcal{H}\theta(\cdot, x_i))(t), \quad i = 0, 1, \dots, \end{aligned}$$

to  $x(t)$  on  $[0, \delta]$ , for  $\delta$  sufficiently small. By Lemma 5.1(i),  $x_i \in C^m[0, \delta]$ , and

$$(5.4) \quad \begin{aligned} x_{i+1}^{(1)}(t) &= \left( \mathcal{D}_1^1 \left( \frac{\partial f}{\partial z}(\cdot, \zeta + x_i) - M \right) x_i^{(1)} \right)(t) \\ &+ \left( \mathcal{D}_1^1 \frac{\partial f}{\partial t}(\cdot, \zeta + x_i) \right)(t) + (\mathcal{D}_1^0(f(\cdot, \zeta + x_i) - Mx_i))(t) \\ &+ \Phi^{(1)}(t)Px(\delta) + d_1^0(t)(f(\delta, \zeta + x_i(\delta)) - Mx_i(\delta)). \end{aligned}$$

Now take  $\delta$  so small that

$$C_1^1 \left\| \frac{\partial f}{\partial z}(\cdot, \zeta + x) - M \right\|_{\delta} \leq C_1^1(b(\delta) + d(\|x\|_{\delta})) < 1.$$

Then, as  $\|x_i - x\|_{\delta} \rightarrow 0$  for  $i \rightarrow \infty$ , the iteration (5.4) converges uniformly and  $x \in C^1[0, \delta]$  with  $\lim_{i \rightarrow \infty} x_i^{(1)}(t) = x^{(1)}(t)$ .

When  $m > 1$ , successive differentiation of (5.3) yields iteration schemes for  $x_i^{(l)}$ ,  $l = 2, \dots, m$ , of the form

$$(5.5) \quad x_{i+1}^{(l)}(t) = \left( \mathcal{D}_i^l \left( \frac{\partial f}{\partial z}(\cdot, \zeta + x_i) - M \right) x_i^{(l)} \right)(t) + r_i^l(t),$$

where  $r_i^l$  depends in a continuous fashion on  $x_i, x_i^{(1)}, \dots, x_i^{(l-1)}$ . For sufficiently small  $\delta$  the same argument as above yields convergence of these iterations; hence  $x \in C^m[0, \delta]$  and the proof for the case when  $R = 0$  is complete.

If  $R \neq 0$ , then it is clear that  $Ry \in C^1[0, \delta]$ . After replacing  $Ry$  in  $f(t, y)$  by  $Ry(t)$ , we may consider (4.1) as a differential system for the unknown  $(I - R)y$  only, and from the above it follows that  $(I - R)y \in C^1[0, \delta]$ . If  $m = 1$  the result is proved. If  $m > 1$ , the fact that  $y \in C^1[0, \delta]$  implies  $Ry \in C^2[0, \delta]$ , which in turn yields  $(I - R)y \in C^2[0, \delta]$ , and so on.

(ii) Setting  $t = 0$  in (5.4) and using (5.2) and Lemmas 5.1, 3.7(iv), yields the result for  $m = 1$ . When  $m > 1$ , note that by Lemmas 5.1, 3.7(iv),  $r_i^l(0) \rightarrow 0$  as  $i \rightarrow \infty$  when  $x_i(0), x_i^{(1)}(0), \dots, x_i^{(l-1)}(0)$  tend to zero and  $(\partial^k f / \partial t^k)(0, \zeta) = 0$ ,  $k = 1, \dots, l$ . The result now follows by induction.  $\square$

**6. Problems on infinite intervals.** Since problems on infinite intervals are a rich source of singular equations, we shall now show how they fit into the framework developed and give two examples.

Consider the linear first order system

$$(6.1) \quad x'(\tau) = S(\tau)(B(\tau)x(\tau) + h(\tau)), \quad 1 \leq \tau < \infty$$

where

$$(i) \quad S(\tau) = \text{diag}(I_1\tau^{\beta_1}, I_2\tau^{\beta_2}, \dots, I_r\tau^{\beta_r})$$

where the  $I_k$  are unit matrices of dimension  $\geq 0$ ,  $\beta_k \geq -1$ ,  $1 \leq k \leq r-1$ ,  $\beta_r < -1$ ,

$$(ii) \quad B \in C[1, \infty),$$

$$\lim_{\tau \rightarrow \infty} B(\tau) = N = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1r} \\ 0 & N_{22} & \cdots & N_{2r} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & N_{rr} \end{bmatrix}$$

where each  $N_{kk}$  is a square matrix of the same size as  $I_k$  and is nonsingular if  $\beta_k \geq -1$ ,

$$(iii) \quad h \in C[1, \infty) \quad \text{and} \quad \lim_{\tau \rightarrow \infty} h(\tau) \text{ exists.}$$

We are interested in "regular" solutions of (6.1), i.e. solutions which tend to a finite limit as  $\tau \rightarrow \infty$ .

The transformation  $\tau = 1/t$  applied to (6.1) yields

$$(6.2) \quad T(t)y' - (M + A(t))y = g(t), \quad 0 < t \leq 1,$$

where  $y(t) = x(1/t)$ ,

$$T(t) = \text{diag} (I_1 t^{\beta_1+2}, \dots, I_{r-1} t^{\beta_{r-1}+2}, I_r),$$

$$M_{lk} = -N_{lk}, \quad (l, k) \neq (r, r), \quad M_{rr} = 0$$

and

$$A(t) = -D(t)(B(1/t) + M), \quad g(t) = -D(t)h(1/t)$$

with

$$D = \text{diag} (I_1, \dots, I_{r-1}, I_r t^{-\beta_r-2}).$$

If  $\beta_r \leq -2$  then  $A \in C[0, 1]$ , and if  $-2 < \beta_r < -1$ , then  $(I - R)A \in C[0, 1]$ ,  $RA \in C[0, 1] \cap L^1(0, 1)$ . So all results of §§ 3 and 5 are valid for (6.2) and give corresponding results for (6.1). (Note the remark at the end of § 3.)

As an example we examine a problem which was considered recently by Franklin and Scott [8]. The fourth order equation

$$(6.3) \quad z^{(4)}(\tau) + a(\tau)z(\tau) = 0, \quad 1 \leq \tau < \infty,$$

describes the horizontal deflection of a pile vertically imbedded in soil ( $\tau - 1$  is the distance from the surface). Since soils usually get stiffer with depth we assume that the foundation coefficient  $a(\tau)$  has the form

$$a(\tau) = \tau^\sigma d(\tau), \quad \sigma \geq 0,$$

with

$$d \in C[1, \infty), \quad 0 < d(\tau), \quad \tau \in [0, \infty); \quad \lim_{\tau \rightarrow \infty} d(\tau) = \rho > 0.$$

We use the transformation suggested in Coddington and Levinson [2, p. 169]

$$x_l(\tau) = \tau^{-(l-1)\kappa} z^{(l-1)}(\tau), \quad \kappa = \sigma/4; \quad l = 1, 2, 3, 4,$$

to rewrite (6.3) as a first order system

$$x'(\tau) = \tau^\kappa B(\tau)x(\tau)$$

where

$$B(\tau) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\kappa\tau^{-(\kappa+1)} & 1 & 0 \\ 0 & 0 & -2\kappa\tau^{-(\kappa+1)} & 1 \\ -d(\tau) & 0 & 0 & -3\kappa\tau^{-(\kappa+1)} \end{bmatrix}.$$



This is a special case of (6.1) with  $r = 1$ ,  $\beta_1 = \sigma/4$  and

$$N = N_{11} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\rho & 0 & 0 & 0 \end{bmatrix}.$$

The matrix  $M = -N$  of the transformed system (6.2) has the eigenvalues  $\lambda_l = \rho^{1/4} e^{(2l-1)\pi i/4}$ ,  $l = 1, 2, 3, 4$ , with

$$\operatorname{Re} \lambda_1 = \operatorname{Re} \lambda_4 > 0, \quad \operatorname{Re} \lambda_2 = \operatorname{Re} \lambda_3 < 0.$$

So  $R = 0$ ,  $\operatorname{rank} P = p = 2$ , and by Theorem 3.1, (6.3) has exactly two linearly independent regular solutions. By Theorem 3.2, two linearly independent conditions must be imposed at  $\tau = 1$  to have a Fredholm alternative. We may prescribe any of the pairs  $(y(1), y'(1))$ ,  $(y(1), y''(1))$ ,  $(y'(1), y''(1))$ ,  $(y'(1), y'''(1))$ , and  $(y''(1), y'''(1))$ . In each case it is clear from the physical interpretation of the problem that homogeneous boundary conditions can only lead to the trivial solution; and the Fredholm alternative then yields the unique solvability in the case of nonzero boundary values.

In the nonlinear case, Theorem 4.1 can be used to examine regular solutions of first order systems

$$(6.4) \quad x'(\tau) = S(\tau)f(\tau, x), \quad 1 \leq \tau < \infty,$$

where  $f$  satisfies conditions analogous to (i)–(iii) in § 4. The resulting existence theory for (6.4) is an extension of work of Chang [1], who considers systems (6.4) with  $r = 1$  and  $\beta_1 = 0$ . An example is furnished by the Blasius problem

$$2z''' + zz'' = 0, \quad 1 \leq \tau < \infty, \quad z(1) = z'(1) = 0, \quad z'(\infty) = 1,$$

which describes the boundary layer on a flat plate; see Schlichting [13]. This problem is, of course, thoroughly understood, but its structure and the transformation to the form (6.4) are typical for a variety of other flow problems. We write

$$z = c + \tau + u$$

where  $c$  is a constant, and introduce the new dependent variables

$$x_1 = \tau u, \quad x_2 = \tau^2 u', \quad x_3 = \tau^3 u'', \quad x_4 = c.$$

Then  $x$  satisfies the first order system

$$(6.5) \quad \begin{aligned} x_1' &= x_1/\tau + x_2/\tau, \\ x_2' &= 2x_2/\tau + x_3/\tau, \\ x_3' &= -\tau x_3/2 + (3/\tau - x_4/2 - x_1/2\tau)x_3, \\ x_4' &= 0. \end{aligned}$$

This is a system of the type (6.4) with  $r = 4$ ,  $N_{11} = N_{12} = (1)$ ,  $N_{22} = (2)$ ,  $N_{23} = (1)$ ,  $N_{33} = (-1/2)$ ,  $N_{ik} = (0)$  otherwise,  $\beta_1 = \beta_2 = -1$ ,  $\beta_3 = 1$ ,  $\beta_4 = -2$ . Theorem 4.1 guarantees the existence of a two parameter family of regular solutions of (6.5) on an interval  $[\bar{\tau}, \infty)$  with  $\bar{\tau}$  sufficiently large. Each solution is uniquely defined by prescribing  $x_3(\bar{\tau})$  and  $x_4(\bar{\tau})$ .

**7. The eigenvalue problem.** The eigenvalue problem we consider is

$$(7.1) \quad \begin{aligned} Ty' - (M + A)y &= \lambda(N + C)y, & y \in C[0, 1] \cap C^1(0, 1] \\ B_0y(0) + B_1y(1) &= 0 \end{aligned}$$

where  $T, M, A, B_0$  and  $B_1$  are as in § 3,

$$N = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1r} \\ 0 & N_{22} & \cdots & N_{2r} \\ 0 & \cdots & 0 & N_{rr} \end{bmatrix}$$

is a constant matrix structured in the same way as  $M$ , and

$$(Cy)(t) = C(t)y(t), \quad C \in C[0, 1],$$

with  $(I - R)C(0) = 0$ .

Employing the notation of § 3, we write (7.1) as

$$\mathcal{L}_\lambda y = \{\mathcal{L} - \lambda(N + C)\}y = 0, \quad y \in \mathcal{D}.$$

Define

$$\text{def}(\mathcal{L}_\lambda) = \text{codimension of the range of } \mathcal{L}_\lambda$$

and

$$\text{nul}(\mathcal{L}_\lambda) = \text{dimension of the nullspace of } \mathcal{L}_\lambda.$$

We assume that  $\text{nul}(\mathcal{L}) = 0$ , i.e.  $\mathcal{L}$  is invertible. Let  $\Omega$  be the open connected set containing zero such that  $\lambda \in \Omega$  implies that all eigenvalues of  $(I - R)(M + \lambda N) + R$  have nonzero real parts. From Theorem 3.2,  $\mathcal{L}_\lambda$  is Fredholm with index zero for  $\lambda \in \Omega$ . Furthermore we have

**LEMMA 7.1.** *For each  $\lambda_0 \in \Omega$  there is an  $\varepsilon > 0$  such that  $\text{nul}(\mathcal{L}_\lambda) = \text{const.}$  for all  $0 < |\lambda - \lambda_0| < \varepsilon$ .*

*Proof.* Clearly,

$$\mathcal{L}_\lambda = \mathcal{L}_{\lambda_0} + (\lambda_0 - \lambda)(N + C).$$

As  $N + C$  is bounded, it is also  $\mathcal{L}_{\lambda_0}$  bounded, and the result follows from Kato [9, Thm. 5.31, p. 241].  $\square$

Define the spectrum

$$\Lambda = \{\lambda \in \Omega \mid \text{nul}(\mathcal{L}_\lambda) > 0\}.$$

Since  $\text{nul}(\mathcal{L}) = 0$  Lemma 7.1 immediately yields

**COROLLARY 7.1.** *Every compact subset of  $\Omega$  contains at most a finite number of eigenvalues.*

We have established the first part of

**THEOREM 7.1.** (i) *The spectrum  $\Lambda$  has no limit point in  $\Omega$ . For  $\lambda \notin \Lambda$ ,  $\mathcal{L}_\lambda^{-1}$  exists and is bounded.*

(ii) *Let*

$$\mathcal{P}_{\lambda_0} = -\frac{1}{2\pi i} \int_{\Gamma_{\lambda_0}} \mathcal{L}_\lambda^{-1}(N + C) d\lambda$$

where  $\lambda_0 \in \Lambda$ ,  $\Gamma_{\lambda_0} = \{\lambda \in \Omega \mid |\lambda - \lambda_0| = \delta\}$  and  $\delta$  is so small that there is no  $\lambda_1 \in \Lambda$  with  $|\lambda_1 - \lambda_0| \leq \delta$ . Then  $\mathcal{P}_{\lambda_0} : C[0, 1] \rightarrow \mathcal{D}$  is a projection with a finite dimensional range, which is invariant under the mapping  $\mathcal{L}_\lambda^{-1}(N + C)$ ,  $\lambda \notin \Lambda$ .

*Proof.* (i) Proceeding as in Dunford and Schwartz [7, pp. 600–601] we can derive the identity

$$(7.2) \quad \mu^2(\mu I - \mathcal{L}^{-1}(N + C)) = \mu I + \mathcal{L}_\lambda^{-1}(N + C), \quad \mu = 1/\lambda; \quad \lambda \notin \Lambda.$$

Hence

$$(7.3) \quad \mathcal{P}_{\lambda_0} = \frac{1}{2\pi i} \int_{\Gamma_{\mu_0}} (\mu - \mathcal{L}^{-1}(N + C))^{-1} d\mu,$$

where  $\mu_0 = 1/\lambda_0$  and  $\Gamma_{\mu_0}$  is defined analogously to  $\Gamma_{\lambda_0}$ . By (7.3)  $\mathcal{P}_{\lambda_0}$  is a projection, and the invariance of range  $(\mathcal{P}_{\lambda_0})$  under the mapping  $\mathcal{L}_\lambda^{-1}(N + C)$  follows from (7.2) and standard properties of spectral projections.

To see that range  $(\mathcal{P}_{\lambda_0})$  is finite dimensional we proceed as follows. It is easily verified that  $\mu_0 I - \mathcal{L}^{-1}(N + C): \mathcal{D} \rightarrow \mathcal{D}$  is Fredholm with  $\text{def}(\mu_0 I - \mathcal{L}^{-1}(N + C)) < \infty$  and  $\text{nul}(\mu_0 I - \mathcal{L}^{-1}(N + C)) < \infty$ . Hence it follows from Kato [9 Thms. 5.10 (p. 233), 5.28 (p. 239)] that  $\mathcal{P}_{\lambda_0}\mathcal{D}$  is finite dimensional. But  $\text{range}(\mathcal{P}_{\lambda_0}) = \mathcal{P}_{\lambda_0}C[0, 1] = \mathcal{P}_{\lambda_0}\mathcal{P}_{\lambda_0}C[0, 1] \subset \mathcal{P}_{\lambda_0}\mathcal{D}$ , whence  $\text{range}(\mathcal{P}_{\lambda_0}) = \mathcal{P}_{\lambda_0}\mathcal{D}$ .  $\square$

Let  $\text{range}(\mathcal{P}_{\lambda_0}) = \text{span}\{\phi_1, \dots, \phi_\beta\}$ . The  $\phi_j$  are generalized eigenfunctions of (7.1) corresponding to the eigenvalue  $\lambda_0$ .

**THEOREM 7.2.** *Assume that  $A, C \in C^m[0, 1]$  and that Condition 3.1 holds. Then  $\phi_j \in C^m[0, 1] \cap C^{m+1}(0, 1)$ ,  $j = 1, \dots, \beta$ . Furthermore, if  $\alpha_r \neq 0$  then  $\phi_j^{(l)}(0) = 0$ ,  $l = 0, \dots, m$ ;  $j = l, \dots, \beta$ .*

*Proof.* As the range of  $\mathcal{P}_{\lambda_0}$  is invariant under  $\mathcal{L}^{-1}(N + C)$ ,

$$\mathcal{L}\phi_j = \sum_{k=1}^{\beta} \alpha_{jk}(N + C)\phi_k, \quad j = 1, \dots, \beta,$$

where the  $\beta \times \beta$  matrix  $(\alpha_{jk})$  has the single eigenvalue  $\lambda_0$  and can be assumed to be in Jordan canonical form. Hence each  $\phi_j$  is contained in a "chain" of elements  $\{\phi_r, \phi_s, \phi_b, \dots\}$  satisfying

$$(7.4) \quad \mathcal{L}_{\lambda_0}\phi_r = 0, \quad \mathcal{L}_{\lambda_0}\phi_s = (N + C)\phi_r, \quad \mathcal{L}_{\lambda_0}\phi_t = (N + C)\phi_s, \dots$$

The result now follows on applying Theorem 5.1 to each equation in (7.4).  $\square$

#### REFERENCES

- [1] K. W. CHANG, *Perturbations of nonlinear differential equations*, J. Math. Anal. Appl., 34 (1971) pp. 418–428.
- [2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [3] F. R. DE HOOG AND R. WEISS, *Difference methods for boundary value problems with a singularity of the first kind*, SIAM J. Numer. Anal., 13 (1976), pp. 775–813.
- [4] ———, *Collocation methods for singular boundary value problems*, Ibid., 15 (1978), pp. 198–217.
- [5] ———, *The application of linear multistep schemes to singular initial value problems*, Math. Comput., 31 (1977), pp. 676–690.
- [6] ———, *The numerical solution of boundary value problems with an essential singularity*, SIAM J. Numer. Anal., 16 (1979), pp. 637–669.
- [7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1967.
- [8] J. FRANKLIN AND R. SCOTT, *The solution of the beam equation with variable foundation coefficient*, Manuscript, Calif. Inst. of Technology, 1977.
- [9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [10] P. RENTROP, *Numerical solution of the singular Ginzburg–Landau equation by multiple shooting*, Computing, 16 (1976), pp. 61–67.

- [11] P. RENTROP, *Eine Taylorreihenmethode zur numerischen Lösung von Zwei-Punkt Randwertproblemen mit Anwendung auf singuläre Probleme der nichtlinearen Schalentheorie*, Doctoral dissertation, Technical University Munich, 1977.
- [12] D. L. RUSSELL, *Numerical solution of singular initial value problems*, SIAM J. Numer. Anal., 7 (1970), pp. 399–417.
- [13] H. SCHLICHTING, *Boundary-Layer Theory*, 6th ed., McGraw-Hill, New York, 1968.

**VALUE SET AT  $x \in \Omega$  FOR AN ARBITRARY DISTRIBUTION WITH APPLICATIONS TO LOCAL EXTREMA OF  $f \in C(\Omega)$  AND A MAXIMUM PRINCIPLE FOR ORDINARY DIFFERENTIAL EQUATIONS\***

R. E. WHITE†

**Abstract.** In this paper we define a value set at  $x \in \Omega \subset \mathbb{R}^n$  which is nonempty for any distribution  $u \in \mathcal{D}'(\Omega)$ . We use this notion to generalize the classical theorems for monotonicity ( $\Omega \subset \mathbb{R}$ ) and local extrema for distributions in  $C(\Omega)$  which do not necessarily have classical derivatives. Also we show how these results are applicable in developing maximum principles for ordinary differential equations which have coefficients or an inhomogeneous term that may be distributions which are represented by locally integrable functions.

**1. Introduction.** The main result of this paper is the definition of a value set at  $x \in \mathbb{R}^n$  for any distribution  $u \in \mathcal{D}'(\Omega)$ . Roughly, a value set at  $x$  of a distribution reflects the range of values for suitable approximating continuous functions. Our definitions will give the following examples: One, the value set at  $x = 0$  for the Heavyside function,  $H(x) = 0$  when  $x < 0$  and  $1$  when  $x \geq 1$ , is equal to  $[0, 1]$ . Two, the value set at  $x = 0$  for the delta functional  $\delta(\phi) \equiv \phi(0)$  where  $\phi \in \mathcal{D}(\Omega)$ , is equal to  $[0, \infty]$ . Three, if  $f$  is continuous at  $x$ , then the value set of  $f$  at  $x$  is  $\{f(x)\}$ . All distributions have a nonempty value set at each  $x \in \Omega$ . Also if  $u \in \mathcal{D}'(\Omega)$  has value at  $x$  as defined by S. Lojasiewicz [6] (also see P. Antosik, J. Mikusinski and R. Sikorski [2]), then the value set is a singleton with element being the value at  $x$ . In particular, the derivative of any distribution will have a value set at  $x$  which we will call the derivative set at  $x$  of the original distribution. Also any regular Mikusinski operator which properly contains  $\mathcal{D}'_+(\Omega)$  as defined by T. K. Boehme [3] will have a nonempty "value set at  $x$ ".

The idea of assigning sets with derivatives that do not exist classically has been touched upon in at least two areas. One, in R. T. Rockafellar [8] the subdifferential of convex functions from  $\mathbb{R}^n$  into  $\mathbb{R}$  is defined as the set of  $\partial(x) \equiv \{x^* \in \mathbb{R}^n \mid \text{for all } z \in \mathbb{R}^n \text{ such that } f(z) \geq f(x) + x^* \cdot (z - x)\}$ . In particular, if  $f$  is differentiable, then  $\partial f(x) = \{\nabla f(x)\}$ . The notion of a subdifferential of convex functions is used to study minmax problems. The notion of value set is defined for all distributions and will be used to study minmax problems for continuous problems. The value sets of the partial derivatives give more information about the function being considered than just the subdifferential. Two, in G. Stampacchia [10] the notion of a second derivative of a distribution being positive in the sense of distributions is used. This is used to establish a weak maximum principle for elliptic differential operators with badly behaved coefficients or inhomogeneous term. For example, consider the well known steady state string problem with point force at  $x$ ,  $Lu \equiv -u'' = \delta(x - x_0)$ . Then  $u$  does not have a second derivative at  $x_0$  and it does not make sense classically to say that  $Lu \geq 0$ . However, it is true for all  $0 \leq \phi \in \mathcal{D}(a, b) \equiv C_c^\infty(a, b)$  that

$$(Lu)(\phi) = \int_a^b -u''(y)\phi(y) dy = \int_a^b u'(y)\phi'(y) dy = \phi(x_0) = \delta(\phi) \geq 0.$$

Because of our more refined tools (value sets) for studying minmax problems, we are able to establish a strong maximum principle for elliptic operators with badly behaved (but not quite as bad) coefficients or inhomogeneous term. Because of these generalizations and the applicability of this notion, it seems the definition of a value set is and will be of significance.

\* Received by the editors May 14, 1976 and in final revised form November 8, 1978.

† North Carolina State University, Department of Mathematics, Raleigh, North Carolina 27650.

In § 2 we review some of the pertinent facts concerning distributions. Section 3 contains the definition of value set of an arbitrary distribution as well as examples and some of the basic properties. In § 4 we demonstrate via value sets that the classical theorems about monotonicity and local extrema may be generalized to continuous functions. Finally in § 5 we use the results of the previous sections to prove the strong maximum principle for the ordinary differential equation  $(pu')' + gu' + qu = f$  where  $0 < m \leq p \in L_\infty(a, b)$ ,  $g \in L_\infty(a, b)$ ,  $q = Q'$ ,  $f = F'$  with  $Q \in L_\infty^{loc}(a, b)$  and  $F \in L_2(a, b)$ .  $qu$  is defined via integration by parts as the derivative of  $Qu - \int^x Qu' \in L_2(a, b)$ . In particular,  $q$  and  $f$  could be the delta functional or could be a function of the form  $x^{-1-\alpha}$ ,  $\alpha > \frac{1}{2}$  and  $\Omega = (-1, 1)$ .

**2. Distributions.** This section contains some of the basic facts about distributions. For more details the reader should consult L. Schwartz [9].

A test function,  $\phi$ , on  $\Omega \subset \mathbb{R}^n$  is any  $C^\infty(\Omega)$  function whose support, the closure in  $\Omega$  of the set of  $x \in \Omega$  such that  $\phi(x) \neq 0$ , is a compact subset of  $\Omega$ . We will denote all such test functions by  $\mathcal{D}(\Omega)$ . A sequence of  $\phi_k \in \mathcal{D}(\Omega)$  is said to converge in  $\mathcal{D}(\Omega)$  to  $\phi$  if and only if (i) support  $\phi_k, \phi \subset K$  where  $K$  is a compact subset of  $\Omega$  independent of  $k$  and (ii)  $(\partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n})\phi_k \rightarrow (\partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n})\phi$  converge uniformly on all compact subsets of  $\Omega$  for all  $\alpha = (\alpha_1, \dots, \alpha_n)$ ,  $\alpha_1, \dots, \alpha_n$  nonnegative integers and  $|\alpha| \equiv \alpha_1 + \dots + \alpha_n$ . A distribution,  $u$ ,  $\Omega \subset \mathbb{R}^n$  is a linear map from  $\mathcal{D}(\Omega) \rightarrow \mathbb{R}$  (or  $\mathbb{C}$ ) which is continuous; i.e., when  $\phi_k \rightarrow \phi$  in  $\mathcal{D}(\Omega)$ , then  $u(\phi_k) \rightarrow u(\phi)$ . The set of all distributions is a linear space when  $(\alpha u + \beta v)(\phi) \equiv \alpha u(\phi) + \beta v(\phi)$  and is denoted by  $\mathcal{D}'(\Omega)$ . Examples include (i)  $u(\phi) \equiv \int_\Omega f(x)\phi(x) dx$  where  $f \in L_1^{loc}(\Omega)$  and often we shall write  $u = f$ , (ii)  $u(\phi) \equiv \phi(0) = \delta(\phi)$  the so called delta "function", and (iii)  $u(\phi) \equiv \sum_{j=0}^\infty \phi^j(j)$  where  $\Omega = \mathbb{R}$ .

The  $\partial^{|\alpha|}/\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n} = \mathcal{D}^\alpha$  derivative of  $u \in \mathcal{D}'(\Omega)$ ,  $\mathcal{D}^\alpha u$ , is defined by  $\mathcal{D}^\alpha u(\phi) \equiv u((-\mathcal{D}^\alpha)\phi)$  and is itself a distribution. Examples include (i)  $\Omega = \mathbb{R}$ ,  $u(\phi) = H(\phi) = \int_0^\infty \phi(x) dx$  and so  $u'(\phi) = u(-\phi') = \int_0^\infty -\phi'(x) dx = \phi(0) = \delta(\phi)$  and (ii)  $\Omega \subset \mathbb{R}^n$  and  $u(\phi) = \int_\Omega f(x)\phi(x) dx$  where  $f \in C^1(\Omega)$  and so for  $|\alpha| = 1$ ,  $\mathcal{D}^\alpha u(\phi) = \int_\Omega f(x)(-\mathcal{D}^\alpha)\phi(x) dx = -f(x)\phi(x)|_{\partial\Omega} + \int_\Omega \mathcal{D}^\alpha f(x)\phi(x) dx = \int_\Omega \mathcal{D}^\alpha f(x)\phi(x) dx$ .

In general the product of distributions is not a distribution. However, the following subsets of  $\mathcal{D}'(\Omega)$  are closed in  $\mathcal{D}'(\Omega)$  under multiplication. First, consider  $\phi, \psi \in C^\infty(\Omega)$  and  $u \in \mathcal{D}'(\Omega)$  and then define  $\psi u(\phi) = u(\psi\phi)$ . Since  $\psi\phi \in \mathcal{D}(\Omega)$ , the definition is well-defined and in fact  $\psi u \in \mathcal{D}'(\Omega)$ . Second, consider  $f \in L_\infty^{loc}(\Omega)$  and  $g \in L_2^{loc}(\Omega)$  and define  $(fg)(\phi) = \int_\Omega f(x)g(x)\phi(x) dx$ . Because  $fg \in L_2^{loc}(\Omega) \subset L_1^{loc}(\Omega)$ , this definition is properly defined. Third, let  $H_{-1}(a, b) = \{q \in \mathcal{D}'(a, b) | q = Q', Q \in L_2(a, b)\}$  and  $H_1(a, b) = \{u \in L_2(a, b) | u' \in L_2(a, b)\}$ . Define  $qu$  as the derivative of  $Qu - \int^x Qu'$ . Since  $u' \in L_2(a, b)$ ,  $u \in C[a, b]$  and so  $Qu \in L_2(a, b)$ . Also  $u', Q \in L_2(a, b)$  implies  $Qu' \in L_1(a, b)$  and so  $\int^x Qu' \in C[a, b]$ . Thus  $Qu - \int^x Qu' \in L_2(a, b)$  and consequently  $qu \in H_{-1}(a, b)$ .

Not all distributions have value at a point. S. Lojasiewicz's [6] definition of value at a point will exist only for certain distributions of finite order. In fact, the existence of this value at  $x_0 \in \Omega$  is equivalent to  $\lim_{\varepsilon \rightarrow 0} u(\Gamma_\varepsilon)$  existing where  $\Gamma_\varepsilon \in C_c^m(S_\varepsilon(x_0))$ ,  $S_\varepsilon(x_0) = \{x \in \Omega | |x - x_0| < \varepsilon\}$ ,  $\int_\Omega \Gamma_\varepsilon(x) dx = 1$ ,  $\Gamma_\varepsilon(x) \geq 0$ ,  $\sup |\mathcal{D}^\alpha \Gamma(x)| = O(\varepsilon^{-m-1})$  where  $|\alpha|$  is the order of  $u = m$  and  $u = \mathcal{D}^\alpha F$ ,  $F \in C(\Omega)$ . In this case the value of  $u$  at  $x_0$  is  $\lim_{\varepsilon \rightarrow 0} u(\Gamma_\varepsilon)$  and must be independent of the choice of  $\Gamma_\varepsilon \in C_c^m(S_\varepsilon(x_0))$ . The definition of symmetric value given by P. Antosik [1] is equivalent to  $\lim_{\varepsilon \rightarrow 0} u(\phi)$  existing where  $\phi \in C_c^\infty(S_\varepsilon(x_0))$ ,  $\phi \geq 0$ ,  $\phi$  is even about  $x_0$  and  $\int_\Omega \phi(x) dx = 1$ . For example let  $\phi_n(x) = (n/c_1) e^{-1/(1-n^2x^2)}$  where  $c_1 = \int_{|x-x_0| \leq 1/n} e^{-1/(1-n^2x^2)} dx$  and let  $u = H$ . In this case  $\lim_{n \rightarrow \infty} u(\phi_n) = \frac{1}{2}$  and in fact if  $\phi_n$  is any such  $\phi$ , then  $u(\phi_n) = \int_0^{1/n} 1\phi_n(x) dx = \frac{1}{2}$  and

consequently the symmetric value of  $u$  at 0 is  $\frac{1}{2}$ . Clearly the value of  $u$  at  $x = 0$  does not exist in the sense of S. Lojasiewicz for the possible limits range from 0 to 1. It is important to note that if  $\phi \in C_c^\infty(S_\varepsilon(x_0))$ ,  $\phi \geq 0$ ,  $\int_\Omega \phi(x) dx = 1$  and  $u$  is continuous at  $x_0$  on  $S_\varepsilon(x)$ , then  $\lim_{\varepsilon \rightarrow 0} u(\phi) = u(x)$ . Finally one should note that the delta “functional” does not have symmetric value and hence does not have value in the sense of S. Lojasiewicz.

**3. Value sets at  $x$  of distributions.** Even though not all distributions have values at a point, all distributions do have a nonempty value set at  $x$ ,  $\mathcal{D}^0(u, x)$ . The value set of  $u \in \mathcal{D}'(\Omega)$  at  $x \in \Omega$  reflects the values of suitable approximating functions of  $u$ . Since any distribution has a value set, any derivative of a distribution has a value set or derivative set of  $u$  at  $x$ .

The value or derivative sets will be subsets of the two point compactification of the real line which is denoted by  $[-\infty, \infty]$ . Let  $T_\varepsilon(x) = \{\phi \in C_c^\infty(S_\varepsilon(x)) | x \in \Omega, \phi \geq 0 \text{ and } \int_\Omega \phi(x) dx = 1\}$ ,  $\mathcal{D}_\varepsilon^\alpha(u, x) \equiv \{u((-\mathcal{D})^\alpha \phi) | \phi \in T_\varepsilon(x)\}$  and  $\mathcal{D}_\varepsilon^\alpha(u, x) \subset [-\infty, \infty]$  be the closure in  $[-\infty, \infty]$  of  $\mathcal{D}_\varepsilon^\alpha(u, x)$ .

**DEFINITION.** The  $\alpha$ -th order derivative set at  $x \in \Omega \subset \mathbb{R}^n$  of  $u \in \mathcal{D}'(\Omega)$  is  $\mathcal{D}^\alpha(u, x) \equiv \bigcap_{\varepsilon > 0} \mathcal{D}_\varepsilon^\alpha(u, x)$ . The value set of  $u$  at  $x$  is  $\mathcal{D}^0(u, x)$ .  $u$  is said to have value at  $x$  if and only if  $\mathcal{D}^0(u, x)$  is a singleton and in this case we denote the element by  $u(x)$  and call it the value of  $u$  at  $x$ . In case  $\Omega \subset \mathbb{R}$  we may define left and right value sets and values at  $x$  of  $u \in \mathcal{D}'(\Omega)$  in the obvious manner.

*Remark.* The elements of  $T_\varepsilon(x)$  are called mollifiers. It is well known that they may be used in the approximation of distribution by smooth functions. This suggests that we use the set of values of approximating smooth functions for the definition of value set. This has two drawbacks. First, it is not clear which set of approximating smooth functions to use. Second, it is more difficult to prove the theorems in this paper.

Before giving some examples, we shall prove the following theorem which gives some of the routine properties of value sets.

**THEOREM 1.** Let  $v, u \in \mathcal{D}'(\Omega)$ .

1.  $\mathcal{D}^\alpha(u, x) \neq \phi$  and is closed interval contained in  $[-\infty, \infty]$ .
2.  $\mathcal{D}^\alpha(u, x) = \mathcal{D}^0(\mathcal{D}^\alpha u, x)$ .
3. If  $u$  has value at  $x$ , then  $\lim_{\varepsilon \rightarrow 0} u(\phi_\varepsilon) = u(x)$  where  $\phi_\varepsilon \in T_\varepsilon(x)$ .
4. The notion of value as defined in this paper is equivalent to the notion of value as defined by S. Lojasiewicz. In particular, if  $u \in L_1^{\text{loc}}(\Omega)$  and  $\lim_{y \rightarrow x} u(y)$  exists, then the value exists.
5. If  $\mathcal{D}^0(u, x) \subset (-\infty, \infty)$  and left and right values at  $x$ ,  $u(x+)$  and  $u(x-)$  exist then  $\mathcal{D}^0(u, x) = [u(x-), u(x+)]$  or  $\mathcal{D}^0(u, x) = [u(x+), u(x-)]$ .
6.  $\mathcal{D}^\alpha(u+v, x) \subset \mathcal{D}^\alpha(u, x) + \mathcal{D}^\alpha(v, x)$ . If  $\mathcal{D}^\alpha u$  has value at  $x$ , then  $\mathcal{D}^\alpha(u+v, x) = \mathcal{D}^\alpha u(x) + \mathcal{D}^\alpha(v, x)$ .
7. If  $u, v \in L_1^{\text{loc}}(\Omega)$ ,  $u$  is continuous at  $x$  and  $\mathcal{D}^0(v, x) \subset (-\infty, \infty)$ , then  $\mathcal{D}^0(uv, x) = u(x)\mathcal{D}^0(v, x)$ .
8. If  $u \in L_\infty^{\text{loc}}(\Omega)$ ,  $v \in L_1^{\text{loc}}(\Omega)$  and  $0 \in \mathcal{D}^0(v, x)$ , then  $0 \in \mathcal{D}^0(uv, x)$ .
9. If  $u, v \in H_1(a, b)$  and  $u, u', v, v'$  have values at  $x$ , then  $(uv)'$  has value at  $x$  equal to  $u(x)v'(x) + u'(x)v(x)$ .
10. With the obvious assumptions, rules 6, 7, 8 and 9 hold for right and left value set and values when  $\Omega \subset \mathbb{R}$ .

*Proofs.* 1.  $\mathcal{D}^\alpha(u, x)$  is an intersection of a nested family of closed sets contained in a compact space. Thus  $\mathcal{D}^\alpha(u, x) \neq \emptyset$  and is closed. Since  $T_\varepsilon(x)$  is convex and  $u$  is linear,  $\mathcal{D}_\varepsilon^\alpha(u, x)$  is an interval and consequently  $\mathcal{D}^\alpha(u, x)$  is an interval.

2. This is just notation since  $\mathcal{D}^\alpha u(\phi) \equiv u((-\mathcal{D})^\alpha \phi)$ .

3. Consider  $\Gamma_\varepsilon(x) = \{\phi_{\varepsilon'} \in T_{\varepsilon'}(x) | \varepsilon \leq \varepsilon'\}$ . Then the intersection  $\bigcap_{\varepsilon > 0} \{u(\phi) | \phi \in \Gamma_\varepsilon(x)\}$  is closed and nonempty. Since  $\Gamma_\varepsilon(x) \subset T_\varepsilon(x)$ , the intersection is also contained in  $\mathcal{D}^0(u, x)$ . Because  $\mathcal{D}^0(u, x) = \{u(x)\}$  and  $u(\phi_\varepsilon) \in \{u(\phi) | \phi \in \Gamma_\varepsilon(x)\}$ ,  $u(\phi_\varepsilon)$  must converge to  $u(x)$  as  $\varepsilon \rightarrow 0$ .

4. In order to see this, we need the following characterization of value in the sense of S. Lojasiewicz which is proved in [2]:  $\lim_{n \rightarrow \infty} u(\delta_n)$  exists and is independent of  $\delta_n \in C_c^\infty(S_{1/n}(x))$  with  $\delta_n \geq 0$  and  $\int_{S_{1/n}(x)} \delta_n(y) dy = 1$ . Property 3 implies that if  $\mathcal{D}^0(u, x)$  is a singleton, then this characterization holds. This characterization certainly implies that  $\mathcal{D}^0(u, x)$  is singleton.

5. Suppose  $u(x-) \leq u(x+)$ . One can show as in the proof of property 3 that for  $\phi_n^- \in \{\phi_n \in C_c^\infty(x - (1/n), x) | \phi_n n \geq 0, \int_{x-1/n}^x \phi_n(y) dy = 1\}$  and  $\phi_n^+ \in \{\phi_n \in C_c^\infty(x, x + (1/n)) | \phi_n n \geq 0, \int_x^{x+1/n} \phi_n(y) dy = 1\}$  we have  $u(\phi_n^-) \rightarrow u(x-)$  and  $u(\phi_n^+) \rightarrow u(x+)$ . Now let  $\psi_n \equiv \lambda \phi_n^+ + (1-\lambda) \phi_n^- \in T_{1/n}(x)$ ,  $0 \leq \lambda \leq 1$  and  $u(\psi_n) = \lambda u(\phi_n^+) + (1-\lambda)u(\phi_n^-) \rightarrow \lambda u(x+) + (1-\lambda)u(x-)$ . Thus  $\mathcal{D}^0(u, x) \supset [u(x-), u(x+)]$ .

In order to show  $\mathcal{D}^0(u, x) \subset [u(x-), u(x+)]$ , let  $v \in \mathcal{D}^0(u, x)$  and let  $\phi_n \in T_{1/n}(x)$  be such that  $u(\phi_n) \rightarrow v$  as  $n \rightarrow \infty$ . Let  $R_k \in C_c^\infty(x, b)$  such that  $R_k(y) \equiv 1$  for  $y \in (x + (1/k), b - (1/k))$  with  $k$  in general much larger than  $n$ . Let  $c_{k,n} \equiv \int_a^b R_k(y) \phi_n(y) dy$ . In a similar way define  $L_k$  and  $d_{k,n}$  for the left side of  $x$ . Since  $R_k \phi_n / c_{k,n}$  and  $L_k \phi_n / d_{k,n}$  are in  $T_{1/n}(x) \cap C_c^\infty(x, x + (1/n))$  and  $T_{1/n}(x) \cap C_c^\infty(x - (1/n), x)$ , respectively,  $u(R_k \phi_n / c_{k,n}) \rightarrow u(x+)$  and  $u(L_k \phi_n / d_{k,n}) \rightarrow u(x-)$  as  $n \rightarrow \infty$ . Also for each fixed  $n$ ,  $d_{k,n} u(L_k \phi_n / d_{k,n}) + c_{k,n} u(R_k \phi_n / c_{k,n}) \rightarrow u(\phi_n)$  as  $k \rightarrow \infty$ . Since  $\int_a^b \phi_n(y) dy = 1$ ,  $c_{k,n} \rightarrow c_n$ ,  $d_{k,n} \rightarrow d_n$  as  $k \rightarrow \infty$  and  $c_n + d_n = 1$ . Either the sequences  $\{c_n\}$  and  $\{d_n\}$  or subsequences of  $\{c_n\}$  and  $\{d_n\}$  converge to  $c$  and  $d$ , respectively, and  $c + d = 1$ . Thus  $u(\phi_n) \rightarrow cu(x-) + du(x+) \in [u(x-), u(x+)]$ . Consequently,  $\mathcal{D}^0(u, x) \subset [u(x-), u(x+)]$ .

6. First, we show that  $\mathcal{D}^\alpha(u+v, x) \subset \mathcal{D}^\alpha(u, x) + \mathcal{D}^\alpha(v, x)$  even when  $\mathcal{D}^\alpha(u, x)$  is not a singleton. Let  $\phi_n \in T_{1/n}(x)$  such that  $(u+v)((-\mathcal{D})^\alpha \phi_n) \rightarrow s \in \mathcal{D}^\alpha(u+v, x)$ . Since  $(u+v)((-\mathcal{D})^\alpha \phi_n) = u((-\mathcal{D})^\alpha \phi_n) + v((-\mathcal{D})^\alpha \phi_n)$  and by an argument similar to that given in the proof of Property 3, we must have  $u((-\mathcal{D})^\alpha \phi_n)$  and  $v((-\mathcal{D})^\alpha \phi_n)$  or subsequences converge to elements in  $\mathcal{D}^\alpha(u, x)$  and  $\mathcal{D}^\alpha(v, x)$ . Thus  $s = r + t$  where  $r \in \mathcal{D}^\alpha(u, x)$  and  $t \in \mathcal{D}^\alpha(v, x)$ .

Second, let  $\mathcal{D}^\alpha(u, x) = \{\mathcal{D}^\alpha u(x)\}$  and show  $\mathcal{D}^\alpha(u+v, x) \supset \mathcal{D}^\alpha u(x) + \mathcal{D}^\alpha(v, x)$ . Let  $r \in \mathcal{D}^\alpha(v, x)$  and  $\phi_n \in T_{1/n}(x)$  be such that  $v((-\mathcal{D})^\alpha \phi_n) \rightarrow r$ . By property 3 we also have  $u((-\mathcal{D})^\alpha \phi_n) \rightarrow \mathcal{D}^\alpha u(x)$ . Since  $(u+v)((-\mathcal{D})^\alpha \phi_n) = u((-\mathcal{D})^\alpha \phi_n) + v((-\mathcal{D})^\alpha \phi_n)$  and  $(u+v)((-\mathcal{D})^\alpha \phi_n) \rightarrow \mathcal{D}^\alpha u(x) + r$ , we have that  $\mathcal{D}^\alpha u(x) + r \in \mathcal{D}^\alpha(u+v, x)$ .

7. First, we show  $\mathcal{D}^0(uv, x) \supset u(x)\mathcal{D}^0(v, x)$ . Let  $r \in \mathcal{D}^0(v, x)$  and  $\phi_n \in T_{1/n}(x)$  such that  $v(\phi_n) \rightarrow r$ . Since  $u$  is continuous at  $x$  and  $u, v \in L_1^{\text{loc}}(\Omega)$ , we have

$$\begin{aligned} |(uv)(\phi_n) - u(x)v(\phi_n)| &= \left| \int_\Omega u(y)v(y)\phi_n(y) dy - \int_\Omega u(x)v(y)\phi_n(y) dy \right| \\ &= \left| \int_\Omega v(y)(u(y) - u(x))\phi_n(y) dy \right| \\ &\leq \varepsilon \left| \int_\Omega v(y)\phi_n(y) dy \right| \leq \varepsilon 2r, \quad n \geq N. \end{aligned}$$

Thus  $u(x)r \in \mathcal{D}^0(uv, x)$ .

Second, in order to show that there is an  $r \in \mathcal{D}^0(v, x) \subset (-\infty, \infty)$  such that  $s = u(x)r$  when  $s \in \mathcal{D}^0(uv, x)$ , let  $\phi_n \in T_{1/n}(x)$  be such that  $(uv)(\phi_n) \rightarrow s$ . Since  $u(x)$  exist,  $u(\phi_n) \rightarrow u(x)$ . Now  $\mathcal{D}^0(v, x)$  is bounded and so either  $v(\phi_n)$  converges or a subsequence



converges. If the latter is the case,  $u(\phi_{n_i})$  still converges to  $u(x)$  and so assume  $v(\phi_n) \rightarrow r$ . Now apply the above inequalities to show  $s = u(x) \cdot r$ .

8. Let  $\phi_n \in T_{1/n}(x)$  be such that  $v(\phi_n) \rightarrow 0$ . Since  $y \in L^\infty_{\text{loc}}(\Omega)$ , for  $y \in \text{support } \phi_n$  there exist  $M$  such that  $-M \leq u(y) \leq M$ . Thus we have  $|\int_{\Omega} u(y)v(y)\phi_n(y) dy| \leq M \cdot |\int_{\Omega} v(y)\phi_n(y) dy|$  and consequently  $(uv)(\phi_n) \rightarrow 0$  and so  $0 \in \mathcal{D}(uv, x)$ .

9. It is clear that  $uv \in H_1(a, b)$  and that  $(uv)' = u'v + uv'$ . Since  $u', v' \in L_2(a, b)$ ,  $u, v \in C[a, b]$ . By properties 4 and 7 we have  $\mathcal{D}^0(u'v, x) = v(x)\mathcal{D}^0(u', x) = v(x)u'(x)$ . By property 6 we have  $\mathcal{D}^0((uv)', x) = v(x)u'(x) + \mathcal{D}^0(uv', x)$ . Again by properties 4 and 7 we have  $\mathcal{D}^0(uv', x) = u(x)v'(x)$  and so  $(uv)'(x) = u'(x)v(x) + v'(x)u(x)$ .

10. These proofs follow by inspection of the previous proofs.

*Examples.* 1. Let  $u = H =$  the Heavyside function. Since  $\lim_{y \rightarrow 0^-} u(y) = 0$  and  $\lim_{y \rightarrow 0^+} u(y) = 1$  exist, by 4 and 5 we have  $\mathcal{D}^0(u, 0) = [0, 1]$ .

2. Let  $u = \delta =$  the delta "function". Clearly, by the proper choice of  $\phi_n \in T_{1/n}(x)$ ,  $\mathcal{D}^0(u, 0) \supset [0, \infty]$ . If there exists  $r \in \mathcal{D}^0(u, 0)$  which is negative, then there are  $\psi_n \in T_{1/n}(0)$  such that  $u(\psi_n) \rightarrow r$ . But  $u(\psi_n) = H(-\psi'_n) = \int_0^\infty -\psi'_n(y) dy = -\psi_n(y)|_{y=0}^{y=\infty} = \psi_n(0) \geq 0$  and so we have a contradiction. Also  $\mathcal{D}^0(u, x) = \{0\}$  when  $x \neq 0$ .

3. Let  $u = \delta'$  and  $v = \delta''$ . Both  $u$  and  $v$  have the same value sets for each  $x \in \mathbb{R}$ , namely,  $\mathcal{D}^0(u, 0) = \mathcal{D}^0(v, 0) = [-\infty, \infty]$  and  $\mathcal{D}^0(u, x) = \mathcal{D}^0(v, x) = \{0\}$  when  $x \neq 0$ . Thus the notion of value set of a distribution is in general not descriptive enough to retrieve the distribution from the collection of all value sets. It was proven in [6] if all the value sets are singletons, then we may retrieve the distribution from the collection of all values.

4. The following examples show that the assumptions in properties 6, 7, 8 and 9 are to some degree necessary. If  $u = 1 - H$  and  $v = H$ , then  $\mathcal{D}(u, 0) = [-\infty, 0]$ ,  $\mathcal{D}(v, 0) = [0, \infty]$  and  $\mathcal{D}(u + v, 0) = \{0\}$ . Also  $\mathcal{D}^0(v, 0) = [0, 1]$ ,  $\mathcal{D}^0(u, 0) = [0, 1]$  and  $\mathcal{D}^0(uv, 0) = \{0\}$ . If  $u = x^{-1/2}$  and  $v = x^{1/2}$ , then  $\mathcal{D}^0(u, 0+) = \{+\infty\}$ ,  $\mathcal{D}^0(v, 0+) = \{0\}$  and  $\mathcal{D}^0(uv, 0+) = \{1\}$ . Also  $\mathcal{D}(u, 0+) = \{-\infty\}$ ,  $\mathcal{D}(v, 0+) = \{+\infty\}$  and  $\mathcal{D}(uv, 0) = \{0\}$ .

5.  $u(\phi) \equiv \sum_{j=0}^i \phi^{(j)}(j)$ .  $\mathcal{D}^0(u, x) = \{0\}$  when  $x$  is not a positive integer or zero.  $\mathcal{D}^0(u, 0) = [0, \infty]$  and  $\mathcal{D}^0(u, j) = [-\infty, \infty]$  when  $j > 0$ .

6. See [6] or [2] for examples when the value exists at  $x$  but the distribution is not continuous at  $x$ .

7. See the examples at the end of the next section for examples of distributions in several variables.

As additional possible examples we note that the notion of a value set may be extended from  $\mathcal{D}'_+(\Omega)$  to the regular Mikusinski operators,  $\mathcal{MR}$ , as defined by T. K. Boehme [3]. For those readers who are familiar with this paper we briefly describe this extension. We shall use the notation of T. K. Boehme. A regular Mikusinski operator is a Mikusinski operator,  $a$ , in which the following are in its equivalence class  $f_n/\phi_n$  where  $f_n, \phi_n \in C =$  the usual convolution algebra of continuous functions and  $\{\phi_n\}$  is an approximate identity. Thus the following sets are nonempty  $\mathbb{D}_\varepsilon(a, x) \equiv \{f_\varepsilon(x) | a = f_\varepsilon/\phi_\varepsilon, f_\varepsilon, \phi_\varepsilon \in C, \phi_\varepsilon \geq 0, \int \phi_\varepsilon = 1\}$ . Consequently,  $\mathbb{D}^0_\varepsilon(a, x) \subset [-\infty, \infty]$  is a nested family of closed subsets of a compact space and so  $\mathbb{D}^0(a, x) \equiv \bigcap_{\varepsilon > 0} \mathbb{D}^0_\varepsilon(a, x)$  may be defined as the nonempty value set of  $a \in \mathcal{MR}$  at  $x$ . If  $a = u \in \mathcal{D}'_+$ , then  $f_\varepsilon(x) = u(x)\phi_\varepsilon(x)$  where  $\phi_\varepsilon \in T_\varepsilon(0)$  and thus  $\mathbb{D}^0_\varepsilon(a, x) \supset \mathcal{D}^0_\varepsilon(u, x)$  and upon intersection we have  $\mathbb{D}^0(u, x) = \mathcal{D}^0(u, x)$ . Also if  $a = \sum_{k=0}^\infty (1/(2k)!)s^k$ , then  $\mathbb{D}^0(a, x) = \{0\}$  when  $x \neq 0$ . It is not clear what  $\mathcal{D}^0(a, 0)$  equals.

A shortcoming of the present notion of value set is that it is not descriptive enough to retrieve the original distribution or regular Mikusinski operator from the collection of value sets. We will not at this time discuss this problem or other obvious problems relating value sets of general distributions or Mikusinski operators to the classical

distributions. However, the present notion of value sets will prove quite useful in the next two sections. We have only developed the properties of value sets which we will need to study local extrema of  $u \in C(\Omega)$  and the maximum principle. The maximum principle was the original motivation for this study. The maximum principle will be a crucial tool in another paper [11] in which monotone methods are used to construct solutions to certain nonlinear problems. This method will be similar to the work of J. Chandra and P. W. Davis [4].

**4. Local extrema of  $u \in C(\Omega)$ .** In this section we use the first and second order derivative sets of  $u$  to test for monotonicity and local extrema. The methods of proof basically follow the classical methods once the operator of differentiation is transposed from  $\mathcal{D}'(\Omega) \rightarrow \mathcal{D}'(\Omega)$  to  $\mathcal{D}(\Omega) \rightarrow \mathcal{D}(\Omega)$ .

**THEOREM 2.** *Let  $u \in L_1(a, b)$ . If  $u$  is increasing (decreasing) almost everywhere, then  $\mathcal{D}(u, x) \subset [0, \infty) (\subset [-\infty, 0])$  almost everywhere.*

*Proof.* Suppose  $u$  is increasing almost everywhere; i.e.,  $0 \leq u(x+h) - u(x)$  for  $h \geq 0$  for almost all  $x \in (a, b)$  and  $h+x \in (a, b)$ . If  $\phi \in T_\varepsilon(x)$  and  $h \leq \varepsilon$ , then as  $h \rightarrow 0$  we have

$$\begin{aligned} 0 &\leq \frac{1}{h} \int_a^b (u(y+h) - u(y))\phi(y) dy \\ &= \int_a^b u(y) \left( -\frac{\phi(y-h) - \phi(y)}{-h} \right) dy \rightarrow \int_a^b u(y)(-\phi'(y)) dy. \end{aligned}$$

Thus for all  $\varepsilon > 0$ ,  $u(-\phi') \geq 0$  and so  $\mathcal{D}(u, x) \subset [0, \infty)$ .

**THEOREM 3.** *Let  $u \in C(\Omega)$  and  $\Omega \subset \mathbb{R}^n$ . If  $x \in \Omega$  gives a local maximum (minimum) of  $u$ , then  $0 \in \mathcal{D}^\alpha(u, x)$  where  $|\alpha| = 1$ .*

*Proof.* If  $0 \notin \mathcal{D}^\alpha(u, x)$  with  $|\alpha| = 1$ ,  $\alpha_i = 1$ , then because  $\mathcal{D}^\alpha(u, x)$  is an interval,  $\mathcal{D}^\alpha(u, x)$  is either contained in  $[d, +\infty)$  or  $[-\infty, -d]$  where  $d > 0$ . Suppose  $0 < r = \min \mathcal{D}^\alpha(u, x)$ . There exists  $\phi_n \in T_{1/n}(x)$  such that  $\int_\Omega u(y)(-1)(\partial/\partial y_i)\phi_n(y) dy \rightarrow r$  as  $n \rightarrow \infty$ . Let  $N$  be such that  $n \geq N$  implies  $\int_\Omega u(y)(-1)(\partial/\partial y_i)\phi_n(y) dy \geq r/2$ . Now as  $|h| \rightarrow 0$ ,  $h = (0, \dots, h_i, \dots, 0)$ ,  $\int_\Omega u(y)(-1)(\phi_n(y-h) - \phi_n(y))/(-h_i) dy \rightarrow \int_\Omega u(y)(-1)(\partial/\partial y_i)\phi_n(y) dy$  and so there exists  $\delta = \delta(n) > 0$  so that  $0 < h_i < \delta(n)$  implies  $\int_\Omega u(y)(-1)(\phi_n(y-h) - \phi_n(y))/(-h_i) dy = \int_\Omega ((u(y+h) - u(y))/h_i)\phi_n(y) dy > r/4$ . Since  $u$  is continuous,  $\int_\Omega ((u(y+h) - u(y))/h_i)\phi_n(y) dy \rightarrow (u(x+h) - u(x))/h_i$  as  $n \rightarrow \infty$ . Because  $u$  has a local maximum at  $x$ ,  $(u(x+h) - u(x))/h_i \leq 0$  for suitably small  $h_i > 0$ . Thus we may choose  $n$  so that  $\int_\Omega ((u(y+h) - u(y))/h_i)\phi_n(y) dy < r/8$ . This is a contradiction to  $r > 0$ .

If  $0 > r = \max \mathcal{D}^\alpha(u, x)$ , then let  $h_i < 0$  and multiply the integrals by  $(-1)$ . This also leads to a contradiction.

Finally we state and prove the main theorem in this section which gives sufficient conditions for extrema. Let  $d_x^k \phi$  represent the  $k$ th order differential at  $x$  of  $\phi$ .

**THEOREM 4.** *Let  $u \in C(\Omega)$ ,  $\Omega \subset \mathbb{R}^n$  and  $h \in \mathbb{R}^n$ . If*

- (i) *for all  $h$  such that  $|h| = 1$  there exist  $\phi_n \in T_{1/n}(x)$  such that as  $n \rightarrow \infty$   $u((d_x^1 \phi_n)h) \uparrow 0 (\downarrow 0)$ ,*
- (ii) *for all  $n$  and for all  $\phi \in T_{1/n}(x)$  there exist  $m > 0$  such that  $u((d_x^2 \phi)h) \leq -m (\geq m)$ , then  $x \in \Omega$  gives a local maximum (minimum) of  $u$ .*

*Proof.* Apply Taylor's theorem to  $\phi_n$  of Assumption (i) to obtain

$$\phi_n(x+h) - \phi_n(x) = (d_x^1 \phi_n)h + \frac{1}{2!}(d_x^2 \phi_n)h + \frac{1}{3!}(d_{x+\theta h}^3 \phi)h$$

where  $\theta h = (\theta_1 h_1, \dots, \theta_n h_n)$   $0 \leq \theta_i \leq 1$  and  $i = 1, \dots, n$ . Since  $u$  is continuous, as  $n \rightarrow \infty$

$u(\phi_n(x+h) - \phi_n(x)) \rightarrow u(x-h) - u(x)$ . Thus it suffices to show that  $u(\phi_n(x+h) - \phi_n(x))$ ,  $n \geq n_0$ , are nonpositive. So consider

$$u(\phi_n(x+h) - \phi_n(x)) = |h|u\left(\frac{d_x^1 \phi_n}{|h|} \frac{h}{|h|}\right) + \frac{1}{2!}|h|^2 u\left(\frac{d_x^2 \phi_n}{|h|} \frac{h}{|h|}\right) + \frac{1}{3!}|h|^3 u\left(\frac{d_x^3 \phi_n}{|h|} \frac{h}{|h|}\right).$$

First, consider the third term on the right side.

$$u\left(\frac{d_x^3 \phi_n}{|h|} \frac{h}{|h|}\right) = \sum_{|\alpha|=3} c_\alpha u\left((-\mathcal{D})^\alpha \phi_n(x+\theta h) \frac{h}{|h|}\right)$$

where  $c_\alpha$  are the coefficients of the third order differential. Let  $\psi \in C_c^\infty(S_{1/(n_0-1)}(x))$  such that  $\psi \equiv 1$  on  $S_{1/n_0}(x)$  and  $n_0$  is to be chosen. Since  $u$  is continuous we may choose  $n_0$  so that for all  $\varepsilon > 0$   $n \geq n_0$  implies

$$\left| \int_\Omega \psi(y) u(y) (-\mathcal{D})^\alpha \phi_n(y+\theta h) dy - u(x) \int_\Omega \psi(y) (-\mathcal{D})^\alpha \phi_n(y+\theta h) dy \right| < \varepsilon.$$

Also note

$$u((-\mathcal{D})^\alpha \phi_n(x+\theta h)) = \int_\Omega u(y) (-\mathcal{D})^\alpha \phi_n(y+\theta h) dy = \int_\Omega \psi(y) u(y) (-\mathcal{D})^\alpha \phi_n(y+\theta h) dy.$$

Since  $\mathcal{D}^\alpha \psi$  is continuous and  $\phi_n \in T_{1/n}(x)$ , as  $n \rightarrow \infty$ ,  $\int_\Omega \psi(y) (-\mathcal{D})^\alpha \phi_n(y+\theta h) dy = \int_\Omega \mathcal{D}^\alpha \psi(y-\theta h) \phi_n(y) dy \rightarrow \mathcal{D}^\alpha \psi(x-\theta h)$ . Thus for all bounded  $|h|$ ,  $u((-\mathcal{D})^\alpha \phi_n(x+\theta h))$  is bounded by a constant which is independent of  $n \geq n_0$ . So choose  $\delta > 0$  so that  $|h| < \delta$  implies

$$\frac{1}{3!} u\left(\frac{d_x^3 \phi_n}{|h|} \frac{h}{|h|}\right) |h| \leq \frac{m}{4}$$

where  $m$  is from assumption (ii).

Second, consider the middle term on the right side. By assumption (ii)  $u((d_x^2 \phi_n)/|h|) \leq -m$  and so we have

$$u(\phi_n(x+h) - \phi_n(x)) \leq |h|u\left(\frac{d_x^1 \phi_n}{|h|} \frac{h}{|h|}\right) + |h|^2 \left(\frac{-m}{2}\right) + |h|^2 \frac{m}{4}.$$

Third, for each fixed direction  $h/|h|$  choose  $\phi_n$  as given in assumption (i) so that  $u((d_x^1 \phi_n)/|h|) \uparrow 0$  as  $n \rightarrow \infty$ . Consequently,  $u(\phi_n(x+h) - \phi_n(x)) \leq (-m/4)|h|^2 \leq 0$  for  $n \geq n_0$ .

**COROLLARY 1.** Let  $u \in C(\Omega)$  and  $\Omega \subset \mathbb{R}^2$ . If for all  $|h|=1$  there exist  $\phi_n \in T_{1/n}(x)$  such that

- (i)  $(u(-\phi_{n_{x_1}})h_1 + u(-\phi_{n_{x_2}})h_2) \uparrow 0$  as  $n \rightarrow \infty$ ,
- (ii)  $u(\phi_{n_{x_1 x_1}})u(\phi_{n_{x_2 x_2}}) - u(\phi_{n_{x_1 x_2}})^2 \geq m' > 0$  for all  $\phi_n$  and either  $u(\phi_{n_{x_1 x_1}}) < 0 (> 0)$  or  $u(\phi_{n_{x_2 x_2}}) < 0 (> 0)$ ,

then  $x$  gives a local maximum (minimum) of  $u$ .

**COROLLARY 2.** Let  $u \in C(a, b)$  and  $\Omega = (a, b) \subset \mathbb{R}^1$ . If there exist  $\phi_n \in T_{1/n}(x)$  such that

- (i)  $u(-\phi'_n) \uparrow 0$  as  $n \rightarrow \infty$ ,
- (ii)  $u(\phi''_n) \leq -m < 0 (\geq m > 0)$  for all  $\phi_n$ ,

then  $x$  gives a local maximum (minimum) of  $u$ .

**COROLLARY 3.** Let  $u \in C(a, b)$  and  $\Omega = (a, b) \subset \mathbb{R}^1$ . If  $x$  gives a local maximum (minimum), then there is a sequence  $\phi_n \in T_{1/n}(x)$  such that

- (i)  $u(-\phi'_n) \uparrow 0$  as  $n \rightarrow \infty$  ( $\downarrow 0$  as  $n \rightarrow \infty$ )
- (ii) either  $u(\phi''_n) \leq 0$  or  $u(\phi''_n) \downarrow 0$  as  $n \rightarrow \infty$  ( $\geq 0$  or  $\uparrow 0$  as  $n \rightarrow \infty$ ).

*Proof.* Theorem 3 gives a sequence  $\phi_n \in T_{1/n}(x)$  such that  $u(-\phi'_n) \uparrow 0$  as  $n \rightarrow \infty$ . Suppose of all such sequences  $u(\phi''_n)$  is not less than or equal to zero. Then there is a sequence  $\phi_n$  such that  $u(\phi''_n) > 0$ . Either  $u(\phi''_n)$  has a subsequence greater than some  $m > 0$  or  $u(\phi''_n)$  has a subsequence that goes to zero. By Theorem 4 the former case yields that  $x$  gives a local minimum which is a contradiction. Note if  $u = \text{constant}$ ,  $u'' = 0$  and  $u(\phi''_n)$  must go to zero.

The next two examples illustrate Corollaries 2 and 1.

*Examples.* 1. Let  $\Omega = \mathbb{R}$  and define  $u(x) = x + 1$  when  $x \leq 0$  and  $u(x) = -2x + 1$  when  $x \geq 0$ .  $\mathcal{D}(u, 0) = [-2, 1]$  and  $\mathcal{D}^2(u, 0) = [-\infty, 0]$ . Thus we must attempt to apply Corollary 3. Let  $\gamma$  be such that  $\int_{-1}^{\gamma} \delta_1(y) dy = 2/3$  where  $\delta_1(y) \equiv (1/c_1) e^{-1/(1-y^2)}$  with  $c_1 = \int_{-1}^1 e^{-1/(1-y^2)} dy$ . Also let  $\delta_n(y) \equiv n\delta_1(ny)$  and  $\phi_n(y) \equiv \delta_n(y + (\gamma/n))$ .

$$u(-\phi'_n(y)) = u'(\phi_n(y)) = \int_{-\infty}^0 \phi_n(y) dy - 2 \int_0^{\infty} \phi_n(y) dy = 0 \uparrow 0 \quad \text{as } n \rightarrow \infty.$$

$$\begin{aligned} u(\phi''_n(y)) &= -u'(\phi'_n(y)) = -\int_{-\infty}^0 \phi'_n(y) dy + 2 \int_0^{\infty} \phi'_n(y) dy \\ &= -3\phi_n(0) = \frac{-3n}{c_1} e^{-1/(1-\gamma^2)} \rightarrow -\infty \end{aligned}$$

as  $n \rightarrow \infty$ . Thus the two conditions of Corollary 2 hold and we may conclude that  $x = 0$  gives a local maximum of  $u$ .

2. Let  $\Omega = (-1, 1) \times (-1, 1) \subset \mathbb{R}^2$  and define  $u(x_1, x_2)$  as follows

$$u(x_1, x_2) \equiv \begin{cases} 1 - x_1 - x_2, & x_1, x_2 \geq 0, \\ 1 - x_1 + x_2, & x_1 \geq 0 \geq x_2, \\ 1 + x_1 + x_2, & 0 \geq x_1, x_2, \\ 1 + x_1 - x_2, & x_2 \geq 0 \geq x_1. \end{cases}$$

$\mathcal{D}^\alpha(u, 0) = [-1, 1]$  when  $|\alpha| = 1$  and  $\mathcal{D}^\alpha(u, 0) = [-\infty, 0]$  when  $|\alpha| = 2$ . Therefore we need to attempt an application of Corollary 1. Let  $\phi_n = \delta_n(x_1, x_2) = (n/c_2) e^{-1/(1-(x_1^2+x_2^2)n^2)}$  where  $c_2 = \int_{\Omega} e^{-1/(1-(x_1^2+x_2^2))} dx_1 dx_2$ . When  $|\alpha| = 1$ ,  $u((-\mathcal{D})^\alpha \phi_n) = \mathcal{D}^\alpha u(\phi_n) = \int_{\Omega} (\partial u / \partial x_i)(y) \phi_n(y) dy = 0 \uparrow 0$  as  $n \rightarrow \infty$  and so  $\sup_{|h|=1} (u(-\phi_{nx_1})h_1 + u(-\phi_{nx_2})h_2) = 0 \uparrow 0$  as  $n \rightarrow \infty$ .

$$\begin{aligned} u(\phi_{nx_1x_1}) &= u_{x_1}(-\phi_{nx_1}) = \int_{-1}^1 \int_{-1}^0 1(-\phi_{nx_1}) dx_1 dx_2 + \int_{-1}^1 \int_0^1 (-1)\phi_{nx_1} dx_1 dx_2 \\ &= 2 \int_{-1}^0 \int_0^1 \phi_{nx_1} dx_1 dx_2 \\ &= 2 \int_1^1 \phi_n(1, x_2) - \phi_n(0, x_2) dx_2 = -2 \int_1^1 \phi(0, x_2) dx_2 \\ &= -2 \frac{c_1}{c_2} < 0, \end{aligned}$$

$$u(\phi_{nx_2x_2}) = -2 \frac{c_1}{c_2} < 0,$$

$$\begin{aligned} u(\phi_{nx_1x_2}) &= u_{x_2}(-\phi_{nx_1}) = \int_{-1}^0 \int_{-1}^1 -\phi_{nx_1} dx_1 dx_2 + \int_0^1 \int_{-1}^1 (-1)(-\phi_{nx_1}) dx_1 dx_2 \\ &= 2 \int_0^1 \int_{-1}^1 \phi_{nx_2} dx_1 dx_2 = 2 \int_0^1 \phi_n(1, x_2) - \phi_n(-1, x_2) dx_2 = 0. \end{aligned}$$

Thus  $u(\phi_{nx_1x_1})u(\phi_{nx_2x_2}) - (u(\phi_{nx_1x_2}))^2 = 4(c_1^2/c_2^2) > 0$  and  $u(\phi_{nx_1x_1}) = -2(c_1/c_2) < 0$  and therefore we may conclude that  $(0, 0)$  gives a local maximum for  $u$ .

**5. A maximum principle.** In this section we consider the differential equation  $Lu \equiv (pu')' + gu' + qu = f$  where  $p, g \in L_\infty(a, b)$ ,  $0 < m \leq p(x)$  for all  $x \in (a, b)$ ,  $q = Q'$ ,  $f = F'$  with  $Q, F \in L_2(a, b)$ . Recall that  $qu$  is defined as the derivative of  $Q(x)u(x) - \int_a^x Q(y)u'(y) dy$  when  $u \in H_1(a, b) \equiv \{u \in L_2(a, b) | u' \in L_2(a, b)\} \subset C([a, b])$ .

DEFINITION. Let  $p, g, q, f$  be as above.  $u \in H_1(a, b)$  is a *weak solution* to  $Lu = f$  if and only if for all  $\psi \in \mathcal{D}(a, b) \equiv C_c^\infty(a, b)$   $(Lu)(\psi) = f(\psi)$  i.e.

$$\int_a^b [p(y)u'(y)(-\psi(y)) + g(y)u(y)\psi(y) + Q(y)u(y)(-\psi'(y)) - Q(y)u'(y)\psi(y)] dy = \int_a^b F(y)(-\psi'(y)) dy.$$

We prove the strong maximum principle as defined below for the above operator  $L$  when  $Q \in L_\infty(a, b)$ . This generalizes the classical strong maximum principle when  $u \in C^2(a, b)$ ,  $p, p', g, q \in C(a, b)$ . The proof of the classical version may be found in M. Protter and H. Weinberger [7]. The proofs of this section closely follow those given for the classical results in [7] once, as in the previous section, the operation of differentiation has been transposed from  $\mathcal{D}'(a, b) \rightarrow \mathcal{D}'(a, b)$  to  $\mathcal{D}(a, b) \rightarrow \mathcal{D}(a, b)$ .

A weak maximum principle as defined below for elliptic operators of more than one variable when  $Q \in L_n(\Omega)$  with  $\Omega \subset \mathbb{R}^n$  has been proved in G. Stampacchia [10]. At the end of this section examples are given that illustrate the importance of the assumptions on  $q$ . In particular, in one variable  $Q$  must be in  $L_\infty(a, b)$  in order that the strong maximum principle holds.

DEFINITION. *Strong maximum principle.* If  $Lu = f$ , for all  $x \in (a, b)$   $\mathcal{D}^0(q, x) \subset [-\infty, 0]$ ,  $\mathcal{D}^0(f, x) \subset [0, \infty]$ , and  $u \neq \text{constant}$ , then  $\sup_{y \in (a, b)} \{0, u(y)\} > u(x)$  for all  $x \in (a, b)$ .

DEFINITION. *Weak maximum principle.* If  $Lu = f$ , for all  $x \in \Omega$   $\mathcal{D}^0(q, x) \subset [-\infty, 0]$ ,  $\mathcal{D}^0(f, x) \subset [0, \infty]$ , and  $u \neq \text{constant}$ , then  $\sup_{y \in (a, b)} \{0, u(y)\} \leq \max\{u(a), u(b)\}$ .

The maximum principle has applications to the question of uniqueness for the linear problem and to certain nonlinear problems. In particular, one often wishes to use fixed point theorems which involve self maps or perhaps to use monotone methods in order to obtain existence or construction of solutions to nonlinear problems. For examples of both, consult Courant and Hilbert volume two [5], or, in the case of monotone methods, see [4] or [11].

The next theorem is perhaps the simplest maximum principle. In the classical case it is trivial. In all that follows  $L$  will be as above with the additional restriction that  $Q \in L_\infty(a, b)$  and  $\mathcal{D}^1(p, x) \subset [-K, \infty]$ ,  $K < \infty$ ,  $\forall x \in (a, b)$ .

THEOREM 5. If  $u \in H_1(a, b)$ ,  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  for all  $x \in (a, b)$  and  $\mathcal{D}^0(Lu, x) \subset [m(x), \infty]$  where  $m(x) > 0$  for all  $x \in (a, b)$ , then  $u$  cannot have a nonnegative maximum in the interior of  $[a, b]$ .

*Proof.* Since  $u' \in L_2(a, b)$ , then by Hölder's inequality  $u \in C[a, b]$ . Thus  $u$  has a maximum which is attained at some  $x \in [a, b]$ . Assume  $x \in (a, b)$ . By Corollary 3 of Theorem 4 there exists  $\phi_n \in T_{1/n}(x)$  such that as  $n \rightarrow \infty$ ,  $u(-\phi'_n) \uparrow 0$  and  $u(\phi''_n) \leq 0$  or  $u(\phi''_n) \downarrow 0$ .

Note we may assume  $q \equiv 0$ . This follows from  $D^0(-qu, x) \subset [0, \infty]$  when  $0 \leq u(x) = \sup_{(a, b)} u$ ,  $u$  is continuous at  $x$  and  $D^0(q, x) \subset [-\infty, 0]$ . Hence  $D^0(Lu - qu, x) \subset [m(x), \infty]$ . For the moment, let  $p$  and  $q$  be continuous at  $x$  and consider  $Lu(\phi_n) = pu'(-\phi'_n) + gu(\phi_n)$ . By a proof similar to property 7 of Theorem 1 and  $p(x) > 0$ , there exists  $N_1 > 0$  such that when  $n \geq N_1$ ,  $pu'(-\phi'_n) \leq m(x)/8$ . By property 8 of Theorem 1,

$gu'(\phi_n) \rightarrow 0$  and thus there exist  $N_2 > 0$  such that when  $n \geq N_2$ ,  $gu'(\phi_n) \leq m(x)/8$ . Thus, when  $n \geq \max\{N_1, N_2\}$ ,  $Lu(\phi_n) \leq m(x)/4$ . This is a contradiction and so  $x = a$  or  $x = b$ .

If  $p$  or  $g$  are not continuous at  $x$ , since  $u$  is a weak solution  $pu' + \int_0^x gu' = F + \text{constant}$ . Thus  $p(x+)u'(x+) - p(x-)u'(x-) = F(x+) - F(x-)$ . Since  $\mathcal{D}^0(Lu, x) \subset [m(x), \infty]$  and  $m(x) > 0$ ,  $F(x+) - F(x-) > 0$ . Since  $u(x) = \sup_{(a,b)} u$ ,  $u'(x+) \leq 0$  and  $u'(x-) \geq 0$ . Since  $p > 0$ , this yields a contradiction.

The main theorem of this section will now be stated and proved by contradicting the above theorem.

**THEOREM 6.** *Let  $u \in H_1(a, b)$  be a solution of  $Lu = f$ . If  $\mathcal{D}^0(f, x) \subset [0, \infty]$  for all  $x \in (a, b)$ ,  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  for all  $x \in (a, b)$  and  $u$  assumes a nonnegative maximum,  $M$ , at  $c \in (a, b)$ , then  $u \equiv M$ .*

*Proof.* By the remarks in the second paragraph of the proof of the previous theorem, it suffices to demonstrate this theorem when  $q = 0$ .

As in the classical case we define  $z(x) = e^{\alpha(x-c)} - 1$  on  $(a_1, d)$  where  $u(d) < u(c)$  and  $c \in (a_1, d)$ . We shall find  $\alpha > 0$  so that  $Lz(\phi) \geq m(x) > 0$  where  $m(x)$  is to be defined and  $\phi \in T_\varepsilon(x)$ . Since  $\mathcal{D}^1(p, x) \subset [-K, \infty]$  and  $p(x) \geq m_1 > 0$ ,  $pz'(-\phi') = (p(y)\alpha^2 e^{\alpha(y-c)})(\phi) - (p(y))(\alpha e^{\alpha(y-c)}\phi') \geq (m_1/2)\alpha^2 e^{\alpha(x-c)} - 2K\alpha e^{\alpha(x-c)}$ , for suitable  $\varepsilon > 0$ . Since  $g \in L_\infty(a, b)$ ,  $|g(y)| \leq K_1 < \infty$ . Thus  $gz'(\phi) \geq -2K_1\alpha e^{\alpha(x-c)}$  for suitable  $\varepsilon > 0$ . Consequently,  $Lz(\phi) \geq (m_1/2)\alpha^2 e^{\alpha(x-c)} - 2K\alpha e^{\alpha(x-c)} - 2K_1\alpha e^{\alpha(x-c)}$  for suitable  $\varepsilon > 0$ . Therefore, we may choose  $\alpha$  large enough so that  $\mathcal{D}^0(Lz, x) \subset [m(x), \infty]$  where  $m(x) \equiv \alpha e^{\alpha(x-c)}[(m_1/2)\alpha - 2K - 2K_1] > 0$ .

Let  $0 < \gamma < (M - u(d))/z(d)$  and consider  $u + \gamma z$ . Since  $\mathcal{D}^0(Lu, x) \subset [0, \infty]$ ,  $\mathcal{D}^0(L(u + \gamma z), x) = \mathcal{D}^0(Lu + \gamma Lz, x) \subset \mathcal{D}^0(Lu, x) + \mathcal{D}^0(\gamma Lz, x) \subset [\gamma m, \infty]$  for all  $x \in (a_1, d)$ . Thus we may apply Theorem 5 to  $u + \gamma z$  on  $(a_1, d)$ . Note  $(u + \gamma z)(d) = u(d) + \gamma z(d) < u(d) + ((M - u(d))/z(d))z(d) = M$  and  $u(c) + \gamma z(c) = M$ . Thus  $\max(u + \gamma z) \geq M = (u + \gamma z)(c)$  and consequently the maximum of  $u + \gamma z$  on  $[a_1, d]$  is attained in the interior of  $[a_1, d]$ . This is a contradiction and so  $u(d) = u(c)$ .

*Remark.* The restrictions of  $Q \in L_\infty(a, b)$  and  $\mathcal{D}^1(p, x) \subset [-K, \infty]$  were needed in order to construct  $z$  such that  $\mathcal{D}^0(Lz, x) \subset [m(x), 0]$  with  $m(x) > 0$ . If we consider the example given by  $Lz \equiv ((2 - H(x-0))z)'$  and let  $\phi_n$  be an even delta sequence about  $x = 0$ , then we obtain  $Lz(\phi_n) \equiv -z'(0)\phi_n(0) \rightarrow -\infty$  as  $n \rightarrow \infty$  unless  $z'(0) = 0$ . Thus we are not able to construct  $z$  such that  $\mathcal{D}^0(Lz, x) \subset [m(x), \infty]$  for  $m(x) > 0$  and for all  $x \in (a, b)$ .

**COROLLARY 1.** *Let  $u \in H_1(a, b)$  be a solution of  $Lu = f$ . If  $\mathcal{D}^0(f, x) \subset [0, \infty]$  for all  $x \in (a, b)$ ,  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  for all  $x \in (a, b)$ ,  $u(a) \leq 0$  and  $u(b) \leq 0$ , then either  $u(x) < 0$  for  $x \in (a, b)$  or  $u \equiv 0$ .*

*Proof.* If  $u \neq 0$ , then by Theorem 6  $u$  cannot have a nonnegative maximum in the interior of  $[a, b]$ . Since the maximum must be at the boundary,  $u(a) \leq 0$  and  $u(b) \leq 0$ , then  $u(x) \leq M \leq 0$ . If  $u(x) = 0$  for  $x \in (a, b)$ , then by Theorem 6  $u \equiv 0$ . Thus either  $u(x) < 0$  for  $x \in (a, b)$  or  $u \equiv 0$ .

**COROLLARY 2.** *Let  $u \in H_1(a, b)$  be a solution of  $Lu = f$ ,  $u(a)$  and  $u(b)$  given. The solution is unique when  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  for all  $x \in (a, b)$ .*

**COROLLARY 3.** *Let  $u \in H_1(a, b)$  be a solution of  $Lu = f$ ,  $u(a)$  and  $u(b)$  given. If  $u(x) \geq 0$  for all  $x \in (a, b)$ ,  $u(a)$  and  $u(b) \leq N$ ,  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  for all  $x \in (a, b)$  and  $\mathcal{D}^0(f - qN, x) \subset [0, \infty]$  for all  $x \in (a, b)$ , then  $0 \leq u(x) \leq N$  for all  $x \in (a, b)$ . An added constraint on  $f$  on  $\mathcal{D}^0(f, x) \subset [-\infty, 0]$  implies that  $u(x) \geq 0$ .*

*Proof.* Consider  $u - N$ .  $L(u - N) = f - qN$ . Apply Corollary 1 with  $u$  replaced by  $u - N$  and  $f$  replaced by  $f - qN$  to conclude that  $u(x) - N \leq 0$  for all  $x \in (a, b)$ . In order to show that  $\mathcal{D}^0(f, x) \subset [-\infty, 0]$  implies  $u(x) \geq 0$ , apply Corollary 1 with  $u$  replaced by  $-u$  and  $f$  replaced by  $-f$ .

We finally give the examples which illustrate the importance of the assumptions on  $q$ .

*Example 1.*  $\mathcal{D}^0(q, x) \subset [-\infty, 0]$  is a necessary restriction.  $Lu \equiv u'' + u$ ,  $u(-\pi) = 0 = u(\pi)$  is the simplest example. Consider the differential equation  $u'' + 2\delta u = 0$ ,  $u(-1) = 0 = u(1)$  on the interval  $[-1, 1]$ . In this case  $q = +2\delta$  and  $\mathcal{D}^0(q, x) \subset [0, \infty]$  if  $x \neq 0$  and  $\mathcal{D}^0(q, 0) = [0, \infty]$ . Let  $u(0)$  be any constant and define

$$u(x) = \begin{cases} +u(0)x + u(0), & -1 \leq x \leq 0, \\ -u(0)x + u(0), & 0 \leq x \leq 1. \end{cases}$$

This is a family of solutions to the given equation which depends on the choice of  $u(0)$ . Thus the solution to the equation  $u'' + 2\delta u = f$ ,  $u(-1)$  and  $u(1)$  given is not unique. See Fig. 1 for the graph of  $u$ .

*Example 2.* This example illustrates  $Q \in L_2(-1, 1) \setminus L_\infty(-1, 1)$  such that the weak maximum principle holds but the strong maximum principle does not hold. Let  $Lu \equiv u'' - 1/9x^{-4/3}u$ ,  $Q \equiv 1/3x^{-1/3}$  and  $u \equiv 2 - x^{2/3} - (1 - |x|)^{2/3}$ . Then  $Lu = f = 2/9(1 - |x|)^{-4/3} + 1/9x^{-4/3}(x^{2/3} + (1 - |x|)^{2/3})$  and so  $\mathcal{D}^0(f, x) \subset [0, \infty]$  for all  $x \in (-1, 1)$  and  $\mathcal{D}^0(f, 0) = \{\infty\}$ . The graph of  $u$  is given in Fig. 2 which clearly shows the desired result.

*Example 3.* At the beginning of these sections we mentioned that G. Stampacchia had proved a weak maximum principle for elliptic operators in more than one variable with  $u \in H_1(\Omega)$  and  $Q \in L_n(\Omega)$ . One reason why this may not in general work for  $\Omega = (a, b)$  is that the product  $qu$  is not defined when  $Q \in L_1(a, b)$ . One must then place further restrictions on  $Q$  or  $u$ . This example illustrates  $u \in L_2(a, b)$  and  $u' \in L_1(a, b)$ , such that neither maximum principle holds for a  $Q \in L_1(a, b) \setminus L_2(a, b)$ . Let  $Lu \equiv u'' - \frac{2}{9}|x|^{-5/3}u$ ,  $Q \equiv \frac{1}{3}|x|^{-2/3}$  and  $u \equiv 1 - |x|^{1/3}$ . Then  $Lu = f = \frac{2}{9}|x|^{4/3}$  and so  $\mathcal{D}^0(f, x) \subset [0, \infty]$  for all  $x \in (-1, 1)$ . Note that  $-\frac{2}{9}x^{-5/3}u$  may be defined as the derivative of an element of  $L_1(a, b)$ . The graph of  $u$  is given in Fig. 3 which shows that neither maximum principle holds.

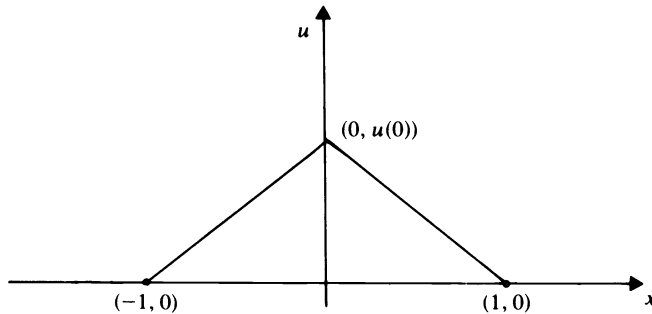


FIG. 1

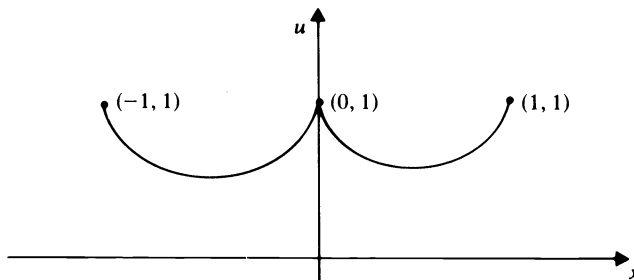


FIG. 2

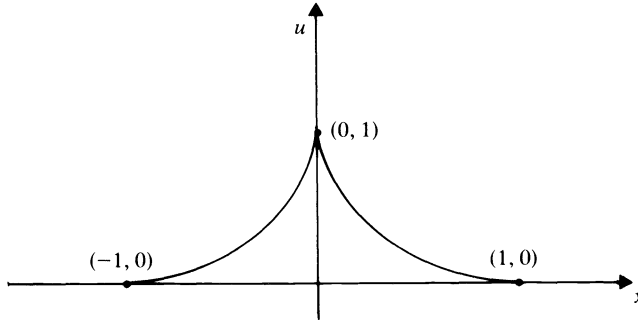


FIG. 3

## REFERENCES

- [1] P. ANTOSIK, *The symmetric value of a distribution at a point*, unpublished, 1974.
- [2] P. ANTOSIK, J. MIKUSINSKI AND R. SIKORSKI, *Theory of Distributions*, The Sequential Scientific Publishers, Warsaw, 1973.
- [3] T. K. BOEHME, *The support of Mikusinski operators*, Trans. Amer. Math. Soc., 196 (1973), pp. 319–334.
- [4] J. CHANDRA AND P. W. DAVIS, *A monotone method for quasilinear boundary value problems*, Arch. Rational Mech. Anal., 54 (1974), pp. 257–266.
- [5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Wiley-Interscience, New York, 1962.
- [6] S. LOJASIEWICZ, *Sur la valeur et le limite d'une distribution dan un point*, Studia Math., 16 (1957), pp. 1–36.
- [7] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
- [8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [9] L. SCHWARTZ, *Theorie des distributions*, Hermann, Paris, 1966.
- [10] G. STAMPACCHIA, *Equations Elliptiques de Second Ordre à Coefficients Discontinus*, Les Presses de l'Université de Montreal, Montreal, 1966.
- [11] R. E. WHITE, *Weak solution of  $(p(x)u'(x))' + g(x)u'(x) + qu(x) = f$  with  $g, f \in H_{-1}[a, b]$ ,  $0 < p(x) \in L_{\infty}[a, b]$ ,  $g[x] \in L_{\infty}[a, b]$  and  $u \in H_1[a, b]$* , this Journal, 10 (1979), pp. 1313–1326.



## THE ZEROS OF THE ODD AND EVEN PARTS OF A HURWITZ POLYNOMIAL\*

AARON FIALKOW†

**Abstract.** This paper is a study of some properties of the zero distributions of Hurwitz polynomials and also of polynomials  $H_r(x)$  of the form  $H_r(x) = x^r F(x) + G(x)$ , where  $F(x)$ ,  $G(x)$  are independent of  $r$ . In particular, it is proved that if a sequence of strict Hurwitz polynomials  $Q_r(z)$  satisfies  $Q_r(z)Q_r(-z) = H_r(x)$ ,  $x = -z^2$ , then an increasing number of the zeros of both the odd part and even part of  $Q_r(z)$  become arbitrarily small and arbitrarily great as  $r \rightarrow \infty$ . This theorem has application in network theory as the guarantor of the validity of a new synthesis method for realizing quite general filters by means of a transformerless, inductance, capacitance ladder network terminated in resistance (LC-R ladders). A special case of the theorem has been used to validate a method of even part synthesis of an arbitrary impedance by means of at most four LC-R ladder networks.

**1. Introduction.** The importance of electric filter networks in the solution of many engineering problems is well known. As a result, there is a considerable literature concerned with the theory and design of filter networks. The actual realization of a preassigned filter function is generally by means of a lossless ladder network terminated in the resistance which represents the load. For arbitrary distributions of passbands and stopbands, this ladder realization frequently requires the unavoidable use of transformers.

A recent investigation [2] has studied the impedance of transformerless, inductance, capacitance two-ports terminated in resistance (LC-R networks). This analysis led to a synthesis criterion and algorithm for a wide class of impedances. Their realization is in the form of an LC-R ladder having considerable element economy relative to the degree of the impedance.

Based on this synthesis procedure, quite general filter functions may be realized without resort to transformers, provided the filter function has either a reflection zero or an attenuation pole at the origin. A statement of this and similar results appears in [3]. The principal theorem may be paraphrased as follows: *Any real rational function having only pure imaginary poles is the characteristic function of a filter which is realizable as an LC-R ladder, provided that the function has a zero or pole at the origin of sufficiently high order.* While the actual synthesis depends upon the methods developed in [2], the guarantee that the procedure must be successful also rests upon some purely mathematical results which are the subject of the present paper.

These results are concerned with the distribution of the zeros of certain polynomials, the distribution of the zeros of their odd and even parts, and their formal structure. Theorem 1 describes the zero distribution of polynomials of the form  $x^r F(x) + G(x)$ , with polynomials  $F(x)$ ,  $G(x)$  independent of  $r$ , for large values of  $r$ . The second theorem and its corollaries finds conditions on the zero distribution of a sequence of Hurwitz polynomials  $Q_r(z)$  sufficient to guarantee that a preassigned number of the (real)  $x$  zeros, ( $x = -z^2$ ) of the odd and even parts of  $Q_r(z)$  be less than (or greater than) a preassigned magnitude. All of these developments culminate in Theorem 3 which is required for the filter application described above. This theorem states that the conclusion of Theorem 2 is true if  $Q_r(z)Q_r(-z) = x^r F(x) + G(x)$ , with  $r$  sufficiently great.

---

\* Received by the editors November 16, 1978.

† Department of Mathematics, Polytechnic Institute of New York, Brooklyn, New York 11201.

**2. The zeros of  $x^r F(x) + G(x)$ .** The required performance specifications of a filter dictate the choice of a characteristic function. The characteristic function of a filter is a real rational function  $S(z)/T(z)$ . It is well known [4, (12, 14)], that a strict Hurwitz polynomial  $Q(z)$  exists such that

$$(1) \quad Q(z)Q(-z) = S(z)S(-z) + T(z)T(-z).$$

Based on [2], the realization of this filter by a transformerless ladder<sup>1</sup> requires that the characteristic function have a zero or pole at  $z = 0$  and further depends upon a guarantee that enough zeros of the odd and even parts of  $Q(z)$  are sufficiently small (or great). There is no obvious, direct relationship between this last condition and the corresponding structure of the characteristic function. A principal goal of the subsequent mathematical analysis is to establish such a connection.

Suppose that the polynomials  $S(z)$ ,  $T(z)$  are prescribed, except for an arbitrary factor  $z^r$  in one of them. Then (1) takes the form

$$Q(z)Q(-z) = x^r F(x) + G(x), \quad x = -z^2,$$

where  $F(x)$  and  $G(x)$  are independent of  $r$ . In this section, we investigate the zeros of this polynomial in  $x$ . The subsequent sections relate the location of these  $x$  zeros to the zeros of the odd and even parts of  $Q(z)$ .

**THEOREM 1.** *Let  $F(x)$ ,  $G(x)$ , with  $F(0)G(0) \neq 0$ , be relatively prime polynomials of degree  $r_1$  and  $r_2$  respectively, and*

$$(2) \quad H_r(x) = x^r F(x) + G(x).$$

*Let  $\Phi$  be any angle of the complex  $x$ -plane, with vertex at the origin, defined by*

$$(3) \quad \Phi: \theta \leq \arg x \leq \theta + \phi, \quad \phi < \frac{\pi}{1 + \min(r_1, r_2)}.$$

*Then, for  $\epsilon, 0 < \epsilon < 1$ , and every positive integer  $c$ , an integer  $r_0(\epsilon, c)$  exists such that for all  $r \geq r_0$ ,  $H_r(x)$  has at least  $c$  zeros which lie within the ring,  $1 - \epsilon \leq |x| \leq 1 + \epsilon$ , but outside the angle  $\Phi$ .*

*Proof.* We first prove a number of lemmas.

**LEMMA 1.** *Let  $G(x) = (x - a)^u G_1(x)$ ,  $G_1(a) \neq 0$  and  $|a| < 1$ . Also let  $C_R$  be the circle  $|x - a| = R$ , where  $R$  is any positive number for which  $C_R$  lies inside the unit circle  $|x| = 1$ , and contains no zeros of  $F(x)G_1(x)$ . Then, for all sufficiently great  $r$ ,  $H_r(x)$  has exactly  $u$  zeros within the circle  $C_R$ .*

*Proof.* Since  $F(x)$ ,  $G(x)$  are relatively prime,  $F(a) \neq 0$ . Consequently, for all sufficiently small  $R$ ,  $C_R$  lies in  $|x| = 1$  and no zeros of  $F(x)G_1(x)$  lie in or on  $C_R$ . Then  $G_1(x)/(F(x))$  is analytic in and on  $C_R$  and not zero there. Applying the maximum modulus theorem to its reciprocal, we conclude that  $G_1(x)/(F(x))$  assumes its minimum modulus  $K$  on the boundary of  $C_R$ . Consequently, on the boundary,

$$\left| \frac{G(x)}{F(x)} \right|_{C_R} = \left| \frac{G_1(x)}{F(x)} \right|_{C_R} \cdot R^u \geq KR^u.$$

Also,

$$|x^r|_{C_R} \leq (|a| + R)^r.$$

---

<sup>1</sup> For any ladder, the zeros of  $T(z)$  must all be pure imaginary [1; p. 184]. This fact plays no role in this paper.

Since  $|a| + R < 1$ , it follows that for all sufficiently great  $r$ ,

$$KR^u > (|a| + R)^r.$$

It follows from Rouché's theorem that  $G(x)/(F(x))$  and  $G(x)/(F(x)) + x^r$  have the same number of zeros in  $C_R$ ; that is,  $H_r(x)$  has exactly  $u$  zeros which approach  $x = a$  as  $r \rightarrow \infty$ .  $\square$

LEMMA 2. Let  $F(x) = (x - b)^v F_1(x)$ ,  $F_1(b) \neq 0$  and  $|b| > 1$ . Also let  $C_{R'}$  be the circle  $|x - b| = R'$ , where  $R'$  is any positive number for which  $C_{R'}$  lies outside the unit circle  $|x| = 1$  and contains no zeros of  $F_1(x)G(x)$ . Then, for all sufficiently great  $r$ ,  $H_r(x)$  has exactly  $v$  zeros within the circle  $C_{R'}$ .

Proof. For sufficiently great  $r$ ,  $r + r_1 - r_2 > 0$ . We assume this is so. Then by means of the transformation,  $y = 1/x$ , from (2) we obtain

$$(4) \quad H'_r(y) \equiv y^{r+r_1} H_r\left(\frac{1}{y}\right) = y^{r'} G'(y) + F'(y),$$

where  $r' = r + r_1 - r_2$ , and  $F'(y)$ ,  $G'(y)$  are polynomials in  $y$  are defined by

$$F'(y) = y^{r_1} F\left(\frac{1}{y}\right), \quad G'(y) = y^{r_2} G\left(\frac{1}{y}\right).$$

Clearly  $F'(0)G'(0) \neq 0$ . Also, since neither 0 nor  $\infty$  is a zero of  $H_r(x)F(x)G(x)$ , the zeros of  $H'_r(y)$ ,  $F'(y)$ ,  $G'(y)$  are also bifinite and the reciprocals of the zeros of  $H_r(x)$ ,  $F(x)$ ,  $G(x)$  respectively. In particular, the factor  $(x - b)^v$  of  $F(x)$  corresponds to the factor  $(by - 1)^v$  of  $F'(y)$ . Also, (4) has the same structure as (2). Hence Lemma 1 may be applied to  $H'_r(y)$  with  $(by - 1)^v$ ,  $F'(y)$  occupying the roles of  $(x - a)^u$ ,  $G(x)$  respectively. When the result is translated from  $y$  to  $x$  variables, we obtain Lemma 2.  $\square$

LEMMA 3. Let the  $a_i$ , of multiplicity  $u_i$ , be those zeros of  $G(x)$  for which  $|a_i| < 1$  and let the  $b_j$ , of multiplicity  $v_j$ , be the zeros of  $F(x)$  for which  $|b_j| > 1$ . Let  $\varepsilon$ ,  $0 < \varepsilon < 1$ , be such that the circular regions

$$C_{a_i}: |x - a_i| \leq \varepsilon, \quad C_{b_j}: |x - b_j| \leq \varepsilon$$

and the ring

$$\mathcal{R}: 1 - \varepsilon \leq |x| \leq 1 + \varepsilon$$

are all disjoint. Then  $\rho_1 > 0$  exists such that for all  $r > \rho_1$ ,  $u_i$  zeros of  $H_r(x)$  are in  $C_{a_i}$ ,  $v_j$  zeros of  $H_r(x)$  are in  $C_{b_j}$  for each  $a_i, b_j$ , and all the remaining zeros of  $H_r(x)$  are in  $\mathcal{R}$ .

Proof. According to Lemma 1 and Lemma 2,  $\rho' > 0$  exists so that for  $r > \rho'$  exactly  $u_i$  zeros of  $H_r(x)$  are in each  $C_{a_i}$  and exactly  $v_j$  zeros are in each  $C_{b_j}$ , and  $De[H_r(x)] = r + r_1$ . Let  $x_1(r)$  be one of the remaining zeros of  $H_r(x)$ . If possible, contrary to Lemma 3, let

$$(5) \quad |x_1(r)| \leq K_1 < 1$$

for an increasing sequence  $\mathcal{S}$  of values of  $r$ . Now a constant  $B > 0$  exists so that  $|F(x_1)| < B$  for all  $x_1$  which lie in the region defined by (5), since a continuous function has bounded modulus on a closed bounded region. Hence

$$(6) \quad |x'_1 F(x_1)| < BK'_1.$$

Since, by hypothesis,  $x_1(r)$ , with  $r > \rho'$ , is outside each  $C_{a_i}$ , a constant  $K > 0$  exists so that, for all  $r > \rho'$

$$(7) \quad |G[x_1(r)]| > K.$$

From (2), (6), (7), for  $r$  belonging to  $\mathcal{S}$  and also  $r > \rho''$ ,

$$0 = |H_r(x_1)| \geq |G(x_1)| - |x_1' F(x_1)| > K - BK_1' > 0.$$

This contradiction proves that (5) cannot be true for  $r > \max(\rho', \rho'')$ . The assumption  $|x_1(r)| \geq K_2 > 1$  may be disproved for values of  $r > \rho'''$  by similar analysis of the zero  $y_1 = 1/x_1$  of  $H_{r'}(y)$ , given by (4). Hence all these remaining zeros  $x_1(r)$  of  $H_r(x)$  eventually enter and remain within  $\mathcal{R}$  for all  $r > \rho_1$ , where  $\rho_1 = \max(\rho', \rho'', \rho''')$ .  $\square$

LEMMA 4. Suppose  $\rho_2 \geq \rho_1$  (of Lemma 3) exists such that, for all  $r > \rho_2$ , the number of zeros of  $H_r(x)$  which are outside an angle  $\Phi_0$  defined by

$$(8) \quad \Phi_0: \theta_0 \leq \arg x \leq \theta_0 + \phi, \quad \phi < \frac{\pi}{k},$$

where  $k$  is a positive integer, does not exceed a fixed integer  $p_0$ . Then the elementary symmetric function  $\Sigma_k$  of order  $k$  of the zeros of  $H_r(x)$ , where  $r > \rho_2$ , cannot be equal to zero.

*Proof.* For  $r > \rho_1$ , in accordance with Lemma 3, each of the circles  $C_{a_i}, C_{b_i}$  contains a fixed number of zeros of  $H_r(x)$ , while the remaining zeros are in  $\mathcal{R}$ . Let  $p_1$  of the zeros of  $H_r(x)$  be in those  $C_{a_i}, C_{b_i}$  which have points in common with  $\Phi_0$ . Further let  $\Psi_1$  be the intersection of  $\Phi_0$  and  $\mathcal{R}$ . Therefore, if  $n$  and  $p$  are the number of zeros of  $H_r(x)$  which are respective points of  $\Psi_1$  and its complement,

$$(9) \quad p \leq p_0 + p_1, \quad n = r + r_1 - p,$$

since  $De[H_r(x)] = r + r_1$ .

An elementary symmetric function of order  $h$  of some variables is the sum of all possible products, without repetition, of  $h$  of these variables. Denote the elementary function of order  $h$  of the  $p$  zeros of  $H_r(x)$  outside  $\Psi_1$  by  $\Sigma'_h$  and the corresponding function of the  $n$  zeros inside  $\Psi_1$  by  $\Sigma''_h$ . Since, by Lemma 3, all the zeros of  $H_r(x)$ , for all  $r$ , lie in a closed bounded region, and the number of them which enter  $\Sigma'_h$  is bounded by (9) for all  $r$ , the  $|\Sigma'_h|$  are each bounded, independent of  $r$ . Thus there exists a number  $B > 1$  so that, for all  $r$ ,

$$(10) \quad \begin{aligned} |\Sigma'_h| &< B, & h = 1, 2, \dots, p, \\ \Sigma'_0 &\equiv 1, & \Sigma'_h \equiv 0, & h > p. \end{aligned}$$

Let  $x_j$  be one of the remaining  $n$  zeros of  $H_r(x)$  given by

$$(11) \quad x_j = r_j e^{i\theta_j}, \quad 1 - \varepsilon \leq r_j \leq 1 + \varepsilon, \quad \theta_0 \leq \theta_j \leq \theta_0 + \phi.$$

If

$$x_{(h)} = r_{(h)} e^{i\theta_{(h)}}$$

is a typical product in  $\Sigma''_h$ , from (11),

$$(12) \quad (1 - \varepsilon)^h \leq r_{(h)} \leq (1 + \varepsilon)^h,$$

$$(13) \quad h\theta_0 \leq \theta_{(h)} \leq h\theta_0 + h\phi.$$

Hence each product term is a point in the intersection  $\Psi_h$  of the ring  $\mathcal{R}_h$ , defined by (12), and the angle  $\Phi_h$ , defined by (13). In view of (8), the vertex opening of  $\Phi_h$  is less than  $\pi$  if  $h \leq k$ . The average or centroid  $\bar{x}_{(h)}$  of all these product terms  $x_{(h)}$  lies within the convex hull of  $\Psi_h$ , and cannot be zero if  $h \leq k$ . Consequently, from (12) and (13),

$$(14) \quad (1 - \varepsilon)^h \cos \frac{h\phi}{2} \leq |\bar{x}_{(h)}| \leq (1 + \varepsilon)^h, \quad 1 \leq h \leq k,$$

where, since  $h\phi/2 < \pi/2$ , all quantities are positive. The total number of  $x_{(h)}$  equals

$$\binom{n}{h} \equiv \frac{n(n-1) \cdots (n-h+1)}{h!}$$

and  $\Sigma_h''$  is the sum of all these  $x_{(h)}$ . Therefore

$$\Sigma_h'' = \binom{n}{h} \bar{x}_{(h)}$$

and, from (14),

$$(15) \quad \binom{n}{h} (1-\varepsilon)^h \cos \frac{h\phi}{2} \leq |\Sigma_h''| \leq \binom{n}{h} (1+\varepsilon)^h, \quad 0 \leq h \leq k.$$

(The inequalities for  $h=0$  are a consequence of  $\Sigma_0'' = 1$ .)

Now the elementary symmetric function  $\Sigma_k$  of all the zeros of  $H_r(x)$  obeys

$$(16) \quad \Sigma_k = \sum_{h=0}^k (\Sigma_h'' \Sigma'_{k-h}) = \Sigma_k'' + \sum_{h=0}^{k-1} (\Sigma_h'' \Sigma'_{k-h}).$$

From (10) and (15),

$$(17) \quad \left| \sum_{h=0}^{k-1} (\Sigma_h'' \Sigma'_{k-h}) \right| < B \sum_{h=0}^{k-1} \binom{n}{h} (1+\varepsilon)^h,$$

$$|\Sigma_k''| \geq \binom{n}{k} (1-\varepsilon)^k \cos \frac{k\phi}{2}.$$

Now  $\lim_{n \rightarrow \infty} \binom{n}{k} / \binom{n}{h} = \infty$  if  $h < k$ . Also, (9) implies that  $n \rightarrow \infty$  iff  $r \rightarrow \infty$ . Finally, from (8),  $h\phi/2 < \pi/2$  for  $h = 0, 1, 2, \dots, k$ . Consequently, a constant  $\rho_2 \geq \rho_1$  exists so that, for all  $r > \rho_2$ ,

$$|\Sigma_k''| > \left| \sum_{h=0}^{k-1} (\Sigma_h'' \cdot \Sigma'_{k-h}) \right|$$

after using (17). But, from (16) and this equation,

$$|\Sigma_k| \geq |\Sigma_k''| - \left| \sum_{h=0}^{k-1} (\Sigma_h'' \Sigma'_{k-h}) \right| > 0. \quad \square$$

We now complete the proof of Theorem 1. According to Lemma 3, if  $r > \rho_1$ , a fixed number of the zeros of  $H_r(x)$  lie within the circles  $C_{a_r}, C_{b_r}$ , while all the remaining zeros are in  $\mathcal{R}$ . Now the coefficients of a polynomial are multiples of the elementary symmetric functions of its zeros. For  $H_r(x)$ , since the term involving  $x^{r-1}$  is missing if  $r > r_2 + 1$ , it follows that  $\Sigma_{(r_1+1)} = 0$  for all sufficiently great  $r$ . Then Lemma 4 proves that as  $r \rightarrow \infty$ , the number of zeros of  $H_r(x)$  inside  $\mathcal{R}$  but outside any angle

$$\Phi_1: \theta \leq \arg x \leq \theta + \phi_1, \quad \phi_1 < \frac{\pi}{1+r_1},$$

cannot remain bounded.

In a similar manner, if we proceed with  $H_{r'}(y)$ , defined by (4), we find that as  $r' \rightarrow \infty$ , the number of  $y$  zeros of  $H_{r'}(y)$  inside the ring  $\mathcal{R}'$  (which is the transform of  $\mathcal{R}$  by  $y = 1/x$ ) but outside any angle in the  $y$  plane

$$\Phi_2: \theta \leq \arg y \leq \theta + \phi_2, \quad \phi_2 < \frac{\pi}{1+r_2},$$

cannot remain bounded. If this result is restated in terms of  $x$  quantities, it is identical with the conclusion of the preceding paragraph, except that  $\phi_1$  is replaced by  $\phi_2$ . The results concerning  $\Phi_1$  and  $\Phi_2$  taken together are equivalent to the conclusion of Theorem 1.  $\square$

**3. Hurwitz polynomials with zeros in prescribed regions.** Let  $Q_r(z)$ ,  $r = 1, 2, 3, \dots$ , be a sequence of strict Hurwitz polynomials. By definition of  $Q_r(z)$ , its zeros lie in the interior of the left half plane. In this section, we show that if  $m_r$  of the zeros of  $Q_r(z)$ , with  $m_r \rightarrow \infty$  as  $r \rightarrow \infty$ , are further restricted to certain subregions of the left half plane, then an increasing number of the zeros of the even and odd parts of  $Q_r(z)$  tend to 0 and  $\infty$  as  $r \rightarrow \infty$ .

THEOREM 2. *Let  $\Psi'$  be the sector of the  $z$  plane*

$$(18) \quad \Psi': |z| \leq R_0, \quad R_0 > 0; \quad \pi - \theta_0 \leq \arg z \leq \pi + \theta_0, \quad 0 \leq \theta_0 < \frac{\pi}{2}.$$

Let  $Q_r(z)$ ,  $r = 1, 2, 3, \dots$ , be an infinite sequence of real, strict Hurwitz polynomials written as

$$(19) \quad Q_r(z) = f_r(x) + zg_r(x), \quad x = -z^2.$$

Suppose  $Q_r(z)$  has  $m_r$  zeros in  $\Psi'$  and that

$$(20) \quad \lim_{r \rightarrow \infty} m_r = \infty.$$

Let positive constant  $x_0$  and positive integer  $c$  be prescribed. Then a constant  $r_0(x_0, c)$  exists such that, for all  $r > r_0$ ,  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros  $x_j$  such that  $0 < x_j < x_0$ .

*Proof.* As  $z = i\omega$  varies along the  $i$ -axis from 0 to  $\infty$ , the corresponding  $x$ , obeying  $x = \omega^2$ , varies on the real axis from 0 to  $\infty$ . From (19), it follows that  $Q_r(i\omega)$  is pure imaginary at a zero of  $f_r(\omega^2)$  and is real at a zero of  $g_r(\omega^2)$ . Thus the angle  $\phi[Q_r(i\omega)]$  of the complex number  $Q_r(i\omega)$  is an odd or even multiple of  $\pi/2$  if and only if  $\omega^2$  is a zero of  $f_r(x)$  or  $g_r(x)$  respectively. Consequently, we study the function  $\phi[Q_r(i\omega)]$ .

Write  $Q_r(z)$  in factored form as

$$Q_r(z) = K \prod_h (z + c_h) \cdot \prod_j (z + R_j e^{i\theta_j})(z + R_j e^{-i\theta_j}),$$

where  $K, c_h, R_j$  are real positive numbers and  $0 < \theta_j < \pi/2$ . The angle function  $\phi$  satisfies the equation

$$(21) \quad \phi[Q_r(i\omega)] = \sum_h \phi(i\omega + c_h) + \sum_j [\phi(i\omega + R_j e^{i\theta_j}) + \phi(i\omega + R_j e^{-i\theta_j})].$$

If we write

$$(22) \quad \tan \gamma_h = \frac{\omega}{c_h}, \quad \tan \beta_{j1} = \frac{\omega + R_j \sin \theta_j}{R_j \cos \theta_j}, \quad \tan \beta_{j2} = \frac{\omega - R_j \sin \theta_j}{R_j \cos \theta_j},$$

then (21) becomes

$$(23) \quad \phi[Q_r(i\omega)] = \sum_h \gamma_h + \sum_j (\beta_{j1} + \beta_{j2}).$$

Also define  $\alpha_0, \alpha_j$  by

$$(24) \quad \tan \alpha_0 = \frac{\omega}{R_0}, \quad \tan \alpha_j = \frac{\omega}{R_j}.$$

Then, using (22) and (24),

$$(25) \quad \tan (\beta_{j1} + \beta_{j2}) = \frac{2\omega R_j \cos \theta_j}{R_j^2 - \omega^2} = \cos \theta_j \cdot \tan 2\alpha_j.$$

As the real variable  $\omega$  increases from 0 to  $\infty$ , each  $\gamma_h$  and each  $\alpha_j$  increase monotonically from 0 to  $\pi/2$ , while the  $\theta_j$  are constants in the range  $0 < \theta_j < \pi/2$ . Thus, using (25),  $(\beta_{j1} + \beta_{j2})$  increases monotonically with  $\omega$ , varying from an initial value 0 to a final value  $\pi$ . Consequently,  $\phi[Q_r(i\omega)]$  is a monotonically increasing function of  $\omega$ , with each of the angles,  $\gamma_h, (\beta_{j1} + \beta_{j2})$ , in (23) making a positive contribution to the total.

According to the hypothesis of the theorem,  $m_r$  of the zeros of  $Q_r(z)$  lie in the sector  $\Psi'$ . Let  $m'_r$  of these zeros be negative real and  $m''_r$  be pairs of conjugate complex zeros, with

$$(26) \quad m'_r + 2m''_r = m_r.$$

For each of the  $m'_r$  negative real zeros,  $c_h \leq R_0$ . Then, from (22) and (24),  $\gamma_h \geq \alpha_0$ . Also, for each of the  $m''_r$  pairs of complex zeros,  $R_j \leq R_0$  and  $\theta_j \leq \theta_0$ . Consequently, from (24),  $\alpha_j \geq \alpha_0$  and  $\cos \theta_j \geq \cos \theta_0$ . Then

$$\tan 2\alpha_j \geq \tan 2\alpha_0 > 0 \quad \text{if } \alpha_0 \leq \frac{\pi}{4}, \quad \alpha_j \geq \frac{\pi}{4},$$

$$0 > \tan 2\alpha_j \geq \tan 2\alpha_0 \quad \text{if } \alpha_0 > \frac{\pi}{4}, \quad \frac{\pi}{2} > \alpha_j > \frac{\pi}{4},$$

$$\tan^{-1} [\cos \theta_0 \tan 2\alpha_0] \leq \frac{\pi}{2} \quad \text{and} \quad \tan^{-1} [\cos \theta \tan 2\alpha] > \frac{\pi}{2} \quad \text{if } \alpha_0 \leq \frac{\pi}{4}, \quad \frac{\pi}{2} > \alpha_j > \frac{\pi}{4}.$$

Hence, in all cases, for any of these  $m''_r$  pairs of zeros, we find, after reference to (25), that

$$\beta_{j1} + \beta_{j2} \geq \tan^{-1} [\cos \theta_0 \tan 2\alpha_0].$$

Then, from (23),

$$(27) \quad \phi[Q_r(i\omega)] \geq m'_r \alpha_0 + m''_r \tan^{-1} [\cos \theta_0 \tan 2\alpha_0].$$

Since  $\alpha_0, \theta_0$  are constant, independent of  $r$ , it follows from (20), (26) and (27) that, for any real positive  $\omega$ ,

$$\lim_{r \rightarrow \infty} \phi[Q_r(i\omega)] = \infty.$$

This equation implies that if  $\omega_0 = x_0^{1/2}$ , a constant  $r_0$  exists such that, for  $r > r_0$ ,  $\phi[Q_r(i\omega_0)] > c\pi$ . Then the discussion in the first paragraph of the proof shows that  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros which obey  $0 < x < x_0$ .  $\square$

**COROLLARY 1.** *Let the sequence of strict Hurwitz polynomials  $Q_r(z)$ , defined by (19), have  $m_r$  zeros in the sector exterior  $\Psi''$  defined by*

$$(28) \quad \Psi'': |z| \geq \frac{1}{R_0}, \quad R_0 > 0; \quad \pi - \theta_0 \leq \arg z \leq \pi + \theta_0, \quad 0 \leq \theta_0 < \frac{\pi}{2},$$

where  $m_r$  obeys (20). Then a constant  $r_0$  exists such that for all  $r > r_0$ ,  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros  $x_k$  with  $x_k > x_0$ , for arbitrary choices of  $x_0 > 0$  and positive integer  $c$ .

*Proof.* We consider the effect of the transformation  $w = 1/z$ . Suppose  $De[Q_r(z)] = n$ . Since  $Q_r(z)$  has no zeros at  $z = 0$  or  $z = \infty$ , the zeros of the polynomial

$$Q'_r(w) = w^n Q_r\left(\frac{1}{w}\right)$$

are the reciprocals of the zeros of  $Q_r(z)$ . Also

$$Q'_r(w) = f'(y) + wg'(y), \quad y = -w^2,$$

where

$$f'(y) = y^{n/2} f\left(\frac{1}{y}\right), \quad g'(y) = y^{(n-2)/2} g\left(\frac{1}{y}\right),$$

if  $n$  is even and

$$f'(y) = y^{(n-1)/2} g\left(\frac{1}{y}\right), \quad g'(y) = y^{(n-1)/2} f\left(\frac{1}{y}\right),$$

if  $n$  is odd. Furthermore the transformation maps  $\Psi''(z)$  onto  $\Psi'(w)$ , defined by (18). It follows that  $Q'_r(w)$  satisfies the hypothesis of Theorem 2. Consequently, for any positive  $1/x_0$ ,  $c$ , at least  $c$  zeros  $y_k$  of  $f'_r(y)$  and of  $g'_r(y)$  obey  $y_k < 1/x_0$ . The corresponding zeros  $x_k$  of  $f_r(x)$  and  $g_r(x)$  equal  $1/y_k$  and so satisfy  $x_k > x_0$ .  $\square$

**COROLLARY 2.** Let  $\Psi_0$  be the intersection of a ring and an angle given by

$$(29) \quad \Psi_0: 0 < R_1 \leq |z| \leq R_2 < \infty; \quad \pi - \theta_0 \leq \arg z \leq \pi + \theta_0, \quad 0 \leq \theta_0 < \frac{\pi}{2}.$$

Let  $Q_r(z)$ ,  $r = 1, 2, 3, \dots$ , be strict Hurwitz polynomials (19), with  $m_r$  zeros in the ring section  $\Psi_0$ , where  $m_r$  obeys (20). Let constant  $x_0$ ,  $0 < x_0 < 1$ , and positive integer  $c$  be prescribed. Then a constant  $r_0$  exists such that, for all  $r > r_0$ ,  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros  $x_j$  so that  $0 < x_j < x_0$  and each have at least  $c$  zeros  $x_k$  so that  $x_k > 1/x_0$ .

*Proof.* The ring section  $\Psi_0$  lies in the sector  $\Psi'$  with circular boundary  $|z| = R_2$ , and also lies in the sector exterior  $\Psi''$  with circular boundary  $|z| = R_1$ . Consequently both these regions have at least  $m_r$  zeros of  $Z_r(z)$ , with  $m_r$  obeying (20). This means that the hypotheses of both Theorem 2 and Corollary 1 are satisfied. Thus positive constants  $r'$  and  $r''$  exist so that for  $r > r'$  and  $r > r''$ , the respective conclusions of the theorem and the corollary are true. Then the conclusion of Corollary 2 follows for  $r_0 = \max(r', r'')$ .  $\square$

**4. Hurwitz polynomials  $Q_r(z)$  which satisfy  $Q_r(z)Q_r(-z) = x^r F(x) + G(x)$ ,  $x = -z^2$ .** If the results of §§ 2 and 3 are combined, we can obtain a property of Hurwitz polynomials having the special structure described by (30) below.

**THEOREM 3.** Let  $Q_r(z)$ ,  $r = 1, 2, 3, \dots$ , be an infinite sequence of real strict Hurwitz polynomials

$$Q_r(z) = f_r(x) + zg_r(x), \quad x = -z^2.$$

Suppose that

$$(30) \quad Q_r(z)Q_r(-z) = H_r(x) = x^r F(x) + G(x),$$

where  $F(x)$  and  $G(x)$  are relatively prime polynomials, independent of  $r$ . Let  $x_0$ ,  $0 < x_0 < 1$  and positive integer  $c$  be prescribed. Then a constant  $r_0(x_0, c)$  exists such that for all



$r > r_0$ ,  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros  $x_j$  so that  $0 < x_j < x_0$ , and each have at least  $c$  zeros  $x_k$  so that  $x_k > 1/x_0$ .

*Proof.* The proof depends upon using Theorem 1 to place an increasing number of zeros of  $H_r(x)$  in a suitable ring segment, and then Corollary 2 to obtain information concerning the zeros of  $f_r(x)$  and  $g_r(x)$ . If  $F(x)$  contains a power of  $x$  as a factor, this factor may be absorbed into  $x^r$ . Hence we may assume, without loss of generality, that  $F(0) \neq 0$ . Also  $G(0) \neq 0$ , else  $Q_r(0) = 0$ , which is impossible since  $Q_r(z)$  is strict Hurwitz. These results and (30) prove that the hypothesis of Theorem 1 is satisfied.

Before applying this theorem, we consider the relevant regions of the  $x$  plane and the  $z$  plane. Choose  $\Phi$  to be the angle of opening  $\phi$  satisfying (3) which is positioned so that it is bisected by the positive  $x$ -axis. Under the map,  $x = -z^2$ ,  $\Phi$  corresponds to two vertical angles of opening  $\phi/2$  in the  $z$  plane which are each bisected by the imaginary axis. The region outside these angles in the left half  $z$  plane satisfies

$$(31) \quad \pi - \theta_0 \leq \arg z \leq \pi + \theta_0, \quad \theta_0 = \frac{\pi}{2} - \frac{\phi}{4}.$$

Since, from (3),  $0 < \phi < \pi$ , it follows that  $\pi/4 < \theta_0 < \pi/2$ . Also corresponding to the ring

$$\mathcal{R}: 1 - \varepsilon \leq |x| \leq 1 + \varepsilon, \quad 0 < \varepsilon < 1,$$

is another ring

$$\mathcal{R}': \sqrt{1 - \varepsilon} \leq |z| \leq \sqrt{1 + \varepsilon}.$$

As a consequence of Theorem 1, the number of zeros of  $H_r(x)$  which lie outside  $\Phi$  but inside  $\mathcal{R}$  increases beyond all bounds as  $r \rightarrow \infty$ . Consequently, the number of zeros of  $Q_r(z)$  which lie within the intersection  $\Psi_0$  of the angle (31) and the ring  $\mathcal{R}'$  is unbounded as  $r \rightarrow \infty$ . We identify  $\Psi_0$  with the region defined by (29). Then the hypothesis of Corollary 2 is satisfied. Hence, after choosing the  $x_0$  and  $c$  of Corollary 2, a constant  $r_0$  exists so that, for all  $r > r_0$ ,  $f_r(x)$  and  $g_r(x)$  each have at least  $c$  zeros  $x_j$  with  $0 < x_j < x_0$  and each have at least  $c$  zeros  $x_k$  with  $x_k > 1/x_0$ .  $\square$

We note that if  $F = G = 1$ , then the  $Q_r(z)$  are the Butterworth polynomials. In this case, Theorem 3 specializes to Theorem 20 of [2]. This theorem is used in [2] to devise an even part synthesis of a general impedance, using at most four LC-R ladder networks.

#### REFERENCES

- [1] N. B. DALABANIAN, *Network Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1958.
- [2] A. D. FIALKOW, *Inductance, capacitance networks terminated in resistance*, IEEE Trans. Circuits and Systems, CS-26 (1979), pp. 603-641.
- [3] ———, *Synthesis by LC-R ladder networks*, Proc. International Symposium on Circuits and Systems, Tokyo, Japan, 1979, pp. 310-311.
- [4] R. SAAL AND E. ULBRICH, *On the design of filters by synthesis*, IRE Trans. on Circuit Theory, vol. CT-5 (1958), pp. 294-327.

## PROPERTIES AND APPLICATIONS OF THE RESOLVENT OPERATOR TO A VOLTERRA INTEGRAL EQUATION IN HILBERT SPACE\*

T. KIFFE† AND M. STECHER†

**Abstract.** This paper discusses the existence, uniqueness, and asymptotic properties of solutions to the equation  $u + a^*Au = f$ , where  $A$  is a positive self-adjoint operator on a Hilbert space. These properties are studied via the resolvent operator for this equation. The authors also consider a nonlinear perturbation of the above.

**1. Introduction.** In this paper we will discuss existence and uniqueness of solutions for the equations,

$$(1.1) \quad u(t) + \int_0^t a(t-s)Au(s) ds = f(t), \quad 0 \leq t \leq T,$$

$$(1.2) \quad u(t) + \int_0^t a(t-s)Au(s) ds + \int_0^t a(t-s)Bu(s) ds \ni f(t),$$

where  $A$  is a positive self-adjoint linear operator densely defined on a Hilbert space  $H$ ,  $B$  is a possibly multiple valued maximal monotone operator, and  $a(t)$  is a real valued function. We will also give some results concerning the asymptotic behavior of the solutions to (1.1).

Clément and Nobel [4] have recently considered (1.1) and established existence and uniqueness results under various hypotheses on the forcing term  $f(t)$ . Their technique is essentially that of constructing a resolvent operator for (1.1), and by use of its properties, deducing that (1.1) has solutions for various  $f$ 's. Using a different analysis of the resolvent operator we have been able to extend their existence results in the case where  $X$  is a Hilbert space and  $A$  is self-adjoint, and also derive some asymptotic properties of the solutions.

Friedman and Shinbrot [6] have also considered existence, uniqueness, and the asymptotic behavior of solutions to (1.1). Their approach is to analyze the resolvent operator of (1.1) using Laplace transforms, while Clément and Nohel's and our's is to analyze the associated scalar resolvents. For related results on linear Volterra equations we refer the interested reader to [9], [14].

Equation (1.2) written as

$$(1.3) \quad u(t) + \int_0^t a(t-s)g(u(s)) ds \ni f(t),$$

where  $g$  is an accretive operator has been studied by various authors [1], [2], [5], [7], [13]. All of the above papers basically require that  $f(t)$  be differentiable. Viewing (1.3) as a nonlinear perturbation of (1.1) and using some of our results for (1.1) we have been able to extend the existence results for (1.3) to include some nondifferentiable forcing terms.

Section 2 of this paper contains the statements of the results for (1.1), while their proofs are in § 3. Equation (1.2) is discussed in § 4. Examples which illustrate our results are worked out in § 5.

\* Received by the editors February 22, 1978 and in revised form January 19, 1979.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

Throughout this paper we will use the following notation:

- (1.4)  $H$  denotes a real Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and norm  $|\cdot|$ ,  
 $L^P[0, T; H] = \{f: [0, T] \rightarrow H \mid f \text{ is strongly measurable and } \int_0^T |f(t)|^P dt < \infty\}$ ,  $1 \leq P < \infty$ ,  
 $L^\infty[0, T; H] = \{f: [0, T] \rightarrow H \mid f \text{ is strongly measurable and } \text{ess sup}_{0 \leq t \leq T} |f(t)| < \infty\}$ ,  
 $((\cdot, \cdot))$  and  $\|\cdot\|$  will denote the inner product and norm respectively on  $L^2[0, T; H]$ ,  
 $B(H)$  is the space of bounded linear operations from  $H$  to  $H$  equipped with the operator norm topology,  
 $A$  denotes a positive, linear, self-adjoint operator from  $H$  to  $H$  with dense domain,  
 $\{E_\lambda\}_{\lambda \geq 0}$  will denote the resolution of the identity determined by  $A$ ,  
 $\mathcal{A}$  will be the usual extension of  $A$  from  $H$  to  $L^2[0, T; H]$  and  $D(\mathcal{A}) = \{u \in L^2[0, T; H] \mid Au \in L^2[0, T; H]\}$ .

For the standard results concerning the resolution of the identity we refer the reader to [16].

**2. The linear equation.** The standard approach to solving (1.1) has been to first consider solutions of the resolvent scalar equations

$$(2.1) \quad r(t, \lambda) + \lambda \int_0^t a(t-\tau)r(\tau, \lambda) d\tau = a(t), \quad 0 \leq t \leq T,$$

$$(2.2) \quad s(t, \lambda) + \lambda \int_0^t a(t-\tau)s(\tau, \lambda) d\tau = 1, \quad 0 \leq t \leq T.$$

If we define resolvent operators  $R(t)$  and  $S(t)$  by

$$(2.3) \quad R(t) = \int_0^\infty r(t, \lambda) dE_\lambda,$$

$$(2.4) \quad S(t) = \int_0^\infty s(t, \lambda) dE_\lambda,$$

then the solution of (1.1) can be written in the form

$$(2.5) \quad u(t) = f(t) - \int_0^t R(t-\tau)Af(\tau) d\tau$$

or

$$(2.6) \quad u(t) = S(t)f(0) + \int_0^t S(t-\tau)f'(\tau) d\tau$$

under various hypotheses of  $f$  and  $f'$  [4]. We begin by stating various properties of the resolvent functions which will be used in studying (1.1).

Throughout we will assume that  $a(t)$  is a real-valued function defined for  $0 < t < \infty$  satisfying

$$(2.7) \quad a \in C(0, \infty), a \in L^1_{loc}(0, \infty), a(t) \text{ is positive and nonincreasing} \\ \text{and } \log a(t) \text{ is convex.}$$

LEMMA 1. *Suppose (2.7) is satisfied and let  $r(t, \lambda)$  and  $s(t, \lambda)$  denote the solutions of (2.1) and (2.2) respectively. Then*

- (i)  $r(t, \lambda) \geq 0, s(t, \lambda) \geq 0$  for  $\lambda \geq 0$  and  $t > 0$ ,
- (ii)  $\sup_{\lambda \geq 0} \lambda^\alpha r(t, \lambda) \leq C_\alpha a(t) [\int_0^t a(\tau) d\tau]^{-\alpha}$  for  $0 \leq \alpha \leq 1, 0 < t$ , where  $C_\alpha$  is a constant depending only on  $\alpha$ ,
- (iii)  $\sup_{\lambda \geq 0} \lambda^\alpha r(t, \lambda) \in L^1_{loc}[0, \infty)$  for  $0 \leq \alpha < 1$ ,  
 $\sup_{\lambda \geq 0} \lambda^\alpha r(t, \lambda) \rightarrow 0$  as  $t \rightarrow \infty$  for  $0 < \alpha \leq 1$ , and  
 $\sup_{\lambda \geq 0} r(t, \lambda) \rightarrow 0$  as  $t \rightarrow \infty$  if  $a(t) \rightarrow 0$  as  $t \rightarrow \infty$ ,
- (iv)  $s(t, \lambda) \leq [1 + \lambda \int_0^t a(\tau) d\tau]^{-1}$  for  $t \geq 0, \lambda \geq 0$ .

We remark that in (iii) if  $\alpha = 1$ , then  $\sup_{\lambda \geq 0} \lambda r(t, \lambda) \notin L^1(0, \delta)$  for any  $\delta > 0$ . The proof of Lemma 1 is based on an inequality due to Gripenberg [8, Thm. 1]. The properties of the operators  $R(t)$  and  $S(t)$  are contained in the next two lemmas.

LEMMA 2. *Suppose (1.4) and (2.7) are satisfied and let  $R(t)$  be defined by (2.3). Then*

- (v)  $A^\alpha R(t) \in B(H)$  for  $0 \leq \alpha \leq 1, 0 < t < \infty$ ,
- (vi)  $A^\alpha R(t) \in L^1_{loc}[0, \infty; B(H)]$  for  $0 \leq \alpha < 1$  and if  $a \in L^1(0, \infty)$  then  $\int_0^\infty A^\alpha R(s) ds = \bar{a} A^\alpha [I + \bar{a} A]^{-1}$  where  $\bar{a} = \int_0^\infty a(s) ds$ ,
- (vii) for each  $x \in H, A^\alpha R(t)x$  is a continuous function of  $t$  for  $0 \leq \alpha \leq 1$  and  $0 < t < \infty$ ,
- (viii)  $A^\alpha R(t) \rightarrow 0$  as  $t \rightarrow \infty$  in the operator norm topology on  $B(H)$  for  $0 < \alpha \leq 1$  and  $R(t)x \rightarrow a(\infty)E_0x$  as  $t \rightarrow \infty$  for each  $x \in H$ , where  $a(t) \rightarrow a(\infty)$  as  $t \rightarrow \infty$ ,
- (ix) if in addition to (2.7) we assume that  $a(t) \rightarrow 0$  as  $t \rightarrow \infty$ , then  $R(t) \rightarrow 0$  as  $t \rightarrow \infty$  in the operator norm topology on  $B(H)$ ,
- (x) if in addition to (1.4) and (2.7) we assume that  $\sigma(A) \subseteq [\lambda_0, \infty)$  for some  $\lambda_0 > 0$ , then also  $R(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

Regarding (vi) it is not known if  $AR(t) \in L^1_{loc}[0, \infty; B(H)]$  under hypotheses (1.4) and (2.7). In (viii) the projection  $E_0$  can be characterized as the projection of  $H$  onto the kernel of  $A$  if zero is an eigenvalue of  $A$  and as the zero projection if zero is not an eigenvalue cf. [16, p. 319]; and in (x),  $\sigma(A)$  denotes the spectrum of  $A$ .

LEMMA 3. *Suppose (1.4) and (2.7) are satisfied and let  $S(t)$  be defined by (2.4). Then*

- (xi)  $S(t) \in B(H)$  for  $0 < t < \infty$  and  $S(t)$  is a continuous function of  $t$  in the operator norm topology on  $B(H)$ ,
- (xii)  $S(t) \rightarrow [I + \bar{a} A]^{-1}$  as  $t \rightarrow \infty$  in the operator norm topology on  $B(H)$  if  $\in L^1(0, \infty)$  and  $\bar{a} = \int_0^\infty a(s) ds$ ; and  $S(t) \rightarrow E_0$  if  $a \notin L^1(0, \infty)$ .

Our next result is concerned with the existence and uniqueness of solutions to (1.1).

THEOREM 1. *Suppose (1.4) and (2.7) are satisfied and let  $f, A^{1+\alpha}f \in L^P[0, T; H]$  for some  $0 < \alpha \leq 1$  and  $1 \leq P \leq \infty$ . Then the function  $u(t)$  defined by (2.5) satisfies  $u(t) \in D(A)$  a.e.  $0 \leq t \leq T; u(t), Au(t) \in L^P[0, T; H]$ , and  $u(t)$  satisfies (1.1). Moreover  $u(t)$  is the unique function having these properties.*

We remark that  $f, A^{1+\alpha}f \in L^P[0, T; H]$  implies that  $Af \in L^P[0, T; H]$  since  $A$  being self-adjoint implies  $|Ax|^2 \leq |x|^2 + |A^{1+\alpha}x|^2$  for  $x \in D(A^{1+\alpha})$ . Theorem 1 provides the existence and uniqueness of a strong solution to (1.1). We can define a weak solution  $u(t)$  of (1.1) as follows.

DEFINITION 1. *A function  $u(t)$  is a weak solution of (1.1) if there exist sequences  $\{u_n(t)\}, \{f_n(t)\}$  where each  $f_n \in L^P[0, T; H]$  and each  $u_n(t)$  is a strong solution of (1.1) with  $f$  replaced by  $f_n$  such that  $f_n \rightarrow f$  and  $u_n \rightarrow u$  in  $L^P[0, T; H]$  as  $n \rightarrow \infty$ .*

It will be clear from the proof of Theorem 1 that (1.1) has a unique weak solution if  $f, A^\alpha f \in L^P[0, T; H]$  for some  $\alpha, 0 < \alpha \leq 1$ .

Clément and Nohel [4] have considered (1.1) in a Banach space  $X$  with the assumption that  $A$  generates a strongly continuous contraction semi-group. They proved existence and uniqueness of a strong solution to (1.1) if  $f, Af, A^2f \in L^p[0, T; X]$  and a weak solution if  $f, Af \in L^p[0, T; X]$ . Our Theorem 1 extends their results by relaxing the restrictions on  $f$  at the expense of assuming that  $X$  is a Hilbert space and  $A$  is self-adjoint. They also have shown the existence and uniqueness of a weak solution  $u(t)$  to (1.1) if  $f \in W^{1,1}[0, T; X]$ . Under our restrictions on  $X$  and  $A$  their representation for the weak solution becomes (2.6) when  $S(t)$  is given by (2.4). We will use this representation for weak solutions when we consider the asymptotic behavior of solutions to (1.1).

One of the purposes of this paper is to study the nonlinear equation (1.2). Our proof of existence and uniqueness of solutions to (1.2) rests heavily on the properties of maximal monotone operators in a Hilbert space and on the properties of solutions to (1.1). For this reason we have considered (1.1) in a Hilbert space setting. Theorem 1 has a significant extension if  $p = 2$ , and this extension, which was proved in [12], is stated below:

**THEOREM 2.** *Suppose  $a(t)$  is of positive type, i.e.,  $a \in L^1_{\text{loc}}[0, \infty)$  and  $\text{Re } \hat{a}(s) \geq 0$  for  $\text{Re } s \geq 0$  where  $\hat{a}(s) = \int_0^\infty e^{-st} a(t) dt$ . Then*

(xiii) *if  $f, Af \in L^2[0, T; H]$ , (1.1) has a unique strong solution  $u(t) \in L^2[0, T; H]$ ,*

(xiv) *if  $f \in L^2[0, T; H]$ , (1.1) has a unique weak solution  $u(t) \in L^2[0, T; H]$  satisfying*

$$(2.8) \quad u(t) + A \left( \int_0^t a(t-s)u(s) ds \right) = f(t) \quad a.e. \ 0 \leq t \leq T.$$

We refer the reader to [15] for properties of functions of positive type. If  $a(t)$  satisfies (2.7) it is well known that  $a(t)$  is of positive type so Theorem 2 handles more general kernels than Theorem 1. Also it should be noted that kernels of the form  $a(t) = e^{-bt} \cos(\gamma t)$  are of positive type for  $b \geq 0$  but do not satisfy hypothesis  $H_4$  of [4].

Theorem 3 below summarizes the asymptotic properties of weak solutions to (1.1).

**THEOREM 3.** *Let (1.4) and (2.7) be satisfied and suppose  $u(t)$  is a weak solution of (1.1):*

(xv) *if  $a \in L^1(0, \infty)$ ,  $(A^\alpha f)(t) \rightarrow x$  as  $t \rightarrow \infty$  and either  $f \in L^\infty[0, \infty; H]$  or  $A^\alpha f \in L^\infty[0, \infty; H]$  for some  $\alpha$ ,  $0 < \alpha \leq 1$ , then  $f(t) - u(t) \rightarrow \bar{a}A^{1-\alpha}[I + \bar{a}A]^{-1}x$ , as  $t \rightarrow \infty$ , where  $\bar{a} = \int_0^\infty a(s) ds$ ,*

(xvi) *if  $a(0) < \infty$ ,  $Af \in L^1[0, \infty; H]$  and either  $A^\alpha f \in L^1[0, \infty; H]$  for some  $\alpha$ ,  $0 \leq \alpha < 1$  or  $\sigma(A) \subseteq [\lambda_0, \infty)$  for some  $\lambda_0 > 0$ , then  $u(t) - f(t) \rightarrow 0$  as  $t \rightarrow \infty$ ,*

(xvii) *if  $a(0) < \infty$ ,  $a(t) \rightarrow 0$  as  $t \rightarrow \infty$  and  $Af \in L^1[0, \infty; H]$  then  $u(t) - f(t) \rightarrow 0$  as  $t \rightarrow \infty$ ,*

(xviii) *if  $A^\alpha f \in L^\infty[0, \infty; H]$  for some  $0 < \alpha \leq 1$  and  $a \in L^1(0, \infty)$ , or if  $A^\alpha f \in L^\infty[0, \infty; H] \cap L^1[0, \infty; H]$  then  $u - f \in L^\infty[0, \infty; H]$ ,*

(xix) *if  $f \in W^{1,1}_{\text{loc}}[0, \infty; H]$  and  $f' \in L^1[0, \infty; H]$  then  $u(t) \rightarrow E_0 f(\infty)$  as  $t \rightarrow \infty$  if  $a \notin L^1(0, \infty)$  and  $u(t) \rightarrow [I + \bar{a}A]^{-1}f(\infty)$  as  $t \rightarrow \infty$  if  $a \in L^1(0, \infty)$  where  $f(t) \rightarrow f(\infty)$  as  $t \rightarrow \infty$ .*

We remark that in the second part of (xviii) it is not assumed that  $a \in L^1(0, \infty)$ .

Friedman and Shinbrot [6] have also considered the existence and asymptotic behavior of solutions to (1.1) in a Banach space setting. Their approach consisted of using Laplace transforms to study (1.1). Our hypotheses on the kernel function are quite different from theirs and we need not assume that  $A$  is invertible.

**3. Proofs.**

*Proof of Lemma 1.* Part (i) is contained in Proposition 1 of [4]. To prove (ii) we first note that if  $r(t, \lambda)$  is defined by (2.1) then  $\lambda r(t, \lambda)$  is the resolvent function associated with the kernel  $\lambda a(t)$ . Hence [8; Thm. 1] gives us the inequality

$$(3.1) \quad 0 < \lambda r(t, \lambda) \leq \lambda a(t) \left[ 1 + \lambda \int_0^t a(s) ds \right]^{-1}, \quad 0 < t < \infty, \quad 0 < \lambda < \infty.$$

Hence we have, for  $0 \leq \alpha \leq 1$ ,

$$(3.2) \quad 0 < \lambda^\alpha r(t, \lambda) \leq a(t) \left[ \lambda^{-\alpha} + \lambda^{1-\alpha} \int_0^t a(s) ds \right]^{-1}.$$

Fixing  $t$  and  $\alpha$  and maximizing the right side of (3.2) we get (ii). Combining (2.7) and (ii) we easily establish (iii). To prove (iv) we first note that

$$(3.3) \quad s(t, \lambda) = 1 - \lambda \int_0^t r(\tau, \lambda) d\tau.$$

By [8, Thm. 1] we have, recalling that  $\lambda r(t, \lambda)$  is the resolvent of  $\lambda a(t)$ , for  $0 < t < \infty$

$$(3.4) \quad \lambda \int_0^t r(\tau, \lambda) d\tau \geq \lambda \left[ \int_0^t a(\tau) d\tau \right] \left[ 1 + \lambda \int_0^t a(\tau) d\tau \right]^{-1}.$$

Combining (3.3) and (3.4) we get (iv). This completes the proof of Lemma 1. The remark following Lemma 1 follows from the observation that (3.4) implies

$$1 \leq \int_0^t \left[ \sup_{\lambda \geq 0} \lambda r(\tau, \lambda) \right] d\tau \quad \text{for any } t > 0.$$

*Proof of Lemma 2.* From well-known results from the theory of self-adjoint operators in a Hilbert space we have

$$(3.5) \quad A^\alpha R(t) = \int_0^\infty \lambda^\alpha r(t, \lambda) dE_\lambda, \quad 0 < t < \infty, \quad 0 \leq \alpha \leq 1.$$

Hence (v) follows directly from (ii) and we have

$$(3.6) \quad \|A^\alpha R(t)\| \leq C_\alpha a(t) \left[ \int_0^t a(s) ds \right]^{-\alpha}.$$

Hence the first part of (vi) follows from (3.6). If  $a \in L^1(0, \infty)$  then

$$\int_0^\infty r(t, \lambda) dt = \left[ \int_0^\infty a(s) ds \right] \left[ 1 + \lambda \int_0^\infty a(s) ds \right]^{-1}$$

and for each  $x \in H$  we have

$$\begin{aligned} \int_0^\infty A^\alpha R(t)x dt &= \int_0^\infty \int_0^\infty \lambda^\alpha r(t, \lambda) dE_\lambda x dt = \int_0^\infty \int_0^\infty \lambda^\alpha r(t, \lambda) dt dE_\lambda x \\ &= \int_0^\infty \frac{\bar{a}\lambda^\alpha}{1 + \bar{a}\lambda} dE_\lambda x = \bar{a}A^\alpha [I + \bar{a}A]^{-1}x. \end{aligned}$$

The change in the order of integration is justified by (ii) and Fubini's theorem. This proves the second part of (vi). Since  $r(t, \lambda)$  is a continuous function of  $t$  for each  $\lambda \geq 0$ , (vii) follows from (ii) and Lebesgue's dominated convergence theorem. The first part of

(viii) follows from (3.6). By (3.1)

$$r(t, \lambda) \rightarrow \begin{cases} a(\infty), & \text{if } \lambda = 0, \\ 0, & \text{if } \lambda > 0, \end{cases} \text{ as } t \rightarrow \infty,$$

so by (ii) and Lebesgue's theorem we get the second part of (viii). Again (ix) follows from (ii) when  $a(t) \rightarrow 0$  as  $t \rightarrow \infty$ . To prove (x) we note that (3.2) implies

$$(3.7) \quad \sup_{\lambda_0 \leq \lambda} r(t, \lambda) \leq a(t) \left[ 1 + \lambda_0 \int_0^t a(s) ds \right]^{-1}, \quad 0 < t < \infty,$$

which immediately implies (x) since now  $R(t) = \int_{\lambda_0}^{\infty} r(t, \lambda) dE_{\lambda}$ .

*Proof of Lemma 3.* By (iv) we have that

$$(3.8) \quad 0 \leq s(t, \lambda) \leq 1, \quad 0 \leq t < \infty, \quad 0 \leq \lambda < \infty.$$

This establishes the first part of (xi). For the second part we have by (3.3) that if  $t_1 < t_2$  then

$$(3.9) \quad |s(t_2, \lambda) - s(t_1, \lambda)| \leq \int_{t_1}^{t_2} \lambda r(\tau, \lambda) d\tau$$

and hence

$$(3.10) \quad \sup_{0 \leq \lambda < \infty} |s(t_2, \lambda) - s(t_1, \lambda)| \leq \int_{t_1}^{t_2} \left[ \sup_{0 \leq \lambda < \infty} \lambda r(\tau, \lambda) \right] d\tau.$$

By (ii) we have  $\sup_{0 \leq \lambda < \infty} \lambda r(t, \lambda) \in L^1(\delta, \infty)$  for any  $\delta > 0$ . Since  $s(t, \lambda)$  is a continuous function of  $t$  for each  $\lambda \geq 0$ , the second part of (xi) follows from (3.10). To prove (xii) we note that by (iv), if  $a(t) \notin L^1(0, \infty)$  then

$$(3.11) \quad s(t, \lambda) \rightarrow \begin{cases} 1 & \text{if } \lambda = 0, \\ 0 & \text{if } \lambda > 0, \end{cases} \text{ as } t \rightarrow \infty$$

and if  $a(t) \in L^1(0, \infty)$ , then it is well-known that

$$(3.12) \quad s(t, \lambda) \rightarrow [1 + \bar{a}\lambda]^{-1} \text{ as } t \rightarrow \infty, \quad 0 \leq \lambda < \infty.$$

By (3.10) we have that the resolvent operator  $S(t)$  converges in the operator norm topology on  $B(H)$  as  $t \rightarrow \infty$  and (3.11), (3.12), and Lebesgue's theorem now imply (xii).

*Proof of Theorem 1.* We wish to show that  $u(t)$  defined by (2.5) is a solution to (1.1) if  $f$  and  $A^{1+\alpha}f$  are both in  $L^p[0, T; H]$ ,  $1 \leq p \leq \infty$ . That  $u$  will then lie in  $L^p[0, T; H]$  follows from (2.8), (vi) and the remark following Theorem 1. We first establish that  $u(t)$  is contained in the domain of  $A$  a.e. Hence we show that  $R*Af$  is in the domain of  $A$ . The calculations below are easily justified by the functional calculus for self-adjoint operators and (vi):

$$(3.13) \quad \begin{aligned} A(R*Af) &= A \int_0^t R(t-s)Af(s) ds = A^{1-\alpha} \int_0^t A^{\alpha}R(t-s)Af(s) ds \\ &= A^{1-\alpha} \int_0^t R(t-s)A^{1+\alpha}f(s) ds = \int_0^t A^{1-\alpha}R(t-s)A^{1+\alpha}f(s) ds. \end{aligned}$$

Thus since  $A^{1+\alpha}R*Af$  exists and  $A$  is a closed operator we must have  $R*Af \in D(A)$

a.e. To see that (2.5) is actually a solution to (1.1) it suffices to show,

$$(3.14) \quad a*(R*Af) = a*f - R*f.$$

This formula is first shown to hold for  $f(t) \equiv x$  and then for

$$f(t) = \begin{cases} 0, & 0 \leq t \leq c, \\ x, & t > c. \end{cases}$$

That it holds for such  $f$ 's follows easily from the functional calculus and (2.1). Linearity and continuity then imply (3.14) for arbitrary  $f$ . The uniqueness of these solutions has been established in [4].

The existence of weak solutions to (1.1) under the hypotheses that  $f$  and  $Af \in L^p[0, T; H]$ ,  $1 \leq p \leq \infty$ , now follows easily from the fact that  $D(A^\alpha)$  is dense in  $D(A^\beta)$  if  $\alpha > \beta$ , (2.5), and (v).

*Proof of Theorem 3.* To prove (xv) with  $A^\alpha f \in L^\infty[0, \infty; H]$  we observe that  $a \in L^1(0, \infty)$  implies  $A^{1-\alpha}R(t) \in L^1[0, \infty; B(H)]$  by (ii) and since  $f(t) - u(t) = \int_0^t A^{1-\alpha}R(t-s)A^\alpha f(s) ds$  the result follows immediately. If  $f \in L^\infty[0, \infty; H]$ , then we can write

$$(3.15) \quad \begin{aligned} & \int_0^t A^{1-\alpha}R(t-s)A^\alpha f(s) ds - \int_0^\infty A^{1-\alpha}R(s)x ds \\ &= \int_0^{t/2} AR(t-s)f(s) ds - \int_0^{t/2} A^{1-\alpha}R(t-s)x ds \\ & \quad + \int_{t/2}^t A^{1-\alpha}R(t-s)[A^\alpha f(s) - x] ds + \int_t^\infty A^{1-\alpha}R(s)x ds. \end{aligned}$$

Since  $AR(s) \in L^1[1, \infty; B(H)]$  by (ii) the result follows from (3.15). To prove (xvi) with  $A^\alpha f \in L^1[0, \infty; H]$  we note that  $a(0) < \infty$  implies  $R(t) \in L^\infty[0, \infty; B(H)]$  and since  $A^{1-\alpha}R(t) \rightarrow 0$  as  $t \rightarrow \infty$  by (viii), (xvi) follows from observing that  $\int_0^t A^{1-\alpha}R(t-s)A^\alpha f(s) ds = \int_0^{t/2} A^{1-\alpha}R(t-s)A^\alpha f(s) ds + \int_{t/2}^t R(t-s)Af(s) ds$ . If  $\sigma(A) \subseteq [\lambda_0, \infty)$  for some  $\lambda_0 > 0$  then  $Af \in L^1[0, \infty; H]$  implies  $A^\alpha f \in L^1[0, \infty; H]$ . To prove (xvii) we note that  $a(0) < \infty$  and  $a(t) \rightarrow 0$  as  $t \rightarrow \infty$  imply that  $R(t) \in L^\infty[0, \infty; B(H)]$  and  $R(t) \rightarrow 0$  as  $t \rightarrow \infty$  by (ix). By (2.5) the result follows.

The first part of (xviii) follows directly from the fact that  $a \in L^1(0, \infty)$  implies  $A^{1-\alpha}R(s) \in L^1[0, \infty; B(H)]$ . To prove the second part of (xviii) we note that  $u(t) - f(t)$  is clearly bounded for  $t \leq 2$ . For  $t > 2$ , write

$$(3.16) \quad f(t) - u(t) = \int_0^{t-1} A^{1-\alpha}R(t-s)A^\alpha f(s) ds + \int_{t-1}^t A^{1-\alpha}R(t-s)A^\alpha f(s) ds.$$

The first term on the right side of (3.16) is bounded since  $A^{1-\alpha}R(s) \in L^\infty[1, \infty; B(H)]$  by (ii) and the second term is bounded by (vi).

To prove (xix) we observe that by [4, Remark 2.3], (1.1) has a unique weak solution given by (2.6). By (xii) we have

$$(3.17) \quad \lim_{t \rightarrow \infty} u(t) = S(\infty)f(0) + S(\infty) \int_0^\infty f'(\tau) d\tau.$$

Hence  $u(t) \rightarrow S(\infty)f(\infty)$  as  $t \rightarrow \infty$  where  $S(t) \rightarrow S(\infty)$  as  $t \rightarrow \infty$ . By (xii)  $S(\infty) = E_0$  if  $a \notin L^1(0, \infty)$  and  $S(\infty) = [I + \bar{a}A]^{-1}$  if  $a \in L^1(0, \infty)$ . This completes the proof of Theorem 3.



**4. A nonlinear perturbation.** Equation (1.3) has been studied by several authors. Their results have been of two types; either  $f \in W^{1,2}[0, T; H]$  and one can differentiate (1.3) cf. [1], [2], [5], [7], [13], or the nonlinear term  $g$  must satisfy either a local boundedness condition [10], or a linear growth condition [11]. By viewing (1.3) as a nonlinear perturbation of (1.1) the present authors have been able to extend the known existence results. Throughout the rest of this section we will assume that  $B$  is a possibly multiple valued maximal monotone operator which satisfies

$$(4.1) \quad |y| \leq c_1|x| + c_2, \quad y \in Bx.$$

$J_\lambda$  will denote  $(I + \lambda B)^{-1}$ , and  $B_\lambda$  will denote the Yosida approximate of  $B$ , i.e.,  $B_\lambda = \lambda^{-1}[I - J_\lambda]$ , cf. [3]. Our next result provides for the existence and uniqueness of solutions to (1.2).

**THEOREM 4.** *Suppose  $a(t)$  is locally absolutely continuous on  $[0, \infty)$ ,  $a'(t)$  is of local bounded variation on  $[0, \infty)$ ,  $a(0) > 0$ , and  $a(t)$  is of positive type. Let  $B$  satisfy (4.1), and let  $A$  satisfy (1.4). If  $f = f_1 + f_2$  where  $f_1 \in D(\mathcal{A})$  and  $f_2 \in W^{1,2}[0, T; H]$ ,  $f_2(0) = 0$ , then equation (1.2) has a unique solution. That is there exists a unique pair of functions  $u(t)$ ,  $w(t)$  such that*

$$(4.2) \quad \begin{aligned} &u, w \in L^2[0, T; H], \\ &u(t) \in D(A) \text{ a.e.}, \quad Au \in L^2[0, T; H], \quad w(t) \in Bu(t) \text{ a.e.}, \\ &u(t) + \int_0^t a(t-s)Au(s) ds + \int_0^t a(t-s)w(s) ds = f(t). \end{aligned}$$

*Proof.* Let  $u_\lambda$  be the unique solution to the following equation,

$$(4.3) \quad u_\lambda + a^*Au_\lambda + a^*B_\lambda u_\lambda = f.$$

That (4.3) has a unique solution  $u_\lambda \in D(\mathcal{A})$  follows from [12, Thm. 2] since  $B_\lambda$  is Lipschitz continuous.

Our next step is to show that the  $u_\lambda$  are uniformly bounded in  $L^2[0, T; H]$ . To this end multiply (4.3) by  $u_\lambda$  and integrate from 0 to  $\delta$  where  $\delta$  satisfies  $c_1\|a\|_{L^1[0,\delta]} < \frac{1}{2}$ . This gives us

$$(4.4) \quad \begin{aligned} \|u_\lambda\|_{L^2[0,\delta]}^2 &\leq ((f, u_\lambda)) + \|a^*B_\lambda u_\lambda\| \|u_\lambda\| \\ &\leq \|f\| \|u_\lambda\| + \|a\|_{L^1(0,\delta)}\{c_1\|u_\lambda\| + c_2\sqrt{\delta}\}\|u_\lambda\|. \end{aligned}$$

Thus

$$(4.5) \quad \|u_\lambda\|_{L^2[0,\delta]} \leq 2\|f\|_{L^2[0,T]} + 2c_2\sqrt{T}\|a\|_{L^1[0,T]},$$

from which we may infer that not only are the norms of  $u_\lambda$  uniformly bounded in  $L^2[0, \delta; H]$  but also in  $L^2[0, T; H]$  since  $\delta$  depends only on  $a(t)$  and  $c_1$  and not on the nonhomogeneous term in (4.3). By picking subsequences of subsequences, if necessary, we may assume the following

$$(4.6) \quad u_\lambda \rightharpoonup u, \quad a^*u_\lambda \rightharpoonup a^*u, \quad Aa^*u_\lambda \rightharpoonup Aa^*u, \quad B_\lambda u_\lambda \rightharpoonup w, \quad a^*B_\lambda u_\lambda \rightharpoonup a^*w.$$

Note that the graph of  $\mathcal{A}$  is closed with respect to weak-weak convergence and once the  $u_\lambda$  are uniformly bounded in  $L^2[0, T; H]$  so are the  $B_\lambda u_\lambda$  by (4.1) and  $|B_\lambda x| \leq |y|$  for any  $y \in Bx$ . Clearly  $u(t)$  and  $w(t)$  satisfy

$$(4.7) \quad u + Aa^*u + a^*w = f.$$

We now need to show that  $w(t) \in Bu(t)$  a.e. and  $u \in D(\mathcal{A})$ . We will first show that

$w(t) \in Bu(t)$  a.e. Rewrite (4.3) with  $\eta$  instead of  $\lambda$ , subtract the two equations, and then multiply by  $Au_\lambda - Au_\eta + B_\lambda u_\lambda - B_\eta u_\eta$  and integrate. Since  $a(t)$  is of positive type we get

$$(4.8) \quad ((u_\lambda - u_\eta, Au_\lambda - Au_\eta)) + ((u_\lambda - u_\eta, B_\lambda u_\lambda - B_\eta u_\eta)) \leq 0,$$

from which we infer by the positivity of  $A$

$$(4.9) \quad ((u_\lambda - u_\eta, B_\lambda u_\lambda - B_\eta u_\eta)) \leq 0,$$

and this in turn gives us

$$(4.10) \quad \overline{\lim}_{\lambda \rightarrow 0} ((u_\lambda, B_\lambda u_\lambda)) \leq ((u, w)).$$

Since  $B_\lambda u_\lambda$  is bounded and equals  $(1/\lambda)(I - J_\lambda)u_\lambda$  we see that  $u_\lambda - J_\lambda u_\lambda \rightarrow 0$  in  $L^2[0, T; H]$ , and since  $B_\lambda J_\lambda u_\lambda \in BJ_\lambda u_\lambda$  we conclude from (4.10) and [3, Prop. 2.5] that  $w(t) \in Bu(t)$  a.e. To see that  $u \in D(\mathcal{A})$  we denote the map  $u \rightarrow u + Aa^*u$  by  $(I + \mathcal{A}\nu)$  and note that  $(I + \mathcal{A}\nu)^{-1}$  exists as a bounded operator on  $L^2[0, T; H]$ , [12]. Moreover  $(I + \mathcal{A}\nu)^{-1}$  maps our  $f$  into  $D(\mathcal{A})$ . Thus multiplying (4.7) by  $(I + \mathcal{A}\nu)^{-1}$  we see that  $u$  satisfies

$$(4.11) \quad u = (I + \mathcal{A}\nu)^{-1}f - (I + \mathcal{A}\nu)^{-1}(a^*w).$$

Moreover  $a^*w \in W^{1,2}[0, T; H]$ ,  $a^*w(0) = 0$ , and  $(I + \mathcal{A}\nu)^{-1}$  takes such functions into  $D(\mathcal{A})$ . This establishes the existence part of the theorem. Uniqueness is proven as in [13].

We remark next that a simple asymptotic result is obtained if the kernel function  $a(t)$  and the nonlinear term  $B$  satisfy

$$(4.12) \quad c_1 \|a\|_{L^1(0,\infty)} < 1, \quad c_2 = 0.$$

If this is true, then (4.3) gives

$$(4.13) \quad \|u_\lambda\|_{L^2(0,\infty;H)} \leq \frac{1}{1 - c_1 \|a\|_{L^1(0,\infty)}} \|f\|_{L^2(0,\infty;H)},$$

from which we may conclude that  $u \in L^2[0, \infty; H]$  and satisfies (4.13) also.

**5. Examples.**

*Example 1.* Let  $H = L^2[0, \pi]$ ,  $A\phi = -d^2\phi/dx^2$  for  $\phi \in H^2[0, \pi]$  and  $\phi$  satisfies the Neumann boundary conditions, that is  $d\phi/dx \in H_0^1[0, \pi]$ . Hence we may write  $\phi$  as

$$(5.1) \quad \phi(x) = \sum_{n=0}^{\infty} \phi_n \cos nx,$$

where  $\sum_{n=1}^{\infty} n^4 |\phi_n|^2 < \infty$ . Let  $a(t) = e^{-ct}$ , then  $\bar{a} = \int_0^{\infty} e^{-ct} dt = 1/c$ . Now suppose  $f(x, t) \in L^2[(0, \pi) \times (0, T)]$  and  $A^\alpha f(\cdot, t) \in L^\infty[0, T; H]$  and that

$$(5.2) \quad \lim_{t \rightarrow \infty} A^\alpha f(\cdot, t) = \psi(x) = \sum_{n=1}^{\infty} \psi_n \cos nx.$$

Then, if  $u(x, t)$  is the solution to (1.1), we have from (xv) that

$$(5.3) \quad \begin{aligned} \lim_{t \rightarrow \infty} f(\cdot, t) - u(\cdot, t) &= \frac{1}{c} A^{1-\alpha} \left( I + \frac{1}{c} A \right)^{-1} \psi \\ &= \sum_{n=1}^{\infty} \frac{n^{2(1-\alpha)}}{c + n^2} \psi_n \cos nx. \end{aligned}$$

We remark that the operator  $A$  does not have an inverse as is needed in the theory developed in [6].

To illustrate Theorem 4 we give the following example.

*Example 2.* Let  $H = L^2(\Omega)$ , where  $\Omega$  is a bounded open subset of  $\mathbb{R}^n$  with smooth boundary. Let  $A = -\Delta$ , with  $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ . Let  $a(t) = \cos t$ . Let

$$B\phi(x) = \operatorname{sgn} \phi(x) = \begin{cases} 1, & \phi(x) > 0, \\ [-1, 1], & \phi(x) = 0, \\ -1, & \phi(x) < 0. \end{cases}$$

Then  $B$  satisfies (4.1). Let  $f(t) = h(t)\psi$  where  $h: [0, T] \rightarrow L^2[0, T]$  and  $\psi \in D(A)$ . Then by Theorem 4 there exist functions  $u$  and  $w$  such that

$$(5.4) \quad u(t) - \int_0^t \cos(t-s)\Delta u(s) ds + \int_0^t \cos(t-s)w(s) ds = h(t)\psi,$$

and  $w(t) \in Bu(t)$  a.e.

#### REFERENCES

- [1] V. BARBU, *On a nonlinear Volterra integral equation on a Hilbert space*, this Journal, 8 (1977), pp. 346–355.
- [2] ———, *Nonlinear Volterra equations in a Hilbert space*, this Journal, 6 (1975), pp. 728–741.
- [3] H. BRÉZIS, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.
- [4] PH. CLÉMENT AND J. A. NOHEL, *Abstract linear and nonlinear Volterra equations preserving positivity*, MRC Technical Summary Report #1716, University of Wisconsin, Madison, 1977.
- [5] M. G. CRANDALL AND J. A. NOHEL, *An abstract functional differential equation and a related nonlinear Volterra equation*, MRC Technical Summary Report #1765, University of Wisconsin, Madison, 1977.
- [6] A. FRIEDMAN AND M. SHINBROT, *Volterra integral equations in Banach space*, Trans. Amer. Math. Soc., 126 (1967), pp. 131–179.
- [7] G. GRIPENBERG, *An existence result for a nonlinear Volterra integral equation in Hilbert space*, this Journal, 9 (1978), pp. 793–805.
- [8] ———, *On positive, nonincreasing resolvents of Volterra equations*, Report-HTKK-MAT-A109, Helsinki University of Technology, 1977.
- [9] K. HANNSGEN, *The resolvent kernel of an integro-differential equation in Hilbert space*, this Journal, 7 (1976), pp. 481–490.
- [10] T. KIFFE AND M. STECHER, *Existence and uniqueness of solutions to abstr. Volterra integral equations*, Proc. Amer. Math. Soc., 68 (1978), 169–175.
- [11] ———,  *$L^2$  solutions of Volterra integral equations*, this Journal, to appear.
- [12] ———, *A characterization of the range of a nonlinear Volterra integral operator*, Nonlinear Equations in Abstract Spaces, Academic Press, 1978, pp. 365–374.
- [13] S. O. LONDON, *On an integral equation in a Hilbert space*, this Journal, 8 (1977), pp. 950–970.
- [14] R. K. MILLER, *Volterra integral equations in a Banach space*, Funckcial. Ekvac., 18 (1975), pp. 163–193.
- [15] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Advances in Math., 22 (1976), pp. 278–303.
- [16] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1974.

## ASYMPTOTIC EXPANSION OF THE HILBERT TRANSFORM\*

R. WONG†

**Abstract.** Asymptotic expansions are obtained for the Hilbert transform

$$H_f(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(t)}{t-x} dt \quad (x \text{ real}),$$

where the bar indicates that the integral is a Cauchy principal value at  $t = x$ . The function  $f(t)$  is locally integrable in  $(-\infty, \infty)$ , continuously differentiable there except possibly at the origin, and decays algebraically at  $\pm\infty$ . Explicit expressions are given for the error terms associated with these expansions. From the explicit expressions, realistic error bounds can be obtained. Two examples are considered to illustrate the use of these results.

**1. Introduction.** Let  $f$  be a locally integrable function on  $(-\infty, \infty)$ . The Hilbert transform of  $f$ , when it exists, is defined by

$$(1.1) \quad H_f(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(t)}{t-x} dt, \quad x \in (-\infty, \infty),$$

where the bar indicates that the integral is a Cauchy principal value at  $t = x$ . This transform plays an important role in the theory of Fourier analysis [12], and also in the study of singular integral equations [8].

For many years, the problem of finding asymptotic expansions of integral transforms has been the subject of intensive study. Among the recent papers dealing with this problem, we mention those of Jones [4], Handelsman and Lew [3], and Olver [11]. As far as we are aware, there is no asymptotic result for the Hilbert transform available in the literature, and it appears to be desirable to obtain such a result: this will be done in the present paper.

We remark that if  $x$  is a complex parameter and  $\text{Im } x > 0$  (or  $\text{Im } x < 0$ ) then the integral in (1.1) is no longer singular, and it has already been considered by Millar [7]. However, Millar's approach fails if  $x$  is real.

By subdividing the range of integration at the origin, we obtain

$$(1.2) \quad H_f(x) = \frac{1}{\pi} \{H_f^-(x) + H_f^+(x)\},$$

where  $H_f^-(x)$  and  $H_f^+(x)$  denote the integrals corresponding to the intervals  $(-\infty, 0)$  and  $(0, \infty)$ , respectively. For definiteness let us restrict  $x$  to be positive. In this case,  $H_f^-(x)$  is simply the Stieltjes transform of  $-f(-t)$ . Asymptotic theory of this transform is fairly complete; see, for instance, the recent article of McClure and Wong [6]. Thus we may confine ourselves to the consideration of the one-sided Hilbert transform

$$(1.3) \quad H_f^+(x) = \int_0^{\infty} \frac{f(t)}{t-x} dt.$$

Throughout this paper we shall assume that the function  $f(t)$  has an asymptotic expansion of the form

$$(1.4) \quad f(t) \sim e^{ict} \sum_{s=0}^{\infty} a_s t^{-s-\alpha}, \quad \text{as } t \rightarrow \infty,$$

\* Received by the editors September 20, 1978 and in revised form January 11, 1979.

† University of Manitoba, Department of Mathematics and Astronomy, Winnipeg, Manitoba, Canada R3T 2N2. This research was supported in part by the National Research Council of Canada under Contract A7359.

where  $0 < \alpha \leq 1$  and  $c$  is a real number.

**2. Main results.** For  $0 < \alpha < 1$  and  $c$  real, we set

$$(2.1) \quad E_{\alpha,c}(x) = \int_0^\infty \frac{e^{ict}}{t^\alpha(t-x)} dt.$$

It can be proved by using contour integration that if  $c \geq 0$  then

$$(2.2) \quad E_{\alpha,c}(x) = \frac{e^{icx}}{x^\alpha} [e^{-i\alpha\pi} \Gamma(1-\alpha) \Gamma(\alpha, icx) + i\pi].$$

Here  $\Gamma(\alpha, z)$  is the incomplete gamma function

$$\Gamma(\alpha, z) = \int_z^\infty t^{\alpha-1} e^{-t} dt \quad (|\arg z| < \pi),$$

whose asymptotic expansion, complete with error bounds, is given in [9, pp. 110–111]. Since  $\overline{E_{\alpha,c}(x)} = E_{\alpha,-c}(x)$ , a similar result holds for  $c < 0$ . If  $c = 0$  then (2.2) reduces to

$$(2.3) \quad E_{\alpha,0}(x) = \int_0^\infty \frac{1}{t^\alpha(t-x)} dt = \frac{\pi \cot \alpha\pi}{x^\alpha}.$$

Let  $\psi_0(t) = f(t)$  and define  $\psi_n(t)$  by

$$(2.4) \quad f(t) = \sum_{s=0}^{n-1} a_s e^{ict} t^{-s-\alpha} + \psi_n(t)$$

for  $n = 1, 2, 3, \dots$ . Put

$$(2.5) \quad \delta_n(x) = \int_0^\infty \frac{t^n \psi_n(t)}{t-x} dt \quad (n = 0, 1, 2, \dots).$$

Since  $\psi_n(t) = O(t^{-n-\alpha})$  as  $t \rightarrow \infty$ , the above Cauchy-principal-value integral exists for each  $n \geq 0$ . Note that  $\delta_0(x) = H_f^+(x)$ .

The following results provide explicit expressions for the error terms associated with the asymptotic expansions of the Hilbert transform.

**THEOREM 1.** *Let  $f(t)$  be a locally integrable function on  $[0, \infty)$  and satisfy (1.4) with  $0 < \alpha < 1$ . Then for any  $n \geq 1$*

$$(2.6) \quad H_f^+(x) = E_{\alpha,c}(x) \sum_{s=0}^{n-1} \frac{a_s}{x^s} - \sum_{s=1}^n \frac{b_s}{x^s} + \frac{1}{x^n} \delta_n(x),$$

where the coefficients  $b_s$  are given by

$$(2.7) \quad b_s = \int_0^\infty t^{s-1} \psi_s(t) dt.$$

*Proof.* For any  $n \geq 1$ , we have

$$(2.8) \quad \psi_n(t) = \psi_{n-1}(t) - a_{n-1} e^{ict} t^{-\alpha-n+1},$$

and

$$(2.9) \quad \delta_n(x) = \int_0^\infty t^{n-1} \psi_n(t) dt + x \int_0^\infty \frac{t^{n-1} \psi_n(t)}{t-x} dt.$$

The first integral is simply the coefficient  $b_n$ . Inserting (2.8) in (2.9) gives

$$\delta_n(x) = b_n + x[\delta_{n-1}(x) - a_{n-1} E_{\alpha,c}(x)],$$

which implies

$$\frac{1}{x^n} \delta_n(x) = -E_{\alpha,c}(x) \frac{a_{n-1}}{x^{n-1}} + \frac{b_n}{x^n} + \frac{1}{x^{n-1}} \delta_{n-1}(x).$$

Repeated application of this identity leads to

$$\frac{1}{x^n} \delta_n(x) = -E_{\alpha,c}(x) \sum_{s=0}^{n-1} \frac{a_s}{x^s} + \sum_{s=1}^n \frac{b_s}{x^s} + H_f^+(x),$$

which is the exact statement of the theorem.

We observe that if  $\alpha = 1$  in (1.4) then the integral  $E_{\alpha,c}(x)$  diverges. However, the above analysis can be extended to include this case. We shall separate the discussion into two cases: (i)  $c = 0$  and (ii)  $c \neq 0$ .

**THEOREM 2.** *Let  $f(t)$  be a locally integrable function on  $[0, \infty)$  and satisfy (1.4) with  $\alpha = 1$  and  $c = 0$ . Then for  $x > 1$  and for any  $n \geq 1$*

$$(2.10) \quad H_f^+(x) = \left(\ln \frac{1}{x}\right) \sum_{s=0}^{n-1} \frac{a_s}{x^{s+1}} - \sum_{s=1}^n \frac{c_s}{x^s} + \frac{1}{x^n} \delta_n(x),$$

where the coefficients  $c_s$  are given by

$$(2.11) \quad c_s = \int_0^1 t^{s-1} \psi_{s-1}(t) dt + \int_1^\infty t^{s-1} \psi_s(t) dt.$$

*Proof.* For any  $n \geq 1$ , we write

$$\delta_n(x) = \int_0^1 \frac{t^n \psi_n(t)}{t-x} dt + \int_1^\infty \frac{t^n \psi_n(t)}{t-x} dt.$$

From (2.4), with  $\alpha = 1$  and  $c = 0$ , we have

$$(2.12) \quad \begin{aligned} \int_0^1 \frac{t^n \psi_n(t)}{t-x} dt &= \int_0^1 \frac{t^n \psi_{n-1}(t)}{t-x} dt - a_{n-1} \ln \left(\frac{x-1}{x}\right) \\ &= \int_0^1 t^{n-1} \psi_{n-1}(t) dt + x \int_0^1 \frac{t^{n-1} \psi_{n-1}(t)}{t-x} dt - a_{n-1} \ln \left(\frac{x-1}{x}\right). \end{aligned}$$

Similar argument gives

$$(2.13) \quad \begin{aligned} \int_1^\infty \frac{t^n \psi_n(t)}{t-x} dt &= \int_1^\infty t^{n-1} \psi_n(t) dt + x \int_1^\infty \frac{t^{n-1} \psi_n(t)}{t-x} dt \\ &= \int_1^\infty t^{n-1} \psi_n(t) dt + x \int_1^\infty \frac{t^{n-1} \psi_{n-1}(t)}{t-x} dt + a_{n-1} \ln(x-1). \end{aligned}$$

Coupling the results (2.12) and (2.13) together, we obtain

$$\delta_n(x) = c_n + a_{n-1} \ln x + x \delta_{n-1}(x),$$

which is equivalent to

$$\frac{1}{x^n} \delta_n(x) = \frac{c_n}{x^n} + (\ln x) \frac{a_{n-1}}{x^n} + \frac{1}{x^{n-1}} \delta_{n-1}(x).$$

Repeated application of this identity yields

$$\frac{1}{x^n} \delta_n(x) = (\ln x) \sum_{s=0}^{n-1} \frac{a_s}{x^{s+1}} + \sum_{s=1}^n \frac{c_s}{x^s} + H_f^+(x).$$

This completes the proof of the theorem.

For the case in which  $c \neq 0$  and  $\alpha = 1$ , we need the identity

$$(2.14) \quad \int_0^\infty \frac{e^{ict}}{t-x} dt = i e^{icx} \quad (c > 0),$$

see [1, p. 251]. A similar result holds for  $c < 0$ . For convenience, we shall also introduce the notation

$$(2.15) \quad E_c(x) = i e^{icx} + E_1(-ic),$$

where  $E_1(z)$  denotes the exponential integral [9, pp. 40–42]. Note that the second term on the right is simply a constant.

**THEOREM 3.** *Let  $f(t)$  be a locally integrable function on  $[0, \infty)$  and satisfy (1.4) with  $\alpha = 1$  and  $c \neq 0$ . Then for any  $n \geq 1$*

$$(2.16) \quad H_f^+(x) = E_c(x) \sum_{s=0}^{n-1} \frac{a_s}{x^{s+1}} - \sum_{s=1}^n \frac{c_s}{x^s} + \frac{1}{x^n} \delta_n(x),$$

where  $c_s$  is given in (2.11).

Since the proof of this result is very similar to that of Theorem 2, we omit it completely.

*Remark.* There seem to be two other approaches to the above problem, which are entirely different from the one given above. One approach is to use the Plemelj formula [5] to write

$$H_f^+(x) = \frac{1}{2} \lim_{\epsilon \rightarrow 0} \int_0^\infty \left[ \frac{1}{t-x+i\epsilon} + \frac{1}{t-x-i\epsilon} \right] f(t) dt.$$

The integrals on the right-hand side are Stieltjes transforms of  $f(t)$  and hence one may use the results already developed in [6]. The disadvantage in this approach is that one meets a formidable difficulty in estimating the error term  $\delta_n(x)$ . (A simple method is given in the next section.) Another approach is to view  $H_f^+(x)$  as a repeated Fourier transform and apply the results available for this transformation. However, in this approach, there arises the question whether asymptotic expansions of Fourier transform can be differentiated (see Condition (ii) in [10]): bear in mind that  $f(t)$  decays only algebraically.

**3. Bounds for  $\delta_n(x)$ .** To show that the expansions obtained in § 2 are indeed asymptotic in nature, one must prove that

$$(3.1) \quad \delta_n(x) = o(1) \quad \text{as } x \rightarrow \infty.$$

We shall, in fact, prove that there exists a positive constant  $M_n$  such that

$$(3.2) \quad |\delta_n(x)| \leq M_n \frac{\ln x}{x^\alpha},$$

for all  $x > e$ .

**THEOREM 4.** *Let  $f(t)$  be a locally integrable function on  $[0, \infty)$  and satisfy (1.4). If, in addition,  $f \in C^1(0, \infty)$  and the asymptotic expansion of  $f'(t)$  is obtained by differentiating (1.4), then the function  $\delta_n(x)$  given in (2.5) satisfies (3.2).*

*Proof.* Write

$$(3.3) \quad \delta_n(x) = \delta_{n,1}(x) + \delta_{n,2}(x) + \delta_{n,3}(x),$$

where the integrals  $\delta_{n,1}$ ,  $\delta_{n,2}$ ,  $\delta_{n,3}$  correspond respectively to the intervals  $(0, x-1)$ ,

$(x - 1, x + 1), (x + 1, \infty)$ . Put

$$(3.4) \quad M_{n,1} = \int_0^1 t^n |\psi_n(t)| dt,$$

$$(3.5) \quad M_{n,2} = \sup \{t^{\alpha+n} |\psi_n(t)| : t \geq 1\}.$$

Under the hypotheses, both numbers  $M_{n,1}$  and  $M_{n,2}$  are finite. To estimate  $\delta_{n,1}(x)$  we further divide the range of integration at  $t = 1$ . It is easy to see that

$$|\delta_{n,1}(x)| \leq \frac{M_{n,1}}{x-1} + M_{n,2} \int_1^{x-1} \frac{1}{t^\alpha(x-t)} dt.$$

In the last integral we make a change of variables  $t = xu$ . The resulting integral is dominated by

$$\frac{1}{x^\alpha} \int_{1/x}^{(x-1)/x} \frac{1}{u(1-u)} du,$$

which is in its turn dominated by  $2 \ln(x - 1)$ . Therefore

$$(3.6) \quad |\delta_{n,1}(x)| \leq 2(M_{n,1} + M_{n,2}) \frac{\ln x}{x^\alpha}.$$

The integral  $\delta_{n,3}(x)$  can be estimated similarly. Here we have

$$|\delta_{n,3}(x)| \leq \frac{1}{x^\alpha} M_{n,2} \int_{1/x}^\infty \frac{1}{(1+u)^\alpha u} du.$$

In the interval  $(x^{-1}, 1)$ , we use the bound  $(1+u)^{-\alpha} \leq 1$ , whereas in the interval  $(1, \infty)$ , we use  $(1+u)^{-\alpha} \leq u^{-\alpha}$ . Thus

$$(3.7) \quad |\delta_{n,3}(x)| \leq \left(1 + \frac{1}{\alpha}\right) M_{n,2} \frac{\ln x}{x^\alpha}.$$

We now turn to the consideration of  $\delta_{n,2}(x)$ . Let  $\varphi_n(t) = t^n \psi_n(t)$  and write

$$(3.8) \quad \delta_{n,2}(x) = \int_{x-1}^{x+1} \frac{\varphi_n(t) - \varphi_n(x)}{t-x} dt = \int_{x-1}^{x+1} \varphi'_n(\xi) dt,$$

where  $\xi$  is between  $t$  and  $x$ . By hypotheses, as  $t \rightarrow \infty$ ,  $\psi_n(t) = O(t^{-n-\alpha})$  and  $\psi'_n(t) = O(t^{-n-\alpha})$ . Put

$$(3.9) \quad M_{n,3} = \sup \{t^{n+\alpha} |\psi'_n(t)| : t \geq 1\}.$$

Then we have

$$|\varphi'_n(t)| \leq (nM_{n,2} + M_{n,3})t^{-\alpha}$$

for all  $t \geq 1$ . (Note that  $M_{n,3}$  is finite). It now follows from (3.8) that

$$(3.10) \quad |\delta_{n,2}(x)| \leq \left(\frac{3}{2}\right)^{\alpha+1} (nM_{n,2} + M_{n,3}) \frac{\ln x}{x^\alpha}.$$

Combination of this result together with (3.6) and (3.7) gives the desired conclusion (3.2).

We remark that the quantities  $M_{n,i}$ ,  $i = 1, 2, 3$ , were introduced merely to provide a rough estimate of the constant  $M_n$  in (3.2). Of course, much better bounds may be obtained in specific instances.



**4. Examples.** Examining the statements in Theorems 1, 2 and 3, one immediately observes that the major difficulty in applying these results resides in the evaluation of the integrals in (2.7) and (2.11). In this section, we shall illustrate some techniques by which one may succeed in calculating these integrals explicitly. For further examples of this nature, we refer to the paper by Grosjean [2].

*Example 1.* From the Cauchy-type nature of the integral (1.3), one is tempted to conjecture that the dominant term of the asymptotic approximation to  $H_f^+(x)$  is of the same order as the function  $f(x)$ . This conjecture is, however, not true as we shall see in this example. Let  $f(t) = \sqrt{t}/(1+t)$ . As  $t \rightarrow \infty$ , we have

$$(4.1) \quad f(t) \sim \sum_{s=0}^{\infty} (-1)^s t^{-s-1/2}.$$

Hence, in the notation of §§ 1 and 2,  $c = 0$ ,  $\alpha = \frac{1}{2}$  and  $a_s = (-1)^s$ . Since

$$\psi_n(t) = \frac{(-1)^n}{t^{n-1/2}(1+t)},$$

the coefficients  $b_s$  are given by

$$b_s = (-1)^s \int_0^{\infty} \frac{dt}{\sqrt{t}(1+t)} dt = (-1)^2 \pi.$$

From (2.6) it now follows that

$$(4.2) \quad H_f^+(x) = -\pi \sum_{s=1}^n \frac{(-1)^s}{x^s} + \frac{1}{x^n} \delta_n(x).$$

A simple calculation shows that

$$M_{n,1} = 0.44, \quad M_{n,2} = \frac{1}{2}, \quad M_{n,3} = \frac{n}{2}.$$

Thus we obtain, from (3.2),

$$(4.3) \quad |\delta_n(x)| \leq [3.38 + 1.84n] \frac{\ln x}{\sqrt{x}}.$$

The estimate (4.3) is rather crude in comparison with the actual result  $|\delta_n(x)| = \pi/(1+x)$ , which can be obtained from the closed form evaluation  $H_f^+(x) = \pi/(1+x)$ ; see [1, p. 251]. We have chosen this example mainly to illustrate the calculation of the coefficients  $b_s$  and to make the observation that dominant approximation in this case is  $O(x^{-1})$  and not  $O(x^{-1/2})$  as one might have expected.

*Example 2.* Consider the integral

$$(4.4) \quad I(x) = \int_0^{\infty} \frac{e^{-u}}{1-xu} du \quad (x > 0).$$

In terms of the Hilbert transform we have

$$I(x) = \int_0^{\infty} \frac{f(t)}{t-x} dt,$$

where

$$f(t) = \frac{1}{t} e^{-1/t} \sim \sum_{s=0}^{\infty} \frac{(-1)^s}{s!} t^{-s-1}, \quad \text{as } t \rightarrow \infty.$$

Thus  $c = 0$ ,  $\alpha = 1$  and  $a_s = (-1)^s/s!$ . To calculate the coefficients  $c_s$ , we note that the first and the second integrals in (2.11) are, respectively, equal to

$$\int_1^{\infty} u^{-s} e^{-u} du + \sum_{k=0}^{s-2} \frac{(-1)^k}{k!(1+k-s)}$$

and

$$\lim_{\epsilon \rightarrow 0} \left[ \int_{\epsilon}^1 u^{-s} e^{-u} du - \sum_{k=0}^{s-2} \frac{(-1)^k}{k!(1+k-s)} + \frac{(-1)^{s-1}}{(s-1)!} \ln \epsilon + \sum_{k=0}^{s-2} \frac{(-1)^k \epsilon^{1+k-s}}{k!(1+k-s)} \right].$$

Adding these two quantities together gives

$$c_s = \lim_{\epsilon \rightarrow 0} \left[ \epsilon^{1-s} E_s(\epsilon) + \frac{(-1)^{s-1}}{(s-1)!} \ln \epsilon + \sum_{k=0}^{s-2} \frac{(-1)^k \epsilon^{1+k-s}}{k!(1+k-s)} \right],$$

where  $E_s(\epsilon)$  is the generalized exponential integral [9, p. 43]. Using the identity [9, p. 43]

$$(4.5) \quad E_s(\epsilon) = \frac{(-\epsilon)^{s-1}}{(s-1)!} \{-\ln \epsilon + \psi(s)\} + \sum'_{k=0}^{\infty} \frac{(-\epsilon)^k}{k!(s-k-1)},$$

we have

$$c_s = \frac{(-1)^s}{s!} \psi(s).$$

In (4.5) the prime on the summation signifies that the term  $k = s - 1$  is omitted, and  $\psi(s)$  denotes the logarithmic derivative of  $\Gamma(s)$ . It now follows from Theorems 2 and 4 that

$$(4.6) \quad I(x) \sim \left( \ln \frac{1}{x} \right) \sum_{s=0}^{\infty} \frac{(-1)^s}{s! x^{s+1}} - \sum_{s=1}^{\infty} \frac{(-1)^s \psi(s)}{s! x^s}.$$

Taking the first two terms in the expansion, we have

$$(4.7) \quad I(x) = \frac{1}{x} \left( \ln \frac{1}{x} \right) - \frac{\gamma}{x} + \frac{1}{x} \delta_1(x),$$

where  $\gamma$  is the Euler constant. A simple calculation gives  $M_1 \leq 12.5$ . Hence

$$|\delta_1(x)| \leq (12.5) \frac{\ln x}{\sqrt{x}}.$$

Similar results can be obtained for higher error terms.

In conclusion, we wish to make the following remark. If the path of integration  $(0, \infty)$  in (1.3) or (2.5) can be deformed into a ray,  $\arg t = \gamma (\gamma \neq 0)$ , then the resulting integral no longer has a singularity at  $t = x$  and hence better error bounds may be obtained; (see the estimates in [6]). However, such deformations presuppose that  $f(t)$  is the restriction of an analytic function and that the asymptotic expansion (1.4) is valid in a sector. In this paper, we have confined ourselves entirely to the real-variable methods.

**Acknowledgment.** I am grateful to the referee for his helpful suggestions.

## REFERENCES

- [1] A. ERDELYI, W. MAGNUS, F. OBERHETTINGER AND F. TRICOMI, *Tables of Integral Transforms*, vol. 2, McGraw-Hill, New York, 1953.
- [2] C. C. GROSJEAN, *On the series expansion of certain types of Fourier integrals in the neighborhood of the origin*, Bull. Soc. Math. Belgique, 17 (1965), pp. 251–418.
- [3] R. A. HANDELSMAN AND J. S. LEW, *Asymptotic expansion of a class of integral transforms with algebraically dominated kernels*, J. Math. Anal., 35 (1971), pp. 405–433.
- [4] D. S. JONES, *Asymptotic behavior of integrals*, SIAM Rev., 14 (1972), pp. 286–317.
- [5] A. KYRALA, *Applied Functions of a Complex Variable*, Wiley-Interscience, New York, 1972.
- [6] J. P. MCCLURE AND R. WONG, *Explicit error terms for asymptotic expansions of Stieltjes transforms*, J. Inst. Maths. Applics., 22 (1978), pp. 129–145.
- [7] R. F. MILLAR, *On the asymptotic behavior of two classes of integrals*, SIAM Rev., 8 (1966), pp. 188–195.
- [8] N. I. MUSKHELISHVILI, *Singular Integral Equations*, Noordhoff, Groningen, 1946.
- [9] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
- [10] ———, *Error bounds for stationary phase approximations*, this Journal, 5 (1974), 19–29.
- [11] ———, *Unsolved problems in the asymptotic estimation of special functions*, Theory and Applications of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975, pp. 99–142.
- [12] E. C. TITCHMARCH, *Fourier Integrals*, Oxford, 1937.

## PRODUCT FORMULAS FOR $q$ -HAHN POLYNOMIALS\*

DENNIS STANTON†

**Abstract.** Product formulas for general  $q$ -Hahn polynomials are derived from counting arguments involving subspaces of a finite vector space.

**1. Introduction.** Product formulas for orthogonal polynomials appear most naturally as first terms of addition theorems. An addition theorem for a family of  $q$ -Hahn polynomials has been found by Dunkl [7]. We shall give a product formula for general  $q$ -Hahn polynomials which will be the first term of a yet undiscovered addition theorem for  $q$ -Hahn polynomials. The derivation will depend upon the  $q$ -Hahn polynomials being functions on the general linear group over a finite field satisfying certain invariance properties. This was first discovered by Delsarte [3]. The problem then is reduced to a combinatorial one involving subspaces of a finite dimensional vector space. In § 2 we give the elementary properties of  $q$ -Hahn polynomials, describe the geometry leading to the polynomials, and state two combinatorial propositions. The product theorem for  $q$ -Hahn polynomials (Theorem 1) is proved in § 3. Using a transformation this formula becomes a linearization formula for dual  $q$ -Hahn polynomials. Sufficient conditions for the positivity of the coefficients are easily found. We also give a sharp bound for some  $q$ -Hahn polynomials.

**2. Preliminaries.** First we briefly describe the circumstances leading to product formulas. All of this material can be found in Dunkl [6].

If  $X$  is a finite set let  $L(X)$  denote the complex valued functions on  $X$  and  $|X|$  denote the cardinality of  $X$ . If  $f, g \in L(X)$ , define the inner product  $\langle f, g \rangle = |X|^{-1} \sum_{x \in X} f(x) \overline{g(x)}$  with  $\langle f, f \rangle = \|f\|_2^2$ . Suppose  $G$  is a finite group with identity  $e$ . For any subgroup  $H$  of  $G$ , let  $L_H(G) = \{f \in L(G) | f(hg) = f(g) \text{ for all } h \in H \text{ and } g \in G\}$ . We can identify  $L_H(G) \cong L(X)$ , if  $X = H \backslash G = \{Hg | g \in G\}$ . Let  $R$  denote right translation,  $(R(g_1)f)(g_2) = f(g_2g_1)$ ,  $f \in L(G)$ ,  $g_1, g_2 \in G$ . Suppose  $V \subset L_H(G) \cong L(X)$  is an irreducible  $G$ -module such that  $V_H = \{f \in V | R(h)f = f \text{ for all } h \in H\}$  and  $V_K = \{f \in V | R(k)f = f \text{ for all } k \in K\}$  are 1-dimensional subspaces of  $V$  for some subgroup  $K$  of  $G$ . Let  $\phi_{HH} \in V_H$  and  $\phi_{HK} \in V_K$  be normalized by  $\|\phi_{HK}\|_2^2 = \|\phi_{HH}\|_2^2 = (\dim V)^{-1}$ . Then the following product formulas hold for any  $g_1, g_2 \in G$ ;

$$(2.1) \quad \phi_{HH}(g_1)\phi_{HH}(g_2) = |H|^{-1} \sum_{h \in H} \phi_{HH}(g_1hg_2),$$

$$(2.2) \quad \overline{\phi_{HK}(e)}\phi_{HK}(g_1)\phi_{HK}(g_2) = (|H||K|)^{-1} \sum_{\substack{h \in H \\ k \in K}} \phi_{HK}(g_1khg_2),$$

$$(2.3) \quad \phi_{HK}(g_1)\overline{\phi_{HK}(g_2)} = |K|^{-1} \sum_{k \in K} \phi_{HH}(g_1kg_2^{-1}),$$

$$(2.4) \quad \phi_{HH}(g_1)\phi_{HK}(g_2) = |H|^{-1} \sum_{h \in H} \phi_{HK}(g_1hg_2).$$

\* Received by the editors September 20, 1978 and in revised form January 10, 1979.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. Much of the material in this paper appears in the author's thesis under the direction of Richard Askey.

In this paper we shall explicitly compute the product formulas (2.1)–(2.4) when  $\phi_{HH}$  and  $\phi_{HK}$  are  $q$ -Hahn polynomials.

The  $q$ -Hahn polynomials [9] are defined in terms of basic hypergeometric series. If  $k$  is a nonnegative integer, let

$${}_3\phi_2\left(\begin{matrix} q^{-k} & a & b \\ c & d \end{matrix} \middle| q; x\right) = \sum_{j=0}^k \frac{(q^{-k}; q)_j (a; q)_j (b; q)_j}{(q; q)_j (c; q)_j (d; q)_j} x^j,$$

where  $(a; q)_j = (1 - a)(1 - aq) \cdots (1 - aq^{j-1})$ ,  $(a; q)_0 = 1$ . If  $k$  and  $N$  are integers,  $0 \leq k \leq N$ , the  $k$ th  $q$ -Hahn polynomial is defined by

$$(2.5) \quad Q_k(q^{-x}; a, b, N; q) = {}_3\phi_2\left(\begin{matrix} q^{-k} & abq^{k+1} & q^{-x} \\ aq & q^{-N} \end{matrix} \middle| q; q\right).$$

This is a polynomial of degree  $k$  in  $\lambda(x) = q^{-x}$ . The orthogonality relation for these polynomials is [2]

$$(2.6) \quad \sum_{x=0}^N Q_k(q^{-x}; a, b, N; q) Q_j(q^{-x}; a, b, N; q) w(x) = \delta_{jk} h_k^{-1},$$

where

$$w(x) = \frac{(q^{-N}; q)_x (aq; q)_x}{(q; q)_x (q^{-N}/b; q)_x} (abq)^{-x},$$

$$h_k = \frac{(q^{-N}/b; q)_{N-k} (aq; q)_k (q^{-N}; q)_k}{(q^{-N-k-1}/ab; q)_{N-k} (q^{-1}; q^{-1})_k (abq^{k+1}; q)_k}.$$

The weight function  $w(x)$  is positive if  $0 < q \neq 1$  and  $0 < a, b$  lie in the same component of the complement of the closed interval from  $q^{-1}$  to  $q^{-N}$ . We shall need a transformation of a terminating  ${}_3\phi_2$  [2]

$$(2.7) \quad {}_3\phi_2\left(\begin{matrix} q^{-k} & a & b \\ c & d \end{matrix} \middle| q; q\right) = \frac{(c/a; q)_k}{(c; q)_k} a^k {}_3\phi_2\left(\begin{matrix} q^{-k} & a & d/b \\ aq^{-k+1}/c & d \end{matrix} \middle| q; \frac{bq}{c}\right).$$

For  $q$ -Hahn polynomials (2.7) implies

$$(2.8) \quad Q_k(q^{-x}; a, b, N; q) = \frac{(q^{-k}/b; q)_k}{(aq; q)_k} (abq^{k+1})^k Q_k(q^{N-x}; b^{-1}, a^{-1}, N; q^{-1}),$$

$$(2.9) \quad Q_k(q^{-x}; a, b, N; q) = \frac{(bq^N; q^{-1})_x}{(1/aq; q^{-1})_x} Q_{N-k}(q^{-x}; q^{-N-1}b^{-1}, q^{-N-1}a^{-1}, N; q).$$

Finally we mention the dual  $q$ -Hahn polynomials  $E_k(q^{-x}; a, b, N; q) = Q_x(q^{-k}; a, b, N; q)$ , which are polynomials in  $\mu(x) = (1 - q^{-x})(1 + abq^{x+1})$  orthogonal with respect to  $h_x$ .

The  $q$ -Hahn polynomials  $Q_k(q^{-x}; q^{n-v-1}, q^{-n-1}, n; q)$  were realized as spherical functions on  $GL(v, GF(q))$  by Delsarte [3]. We recall the situation. Let  $E$  be a vector space of dimension  $v$  over  $GF(q)$  with a fixed basis  $\{e_1, \dots, e_v\}$ . If  $A$  is any subset of a

vector space, we denote by  $\text{sp}A$  the span of  $A$ . Let  $X_n, 0 \leq 2n \leq v$ , denote the set of  $n$ -dimensional subspaces of  $E$ . In terms of the  $q$ -binomial coefficient [1] we have

$$|X_n| = \begin{bmatrix} v \\ n \end{bmatrix}_q = \frac{(q^v; q^{-1})_n}{(q; q)_n}.$$

Fix a base point  $\omega_n \in X_n, \omega_n = \text{sp}\{e_1, \dots, e_n\}$ . Then  $G = GL(v, q)$  acts transitively on  $X_n$  with isotropy subgroup  $H_n = \{g \in G | \omega_n g = \omega_n\}$ . The permutation representation of  $G$  on  $L(X_n)$  is multiplicity free with  $n+1$  irreducible constituents,  $L(X_n) = \sum_{k=0}^n V_k$ . The modules  $V_k$  were first described by Steinberg [10], and their relation to  $q$ -Hahn polynomials is discussed by Dunkl [7]. The spherical functions are the  $H_n$ -invariant functions in each  $V_k$ . They are constant on the  $H_n$ -orbits on  $X_n$ , which are  $\Omega_{n,x} = \{\alpha \in X_n | \dim(\alpha \cap \omega_n) = n-x, 0 \leq x \leq n\}$ . For  $V_k$  we have [7]  $\dim V_k = \begin{bmatrix} v \\ k \end{bmatrix}_q - \begin{bmatrix} v \\ k-1 \end{bmatrix}_q$  and  $\phi_{HH}(g) = Q_k(q^{-x}; q^{n-v-1}, q^{-n-1}, n; q)$ , where  $n-x = \dim(\omega_n g \cap \omega_n)$ .

If  $n \leq m \leq v-n$ , the  $H_m$ -invariant functions on  $X_n$  are constant on  $\Delta_{n,x} = \{\alpha \in X_n | \dim(\alpha \cap \omega_m) = n-x, 0 \leq x \leq n\}$ . It is straightforward to verify [7] that the  $H_m$ -invariant elements of  $V_k$  are spanned by  $\phi_{HK}(g) = Q_k(q^{-x}; q^{m-v-1}, q^{-m-1}, n; q)$ , where  $x = n - \dim(\omega_n g \cap \omega_m)$ . Henceforth we shall have  $H = H_n$  and  $K = H_m$ .

Before working out the product formulas (2.1)–(2.4), we state two enumerative propositions involving  $q$ -binomial coefficients.

PROPOSITION 1. Let  $V_N$  and  $W_M$  be  $N$  and  $M$ -dimensional vector spaces over  $GF(q)$ .

- (i) The number of pairs of  $n$ -dimensional subspaces  $(V_n, W_n), V_n \subset V_N, W_n \subset W_M$ , is  $\begin{bmatrix} N \\ n \end{bmatrix}_q \begin{bmatrix} M \\ n \end{bmatrix}_q$ .
- (ii) For any pair  $(V_n, W_n)$  there are  $|GL(n, q)| = (q; q)_n q^{\binom{n}{2}} (-)^n$  nonsingular linear transformations from  $V_n$  to  $W_n$ .
- (iii) The number of  $k$ -dimensional subspaces  $W_k \subset W_M$  such that  $W_k \cap W_n = \{0\}$  is  $\begin{bmatrix} M-n \\ k \end{bmatrix}_q q^{kn}$ .
- (iv) The number of linear transformations  $g$  from  $V_n$  to  $W_k$  such that  $\dim(V_n g) = j$  is  $a(n, k, j) = \begin{bmatrix} n \\ j \end{bmatrix}_q \begin{bmatrix} k \\ j \end{bmatrix}_q (q; q)_j q^{\binom{j}{2}} (-)^j$ .

PROPOSITION 2. Let  $t, n, m, l, k, j$ , and  $i$  be integers such that  $0 \leq t \leq n \leq m, 0 \leq j \leq m-n+t, j-t+l-k \leq i \leq j$ , and  $\max(l-t, 0) \leq k \leq \min(l, n-t)$ . Let  $\{e_1, \dots, e_{m+i}\}$  be a basis for  $V$  over  $GF(q)$ , and let  $A_l$  be a fixed  $l$ -dimensional subspace of  $\text{sp}\{e_1, \dots, e_n\}$  such that  $\dim(A_l \cap \text{sp}\{e_1, \dots, e_{n-t}\}) = k$ . Then the number of  $j$ -dimensional subspaces  $B_j$  of  $V$  such that  $\dim((A_l + B_j) \cap \text{sp}\{e_1, \dots, e_{n-t}, e_{n+1}, \dots, e_{m+i}\}) = k+i$  and  $B_j \cap \text{sp}\{e_1, \dots, e_n\} = \{0\}$  is  $\begin{bmatrix} m-n+t \\ j \end{bmatrix}_q a(j, t-l+k, j-i) q^{i(n-t+l-k)}$ .

Proof. If  $B_j \cap \text{sp}\{e_1, \dots, e_n\} = \{0\}$ , the projection of  $B_j$  on  $\text{sp}\{e_{n+1}, \dots, e_{m+i}\}$  is a  $j$ -dimensional subspace  $F_j$ . Fix  $F_j$  with a basis  $\{f_1, \dots, f_j\}$ . Then  $B_j$  is uniquely determined by  $\{z_1, \dots, z_j\} \subset \text{sp}\{e_1, \dots, e_n\}$  such that  $\{f_1 + z_1, \dots, f_j + z_j\}$  is a basis for  $B_j$ . Let  $C_{t-l+k}$  be a fixed  $(t-l+k)$ -dimensional complement in  $\text{sp}\{e_{n-t+1}, \dots, e_n\}$  to the projection of  $A_l$  on  $\text{sp}\{e_{n-t+1}, \dots, e_n\}$ . If  $\dim((A_l + B_j) \cap \text{sp}\{e_1, \dots, e_{n-t}, e_{n+1}, \dots, e_{m+i}\}) = k+i$ , the projection of  $\text{sp}\{z_1, \dots, z_j\}$  on  $C_{t-l+k}$  is  $(j-i)$ -dimensional. The projection of  $\text{sp}\{z_1, \dots, z_j\}$  on  $\text{sp}\{e_1, \dots, e_{n-t}\} + A_l$  is arbitrary. This gives  $a(j, t-l+k, j-i) q^{i(n-t+l-k)}$  possible choices for  $\{z_1, \dots, z_j\}$ .  $\square$

**3. The product formula for q-Hahn polynomials.** In this section we shall explicitly calculate the product formula (2.2) for q-Hahn polynomials. For any  $0 \leq l \leq n$ , let  $g_l \in G$  be defined on the basis  $\{e_1, \dots, e_v\}$  by  $e_i g_l = e_{v+1-i}$ ,  $1 \leq i \leq l$ ,  $v-l+1 \leq i \leq v$ ,  $e_i g_l = e_i$ ,  $l+1 \leq i \leq v-l$ , and  $g_0 = \text{identity}$ . Fix  $0 \leq s, t \leq n$  and let  $g_1 = g_s$  and  $g_2 = g_t$  in the product formula (2.2). In order to evaluate the argument of the q-Hahn polynomial on the right hand side of (2.2), we need to find  $\dim(\omega_n g_s k h g_t \cap \omega_m) = \dim(\omega_n g_s k h \cap \omega_m g_t)$ . This is the intent of the next two lemmas.

LEMMA 1. Let  $\omega_n$  and  $g_s$  be as above. The number of  $k \in K$  such that  $\dim(\omega_n g_s k \cap \omega_n) = n - s - \alpha$ ,  $0 \leq \alpha \leq n - s$ , is  $\begin{bmatrix} n-s \\ n-s-\alpha \end{bmatrix}_q \begin{bmatrix} n \\ n-s-\alpha \end{bmatrix}_q |GL(n-s-\alpha, q)| \begin{bmatrix} m-n \\ \alpha \end{bmatrix}_q q^{\alpha n} |GL(\alpha, q)| |GL(m-n+s, q)| q^{(m-n+s)(n-s)} |GL(v-m, q)| q^{m(v-m)}$ .

*Proof.* Clearly  $\omega_n g_s = \text{sp}\{e_{s+1}, \dots, e_n, e_{v-s+1}, \dots, e_v\}$ , so that  $\omega_m \cap \omega_n g_s = \text{sp}\{e_{s+1}, \dots, e_n\}$ . Since  $k \in K$  fixes  $\omega_m \supset \omega_n$ ,  $\omega_n g_s k \cap \omega_n = \text{sp}\{e_{s+1}, \dots, e_n\} k \cap \omega_n$ . Choose  $(n-s-\alpha)$ -dimensional subspaces  $A_{n-s-\alpha}$  and  $A'_{n-s-\alpha}$  of  $\text{sp}\{e_{s+1}, \dots, e_n\}$  and  $\omega_n$  respectively such that  $A_{n-s-\alpha} k = A'_{n-s-\alpha} = \omega_n g_s k \cap \omega_n$ . By Proposition 1(i)-(ii) there are  $\begin{bmatrix} n-s \\ n-s-\alpha \end{bmatrix}_q \begin{bmatrix} n \\ n-s-\alpha \end{bmatrix}_q |GL(n-s-\alpha, q)|$  possible choices for  $k$ . To extend  $k$  to  $\text{sp}\{e_{s+1}, \dots, e_n\}$ , let  $B_\alpha$  be a fixed complement of  $A_{n-s-\alpha}$  in  $\text{sp}\{e_{s+1}, \dots, e_n\}$  and  $B'_\alpha$  be an  $\alpha$ -dimensional subspace of  $\omega_m$  such that  $\omega_n \cap B'_\alpha = \{0\}$ . If  $k$  maps  $B_\alpha$  to  $B'_\alpha$ , by Proposition 1(iii) there are  $\begin{bmatrix} m-n \\ \alpha \end{bmatrix}_q q^{\alpha n}$  possible choices for  $B'_\alpha$  and thus  $\begin{bmatrix} m-n \\ \alpha \end{bmatrix}_q q^{\alpha n} |GL(\alpha, q)|$  possible choices for  $k$ . Finally extending  $k$  to  $\omega_m = \omega_m k$  and to  $E$  we obtain from Proposition 1(ii)-(iii)  $|GL(m-n+s, q)| q^{(m-n+s)(n-s)} |GL(v-m, q)| \cdot q^{(v-m)m}$  possible choices for  $k$ .  $\square$

Given  $k \in K$  as in Lemma 1, we now consider  $\omega_n g_s k h$ . Since  $h$  fixes  $\omega_n$ ,  $\dim(\omega_n g_s k h \cap \omega_n) = n - s - \alpha$ . This implies that  $\dim(\omega_n g_s k h \cap \text{sp}\{e_{t+1}, \dots, e_n\}) = n - s - \alpha - \beta$ , for some  $\beta$ ,  $0 \leq n - s - \alpha - \beta \leq n - t$ ; and that  $\dim(\omega_n g_s k h \cap \omega_m g_t) = n - s - \alpha - \beta + \gamma$  for some  $\gamma$ ,  $0 \leq \gamma \leq s + \alpha$ .

LEMMA 2. If  $k \in K$  satisfies the hypothesis of Lemma 1, then the number of  $h \in H$  such that  $\dim(\omega_n g_s k h \cap \text{sp}\{e_{t+1}, \dots, e_n\}) = n - s - \alpha - \beta$  and  $\dim(\omega_n g_s k h \cap \omega_m g_t) = n - s - \alpha - \beta + \gamma$  is

$$\begin{aligned} & \sum_{\theta} \begin{bmatrix} n-t \\ n-s-\alpha-\beta \end{bmatrix}_q \begin{bmatrix} n-s-\alpha \\ n-s-\alpha-\beta \end{bmatrix}_q |GL(n-s-\alpha-\beta, q)| \begin{bmatrix} t \\ \beta \end{bmatrix}_q q^{\beta(n-t)} |GL(\beta, q)| \\ & \cdot q^{(s+\alpha)(n-s-\alpha)} |GL(s+\alpha, q)| \begin{bmatrix} m-n+t \\ \gamma+\theta \end{bmatrix}_q \begin{bmatrix} s+\alpha \\ \gamma+\theta \end{bmatrix}_q |GL(\gamma+\theta, q)| q^{(n-t+\beta)(\gamma+\theta)} \\ & \cdot a(\gamma+\theta, t-\beta, \theta) \begin{bmatrix} v-m-t \\ s+\alpha-\gamma-\theta \end{bmatrix}_q q^{(s+\alpha-\gamma-\theta)(m+t)} \\ & \cdot |GL(s+\alpha-\gamma-\theta, q)| q^{(v-n-\alpha-s)(n+\alpha+s)} \\ & \cdot |GL(v-n-\alpha-s, q)|, \end{aligned}$$

where  $a(\gamma+\theta, t-\beta, \theta)$  is given by Proposition 1(iv).

*Proof.* From Lemma 1  $A'_{n-s-\alpha} = \omega_n g_s k \cap \omega_n$ , and since  $h$  fixes  $\omega_n$ ,  $\omega_n g_s k h \cap \text{sp}\{e_{t+1}, \dots, e_n\} = A'_{n-s-\alpha} h \cap \text{sp}\{e_{t+1}, \dots, e_n\}$ . Let  $C'_{n-s-\alpha-\beta}$  and  $C''_{n-s-\alpha-\beta}$  be  $(n-s-\alpha-\beta)$ -dimensional subspaces of  $A'_{n-s-\alpha}$  and  $\text{sp}\{e_{t+1}, \dots, e_n\}$  respectively. If  $h$  maps  $C'_{n-s-\alpha-\beta}$  to  $C''_{n-s-\alpha-\beta}$ , there are  $\begin{bmatrix} n-t \\ n-s-\alpha-\beta \end{bmatrix}_q \begin{bmatrix} n-s-\alpha \\ n-s-\alpha-\beta \end{bmatrix}_q |GL(n-s-\alpha-\beta, q)|$  possible choices for  $h$  by Proposition 1(i)-(ii). As in Lemma 1, fix a  $\beta$ -dimensional complement  $D'_\beta$  to  $C'_{n-s-\alpha-\beta}$  in  $A'_{n-s-\alpha}$ , and let  $D''_\beta$  be a  $\beta$ -dimensional subspace of  $\omega_n$  such that  $D''_\beta \cap \text{sp}\{e_{t+1}, \dots, e_n\} = \{0\}$ . If  $h$  maps  $D'_\beta$  to  $D''_\beta$ , there are  $\begin{bmatrix} t \\ \beta \end{bmatrix}_q q^{\beta(n-t)} |GL(\beta, q)|$  possible extensions of  $h$  to  $A'_{n-s-\alpha}$  by Proposition 1(ii)-(iii). Extending  $h$  to  $\omega_n$ , by Proposition 1(ii)-(iii) there are  $q^{(s+\alpha)(n-s-\alpha)} |GL(s+\alpha, q)|$  possible choices for  $h$ .

Let  $Z'_{s+\alpha} \subset \omega_n g_s k$  be a fixed complement to  $A'_{n-s-\alpha}$ . We shall define  $h$  on  $Z'_{s+\alpha}$ . If  $\dim(\omega_n g_s k h \cap \omega_m g_t) = n-s-\alpha-\beta+\gamma$ , then  $\dim(\omega_n g_s k h \cap \text{sp}\{e_1, \dots, e_m, e_{v-t+1}, \dots, e_v\}) = n-s-\alpha-\beta+\gamma+\theta$  for some  $\theta$ ,  $n-s-\alpha-\beta+\gamma \leq n-s-\alpha-\beta+\gamma+\theta \leq n$ . Choose a  $(\theta+\gamma)$ -dimensional subspace  $F''_{\theta+\gamma}$  of  $\text{sp}\{e_1, \dots, e_m, e_{v-t+1}, \dots, e_v\}$  such that  $\dim((F''_{\theta+\gamma} + C''_{n-s-\alpha-\beta} + D''_\beta) \cap \omega_m g_t) = n-s-\alpha-\beta+\gamma$  and  $F''_{\theta+\gamma} \cap \omega_n = \{0\}$ . By Proposition 2 ( $j = \theta + \gamma$ ,  $l = n-s-\alpha$ ,  $k = n-s-\alpha-\beta$ ,  $i = \gamma$ ,  $A_l = C''_{n-s-\alpha-\beta} + D''_\beta$ ,  $V = \text{sp}\{e_1, \dots, e_m, e_{v-t+1}, \dots, e_v\}$ ) the number of such  $F''_{\theta+\gamma}$  is  $\begin{bmatrix} m-n+t \\ \theta+\gamma \end{bmatrix}_q a(\theta+\gamma, t-\beta, \theta) q^{(\theta+\gamma)(n-t+\beta)}$ . Let  $F'_{\theta+\gamma}$  be a  $(\theta+\gamma)$ -dimensional subspace of  $Z'_{s+\alpha}$  and let  $h$  map  $F'_{\theta+\gamma}$  to  $F''_{\theta+\gamma}$ . There are  $\begin{bmatrix} s+\alpha \\ \theta+\gamma \end{bmatrix}_q |GL(\theta+\gamma, q)|$  possible choices for  $h$ . Choose an  $(s+\alpha-\theta-\gamma)$ -dimensional subspace  $G''_{s+\alpha-\theta-\gamma}$  of  $E$  such that  $G''_{s+\alpha-\theta-\gamma} \cap \text{sp}\{e_1, \dots, e_m, e_{v-t+1}, \dots, e_v\} = \{0\}$ . By Proposition 1(iii) there are  $\begin{bmatrix} v-m-t \\ s+\alpha-\theta-\gamma \end{bmatrix}_q q^{(s+\alpha-\gamma-\theta)(m+t)}$  such  $G''_{s+\alpha-\theta-\gamma}$ , and any of the  $|GL(s+\alpha-\theta-\gamma, q)|$  transformations from a fixed complement of  $F'_{\theta+\gamma}$  in  $Z'_{s+\alpha}$  to  $G''_{s+\alpha-\theta-\gamma}$  will complete the definition of  $h$  on  $Z'_{s+\alpha}$ .

Extending  $h$  to  $E$  there are  $q^{(v-n-\alpha-s)(n+\alpha+s)} |GL(v-n-\alpha-s, q)|$  choices. By collecting terms and summing on  $\theta$  Lemma 2 is established.  $\square$

The  $q$ -Hahn polynomials in the product formula (2.2) are normalized

by  $\|cQ_k\|_2^2 = \|\phi_{HK}\|_2^2 = (\dim V_k)^{-1}$ . Using  $\dim V_k = \begin{bmatrix} v \\ k \end{bmatrix}_q - \begin{bmatrix} v \\ k-1 \end{bmatrix}_q$ ,  $|\Delta_{x,k}| = \begin{bmatrix} m \\ n-x \end{bmatrix}_q \begin{bmatrix} v-m \\ x \end{bmatrix}_q q^{(m-n+x)x}$ , and the orthogonality relation (2.6) with  $a = q^{m-v-1}$ ,  $b = q^{-m-1}$ , and  $N = n$ , we obtain

$$c\bar{c} = \frac{\begin{bmatrix} n \\ k \end{bmatrix}_q \begin{bmatrix} v-m \\ k \end{bmatrix}_q}{\begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} v-m \\ k \end{bmatrix}_q} q^{(m-n)k}.$$

After changing  $\gamma$  to  $s+\alpha-\gamma-\theta$  and collecting terms in Lemma 1 and Lemma 2 the product formula (2.2) can be stated.

**THEOREM 1.** *Let  $k, s, t, n, m$ , and  $v$  be integers and  $q$  be a prime power. If*



$0 \leq k, s, t \leq n \leq m \leq v - n$ , then

$$\begin{aligned}
 (3.1) \quad & \frac{\begin{bmatrix} n \\ k \end{bmatrix}_q \begin{bmatrix} v - m \\ k \end{bmatrix}_q}{\begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} v - n \\ k \end{bmatrix}_q} q^{(m-n)k} Q_k(q^{-s}; q^{m-v-1}, q^{-m-1}, n; q) Q_k(q^{-t}; q^{m-v-1}, q^{-m-1}, n; q) \\
 &= \sum_{\alpha, \beta, \gamma, \theta} \frac{\begin{bmatrix} m - n \\ \alpha \end{bmatrix}_q \begin{bmatrix} n - t \\ n - s - \alpha - \beta \end{bmatrix}_q}{\begin{bmatrix} m \\ n - s \end{bmatrix}_q \begin{bmatrix} v - n \\ s + \alpha \end{bmatrix}_q} \begin{bmatrix} t \\ \beta \end{bmatrix}_q \begin{bmatrix} m - n + t \\ s + \alpha - \gamma \end{bmatrix}_q \begin{bmatrix} v - m - t \\ \gamma \end{bmatrix}_q \begin{bmatrix} t - \beta \\ \theta \end{bmatrix}_q \begin{bmatrix} s + \alpha - \gamma \\ \theta \end{bmatrix}_q \\
 &\cdot (q; q)_\theta q^{\binom{\theta}{2}} (-1)^\theta q^A Q_k(q^{-\theta - \beta - \gamma}; q^{m-v-1}, q^{-m-1}, n; q),
 \end{aligned}$$

where  $A = (s + \alpha)(2\beta - t + \alpha - \gamma) + \beta(\beta - t) - \gamma(\beta - \gamma - 2t + n - m)$ .

**COROLLARY.** The product formula (3.1) holds if  $0 \leq k, s, t \leq n$  are integers,  $q$  is any complex number  $|q| \neq 1$ , and  $q^{m-v-1}$  and  $q^{-m-1}$  are replaced by complex numbers  $a$  and  $b$  respectively,  $a, b \neq q^{-l}$ ,  $ab \neq q^{-n-1-l}$ ,  $l = 1, 2, \dots, n$ .

*Proof.* Both sides of (3.1) are rational functions of  $q, a$ , and  $b$ . We have equality for infinitely many values of  $q, a$ , and  $b$ ; and the stated conditions on  $q, a$ , and  $b$  are sufficient to avoid the poles. The sums remain finite since  $s, t$ , and  $n$  are integers.  $\square$

Theorem 1 implies product formulas for various limiting cases of  $q$ -Hahn polynomials. Its major importance is that it is the first term of an addition formula for  $q$ -Hahn polynomials. For  $m = n$  this is a result of Dunkl [7]. As  $q \rightarrow 1$  the product formula for Hahn polynomials is obtained [5], [6].

We could use the positivity of the kernel in Theorem 1 for  $q > 1$  and either  $0 < a, b < q^{-2n}$  or  $a, b = q^{-n-l}$ ,  $l = 1, 2, \dots, n$ , to obtain a bound for the  $q$ -Hahn polynomials. However, a sharp bound can be derived from (2.7) with  $a = q^{-x}$ ,  $b = abq^{k+1}$ ,  $c = q^{-n}$ , and  $d = aq$ . The result is

$$(3.2) \quad |Q_k(q^{-x}; a, b, n; q)| \leq \max_{0 \leq j \leq k} \left| \frac{(bq^{-k}; q^{-1})_j}{(1/aq; q^{-1})_j} \right|, \quad x = 0, 1, \dots, n, q > 0.$$

If  $0 < a, b < q^{-n}$  and  $q > 1$  (3.2) implies

$$(3.3) \quad |Q_k(q^{-x}; a, b, n; q)| \leq \max \{1, |Q_k(q^{-n}; a, b, n; q)|\}, \quad x = 0, 1, \dots, n,$$

and thus for  $0 < a \leq b < q^{-n}$  and  $q > 1$

$$(3.4) \quad |Q_k(q^{-x}; a, b, n; q)| \leq 1, \quad x = 0, 1, \dots, n.$$

By using (2.8) and (2.9) the bound (3.3) gives a bound for any of the  $q$ -Hahn polynomials for  $q > 0$ . For  $q = 1$ , see Dunkl [4] and Gasper [8].

The analytic continuation can be done in another way to obtain a linearization formula for some dual  $q$ -Hahn polynomials. George Gasper has pointed out to the author that by iterating (2.7) to obtain

$$\frac{\begin{bmatrix} n \\ k \end{bmatrix}_q \begin{bmatrix} v - m \\ k \end{bmatrix}_q}{\begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} v - n \\ k \end{bmatrix}_q} q^{(m-n)k} {}_3\phi_2 \left( \begin{matrix} q^{-k} & q^{k-v-1} & q^{-s} \\ & q^{m-v} & q^{-n} \end{matrix} \middle| q; q \right) = {}_3\phi_2 \left( \begin{matrix} q^{-k} & q^{k-v-1} & q^{n-m-s} \\ & q^{n-v} & q^{-m} \end{matrix} \middle| q; q \right),$$

the  $k$ -dependence in (3.1) lies entirely in the polynomials. To retain the dual polynomials if  $m - n = j$ ,  $s$ , and  $t$  remain nonnegative integers, we can replace  $q^m$  and  $q^v$  by

$A$  and  $B$  if  $A \neq q^l$ ,  $B/A \neq q^{-j+l}$ ,  $l = 0, 1, \dots, s+j-1$ . We can replace  $q^k$  by  $x$  if  $B/A \neq q^{t+l}$ ,  $l = 0, 1, \dots, s+j-1$ . Then (3.1) can be interpreted as a linearization or a mixed linearization formula for dual  ${}_3\phi_2$  polynomials with sufficient conditions on  $A, B, q$ , and  $j$  to make the coefficients positive.

The product formula (2.1) follows from Theorem 1 with  $m = n$ . We now state the mixed product formulas (2.4) and (2.3).

**THEOREM 2.** *Let  $k, s, t, n, m$ , and  $v$  be integers and  $q$  be a prime power. If  $0 \leq k, s, t \leq n \leq m \leq n - v$ , then*

$$\begin{aligned}
 & Q_k(q^{-s}; q^{n-v-1}, q^{-n-1}, n; q) Q_k(q^{-t}; q^{m-v-1}, q^{-m-1}, n; q) \\
 (3.5) \quad &= \sum_{\alpha, \beta, \gamma} \frac{\begin{bmatrix} n-t \\ n-s-\alpha \end{bmatrix}_q \begin{bmatrix} t \\ \alpha \end{bmatrix}_q \begin{bmatrix} m-n+t \\ s-\beta \end{bmatrix}_q}{\begin{bmatrix} n \\ s \end{bmatrix}_q \begin{bmatrix} v-n \\ s \end{bmatrix}_q} \begin{bmatrix} v-m-t \\ \beta \end{bmatrix}_q \begin{bmatrix} t-\alpha \\ \gamma \end{bmatrix}_q \begin{bmatrix} s-\beta \\ \gamma \end{bmatrix}_q (q; q)_\gamma (-)^\gamma \\
 & \cdot q^{\binom{2}{2}} q^A Q_k(q^{-\alpha-\beta-\gamma}; q^{m-v-1}, q^{-m-1}, n; q),
 \end{aligned}$$

where  $A = (\alpha + s - \beta)(s + \alpha - t) + \beta(m + t - n + \beta) - s^2$ .

$$\begin{aligned}
 & \frac{\begin{bmatrix} n \\ k \end{bmatrix}_q \begin{bmatrix} v-m \\ k \end{bmatrix}_q}{\begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} v-n \\ k \end{bmatrix}_q} q^{(m-n)k} Q_k(q^{-s}; q^{m-v-1}, q^{-m-1}, n; q) Q_k(q^{-t}; q^{m-v-1}, q^{-m-1}, n; q) \\
 (3.6) \quad &= \sum_{\alpha, \beta, \gamma} \frac{\begin{bmatrix} n-t \\ n-s-\alpha \end{bmatrix}_q \begin{bmatrix} m-n+t \\ \alpha \end{bmatrix}_q}{\begin{bmatrix} m \\ n-s \end{bmatrix}_q \begin{bmatrix} v-m \\ s \end{bmatrix}_q} \begin{bmatrix} t \\ s-\beta \end{bmatrix}_q \begin{bmatrix} v-m-t \\ \beta \end{bmatrix}_q \begin{bmatrix} m-n+t-\alpha \\ \gamma \end{bmatrix}_q \begin{bmatrix} s-\beta \\ \gamma \end{bmatrix}_q \\
 & \cdot (q; q)_\gamma (-)^\gamma q^{\binom{2}{2}} q^A Q_k(q^{-\alpha-\beta-\gamma}; q^{n-v-1}, q^{-n-1}, n; q),
 \end{aligned}$$

where  $A = (\alpha + s - \beta)(s + \alpha - t) + \beta(m - n + t + \beta) - s(m - n + s)$ .

As in (3.1) we can analytically continue  $q^n, q^m, q^v$ , and  $q^k$  in (3.5) and transform (3.6) to obtain a linearization formula for dual polynomials.

Finally we mention a family of  $q$ -Krawtchouk polynomials obtained from the  $q$ -Hahn polynomials by letting  $a \rightarrow 0, b \rightarrow \infty$ , and  $ab = -c$ . For various values of  $c$  these are spherical functions on the other infinite families of Chevalley groups over a finite field. The analogous geometry and product formulas for these polynomials will be described in a forthcoming paper.

**Acknowledgment.** The author would particularly like to thank Professor Dunkl for access to his preprints.

REFERENCES

[1] G. ANDREWS, *Applications of basic hypergeometric functions*, SIAM Rev., 16 (1974), pp. 441-484.  
 [2] G. ANDREWS AND R. ASKEY, *The classical orthogonal polynomials and their discrete and  $q$ -analogues*, to appear.  
 [3] PH. DELSARTE, *Hahn polynomials, discrete harmonics, and  $t$ -designs*, MBLÉ, Report 295, April 1975.  
 [4] C. DUNKL, *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups*, Indiana Univ. Math. J., 25 (1976), pp. 335-358.  
 [5] ———, *An addition theorem for Hahn polynomials: the spherical functions*, this Journal, 9 (1978), pp. 627-637.

- [6] C. DUNKL, *Spherical functions on compact groups and applications to special functions*, Symposia Mathematica, 22 (1977), pp. 145–161.
- [7] ———, *An addition theorem for some  $q$ -Hahn polynomials*, Monatsh. Math., 85 (1978), pp. 5–37.
- [8] G. GASPER, *Positivity and special functions*, Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 375–433.
- [9] W. HAHN, *Über Orthogonalpolynome, die  $q$ -Differenzgleichungen genügen*, Math. Nachr., 2 (1949), pp. 4–34.
- [10] R. STEINBERG, *A geometric approach to the representations of the full linear group over a Galois field*, Trans. Amer. Math. Soc., 71 (1951), pp. 274–282.

## A NEUTRAL FUNCTIONAL DIFFERENTIAL EQUATION OF LURIE TYPE\*

E. N. CHUKWU†

**Abstract.** The problem of Lurie is posed for systems described by a functional differential equation of neutral type. Sufficient conditions are obtained for absolute stability for the controlled system if it is assumed that the uncontrolled plant equation is uniformly asymptotically stable. Both the direct and indirect control cases are treated.

**1. Introduction.** Consider a system of real ordinary differential equations

$$(1) \quad \begin{aligned} \frac{dx}{dt} &= Ax + bf(\sigma), \\ \frac{d\sigma}{dt} &= c^T x - rf(\sigma) \end{aligned}$$

in which  $f: (-\infty, \infty) \rightarrow (-\infty, \infty)$  is sectionally continuous with  $\sigma f(\sigma) > 0$  for  $\sigma \neq 0$ ,  $f(0) = 0$ ,  $A$  is an  $n \times n$  matrix,  $c$  and  $b$  are constant  $n$ -vectors and  $r$  is a scalar. The problem of Lurie consists of finding a necessary and sufficient condition for every solution  $(\sigma(t), x(t))$  of (1) to tend to  $(0, 0)$  as  $t \rightarrow \infty$  whenever it is assumed that the uncontrolled equation

$$(2) \quad \frac{dx}{dt} = Ax$$

is uniformly asymptotically stable in the large (cf. [1, p. 9]). The entire monograph by Lefschetz was devoted to this problem. Recently, Somolinos [2] has generalized this problem of Lurie to functional differential equation of retarded type. In this paper we shall treat the problem of Lurie when the system is described by functional differential equation of neutral type. We shall assume that the uncontrolled system is uniformly asymptotically stable. Utilizing a theorem of Cruz and Hale in [3] which ensures the existence of a Liapunov functional, we then obtain conditions for the uniform asymptotic stability of the feedback system.

**2. Notations and preliminary results.** Let  $E^n$  be a real  $n$ -dimensional Euclidean vector space with norm  $|\cdot|$ . Let  $h \geq 0$  be a given real number. Let  $C$  be the space  $C([-h, 0], E^n)$  of continuous functions taking  $[-h, 0]$  into  $E^n$  with  $\|\phi\|$ ,  $\phi \in C$  defined by  $\|\phi\| = \sup \{|\phi(\theta)|: -h \leq \theta \leq 0\}$ . For any continuous function  $x(\theta)$  on  $-h \leq \theta \leq t_1$ ,  $t_1 > 0$  and a fixed  $t$ ,  $0 \leq t \leq t_1$ ,  $x_t$  denotes the function  $x_t(\theta) = x(t + \theta)$ ,  $-h \leq \theta \leq 0$ . Let  $D(\cdot): [t_0, \infty) \times C \rightarrow E^n$  be a continuous function defined by

$$(3) \quad D(t)\phi = \phi(0) - g(t, \phi), \quad \text{for } t \in [t_0, \infty) \equiv I, \quad \phi \in C,$$

where

$$g: [t_0, \infty) \times C \rightarrow E^n,$$

---

\* Received by the editors August 18, 1977 and in revised form June 20, 1978. This research was supported by National Aeronautics and Space Administration under Contract NSG 1445.

† Cleveland State University, Cleveland, Ohio 44115, U.S.A. Now on leave of absence as Reader, University of Jos, Jos, Nigeria.

is continuous,  $g(t, \phi)$  is linear in  $\phi$  and is given by

$$(4) \quad g(t, \phi) = \int_{-h}^0 [d_s \mu(t, s)] \phi(s).$$

The function  $\mu(t, s)$  is an  $n \times n$  matrix  $t \in I, s \in [-h, 0]$ , with elements of bounded variation in  $s$  which satisfy the following condition:

$$(5) \quad \left| \int_{-\theta}^0 [d_s \mu(t, s)] \phi(s) \right| \leq l(\theta) \sup_{-\theta \leq r \leq 0} |\phi(r)|,$$

for all  $t \in I, \phi \in C$ , where  $l$  is continuous nondecreasing for  $\theta \in [0, h], l(0) = 0$ .

Let  $A: I \times C \rightarrow E^n$  be continuous and consider the equation

$$(6) \quad \begin{aligned} \frac{d}{dt}(D(t)x_t) &= A(t, x_t), \\ x_{t_0} &= \phi, \quad t_0 \in I. \end{aligned}$$

The following theorem ensures the existence of a Liapunov functional when (6) is uniformly asymptotically stable.

**THEOREM 2.1** [3]. *Let  $D(t)$  and  $A(t, \cdot)$  be bounded linear operators from  $C$  into  $E^n$  such that for some constant  $L > 0$ , for all  $\phi \in C$ , for all  $t \geq t_0$ ,*

$$|D(t)\phi| \leq L\|\phi\|.$$

*If (6) is uniformly asymptotically stable, then there exist positive constants  $M, \alpha$  and a continuous scalar function  $V$  on  $I \times C$  such that*

$$(7) \quad \begin{aligned} (i) \quad &|D(t)\phi| \leq V(t, \phi) \leq M\|\phi\|, \\ (ii) \quad &\dot{V}(t, \phi) \leq -\alpha V(t, \phi), \\ (iii) \quad &|V(t, \phi) - V(t, \psi)| \leq K\|\phi - \psi\|, \end{aligned}$$

*for all  $t \geq t_0, \phi, \psi \in C; \dot{V}$  is the usual upper right hand derivate along the solutions of (6).*

In Theorem 2.1 it is assumed that  $D(t)$  and  $A(t, \cdot)$  are linear. However, Cruz and Hale [3] stated a similar result when  $A(t, \phi)$  is not linear in  $\phi$ , but  $g(t, \phi)$  in (3) satisfies

$$|g(t, \phi)| \leq L\|\phi\|, \quad \text{for all } t \geq t_0.$$

We now state the result and point out the required lemma needed to carry out the proof in [3]. It was communicated to the author by Professor J. K. Hale.

**THEOREM 2.2.** *Let  $A(t, 0) = 0$ , and let  $A(t, \phi)$  be uniformly locally Lipschitz in  $\phi$  uniformly with respect to  $t$ , with Lipschitz constant  $N$ . Let  $D$  satisfy locally the condition*

$$|D(t)\phi| \leq K\|\phi\|,$$

*for all  $t \geq t_0$ , for some  $K$ .*

Assume that the null solution of (6) is uniformly asymptotically stable. Then there exist a  $S_0 > 0$ , a  $M = M(S_0) > 0$ , positive definite functions  $b(u), c(u)$ , on  $0 \leq u \leq S_0$  and a scalar function  $V(t, \phi)$  defined and continuous for  $t \in I \times C, \|\phi\| \leq S_0$  such that

$$(a) \quad |D(t)\phi| \leq V(t, \phi) \leq b(\|\phi\|)$$

$$(b) \quad \dot{V}(t, \phi) \leq -c(|D(t)\phi|)$$

$$(c) \quad |V(t, \phi_1) - V(t, \phi_2)| \leq M\|\phi_1 - \phi_2\|$$

for all  $t \geq t_0, \phi_1, \phi_2 \in C, \|\phi_i\| \leq S_0, i = 1, 2$ . The condition (b) can be replaced by

$$(b') \quad \dot{V}(t, \phi) \leq -\beta V(t, \phi), \beta > 0.$$

*Remark.* The problem with the proof of Theorem 7.2 in [3] is contained in verifying (c). The following lemma is needed.

**LEMMA (Hale).** *In (6) assume that  $D$  satisfies the conditions of Theorem 2.2. Then for any  $r_0 > 0$ , there is a constant  $L = L(r_0)$  such that*

$$\|x_t(t_0, \phi_1) - x_t(t_0, \phi_2)\| \leq e^{L(t-t_0)} \|\phi_1 - \phi_2\|$$

for all  $t \geq t_0, \phi_1, \phi_2$  for which

$$\|x_t(t_0, \phi_1)\| \leq r_0, \quad \|x_t(t_0, \phi_2)\| \leq r_0.$$

*Remark.* The proof is not as easy as for retarded equations since one cannot apply the Gronwall inequality directly. One must take small steps in time and make careful estimates using properties of  $D(t)$ .

To prove Theorem 2.2, set

$$V(t, \phi) = \sup_{s \geq 0} |D(t+s)x_{t+s}(t_0, \phi)| e^{a(t)},$$

and proceed as Hale [4, p. 310]. Our lemma replaces the inequality on page 310, bottom line.

The first case considered is the indirect control system

$$\begin{aligned} \frac{d}{dt}(D(t)x_t) &= A(t, x_t) + bf(\sigma), \quad t \geq t_0, \\ (8) \quad \sigma(t) &= B(t, x(t)) - rf(\sigma), \\ x_{t_0} &= \phi, \quad t_0 \in I, \end{aligned}$$

in which  $A$  is as above,  $B(t, y)$  is a scalar continuous function in  $t \geq 0, y \in E^n$ , and  $f$  is a scalar function which is continuous.

**DEFINITION.** The operator  $D$  in (3) is uniformly stable if there are constants  $\alpha > 0, \beta > 0$  such that the solution of the "difference equations"

$$D(t)x_t = 0, \quad x_{t_0} = \phi, \quad D(t_0)\phi = 0,$$

satisfies  $\|x_t\| \leq \beta e^{-\alpha(t-t_0)} \|\phi\|, t \geq t_0$ .

**3. Main theorems.**

**THEOREM 3.1.** *Assume that in (8) the uncontrolled system (6) is uniformly asymptotically stable. Let  $\alpha$  and  $K$  be as given by Theorem 2.1. Assume that  $A(t, \cdot)$  and  $D(t)$  are bounded linear operators from  $C$  into  $E^n$  such that  $|D(t)\phi| \leq M\|\phi\|$  for all  $t \geq t_0, \phi \in C$ . Assume that:*

$$(i) \quad \int_0^\sigma f(s) ds \rightarrow 0, \quad \text{as } |\sigma| \rightarrow \infty;$$

there exists a positive constant  $c$  such that

$$(ii) \quad |B(t, x(t))| \leq c(|D(t)x_t|)$$

for all  $t \in I$ , where  $x$  is continuous;

(iii) for all  $\theta \in [0, h]$  the relation

$$4\alpha r > \left( c + \frac{K|b|}{1-l(\theta)} \right)^2,$$

holds where  $l$  is defined in (5);

(iv) the operator  $D$  is uniformly stable.

Then (8) is uniformly asymptotically stable.

*Proof.* Since (6) is uniformly asymptotically stable, there exists a Liapunov functional for (6) given by Theorem 2.1. Let  $\dot{V}_{(8)}$  denote the derivative of  $V$  along the solutions of (8). Let  $y = y(t_0, \phi)$ ,  $x = x(t_0, \phi)$  be the solutions of (8) and (6) respectively, then the relations (7) imply that

$$(9) \quad \dot{V}_{(8)}(t, \phi) \leq \dot{V}_{(6)}(t, \phi) + K \lim_{h \rightarrow 0} \frac{1}{h} |y_{t+h}(t, \phi) - x_{t+h}(t, \phi)|.$$

But then

$$D(t+h)(y_{t+h} - x_{t+h}) = \int_t^{t+h} bf(\sigma) ds,$$

for any  $h \geq 0$ . Since  $g$  satisfies (5) we have that there exists an  $h_0 > 0$  such that

$$|y_{t+h} - x_{t+h}| \leq \frac{1}{1-l(h_0)} \int_t^{t+h} |bf(\sigma)| ds,$$

for  $0 \leq h \leq h_0$ . We now use this inequality in (9) to obtain

$$(10) \quad \dot{V}_{(8)}(t, \phi) \leq \dot{V}_{(6)}(t, \phi) + \frac{K}{1-l(h_0)} |bf(\sigma)|.$$

Hence, by (7(ii))

$$\dot{V}_{(8)}(t, \phi) \leq -\alpha V + \frac{K}{1-l(h_0)} |bf(\sigma)|.$$

Define  $W = V^2/2 + \int_0^\sigma f(s) ds$ . The derivative of  $W$  along the solutions of (8) satisfies

$$\dot{W} \leq -\alpha V^2 - r|f(\sigma)|^2 + V \left( \frac{K}{1-l(h_0)} |bf(\sigma)| \right) + |f(\sigma)B|.$$

By conditions (ii) of Theorem 3.1 and (i) of Theorem 2.1 we obtain from this that

$$(11) \quad \dot{W} \leq -\alpha V^2 - r|f(\sigma)|^2 + V \left( \frac{K|b|}{1-l(h_0)} + c \right) |f(\sigma)|.$$

The right hand side of (11) is a quadratic form in  $V$  and  $|f(\sigma)|$ . It is obviously negative definite by condition (iii). Hence, there exists a positive number  $\gamma$  such that

$$\dot{W} \leq -\gamma(V^2 + |f(\sigma)|^2).$$

From this it follows that

$$\dot{W} \leq -\gamma|D(t, \phi)|^2,$$

so that the second condition of (4.2) in Theorem 4.1 of Cruz and Hale [3] is met for the Liapunov function  $W$ . Trivially, also the other conditions in (4.2) are satisfied. Because  $D$  is a uniformly stable operator the operator  $\bar{D}$  given by

$$\bar{D}\psi = \bar{\psi}(0) - \bar{g}(t, \bar{\psi}),$$

where

$$\bar{\psi} = \begin{bmatrix} \psi \\ \sigma \end{bmatrix}, \quad \bar{g} = \begin{bmatrix} g \\ 0 \end{bmatrix}$$

is uniformly stable. Therefore, by Theorem 4.1 of [3] the system

$$\frac{d}{dt}(\bar{D}(t)y_t) = \bar{g}(t, y_t)$$

is uniformly asymptotically stable. Here

$$\bar{D}(t)y_t = \begin{pmatrix} D(t)x_t \\ \sigma \end{pmatrix},$$

$$\bar{g}(t, y_t) = \begin{bmatrix} A(t, x_t) + bf(\sigma) \\ B(t, x(t) - rf(\sigma)) \end{bmatrix}.$$

The proof is complete.

**THEOREM 3.2.** Consider (8), and assume that  $A(t, 0) = 0$ ,  $A(t, \phi)$  is locally Lipschitz in  $\phi$  uniformly with respect to  $t$ , and the operator  $D$  satisfies

$$|D(t)\phi| \leq M\|\phi\|,$$

locally in  $\phi \in C$ , for all  $t \geq t_0$  and some  $M$ . Assume that  $D$  is uniformly stable and that (6) is uniformly asymptotically stable. Let  $K$  and  $\beta$  be as given by Theorem 2.2 and assume that

(i) For all  $\theta \in [0, h]$  the relation

$$4\beta r > \left( c + \frac{K|b|}{1-l(\theta)} \right)^2$$

holds where  $l$  is defined in (5), and where  $c$  is a constant such that

(ii)  $|B(t, x(t))| \leq c|D(t)x_t|$

for all continuous  $x$  and all  $t \in I$ .

(iii)  $\int_0^\sigma f(s) ds \rightarrow \infty$ , as  $|\sigma| \rightarrow \infty$ .

Then there exists a  $\delta_0 > 0$ , such that for any  $\varepsilon$ ,  $0 < \varepsilon < \delta_0$  and any  $t_0 \geq 0$ , there is a  $\delta = \delta(\varepsilon)$  such that  $\|\sigma\| < \delta$  implies  $\|x_t(t_0, \phi)\| < \varepsilon$  for all  $t \in [t_0, \infty)$ ; and for any  $\eta > 0$ ,  $0 \leq \eta \leq \delta_0$ , there exists a  $T(\eta) > 0$ , such that  $\|\phi\| \leq \delta$ , implies  $\|x_t(t_0, \phi)\| \leq \eta$ , if  $t \geq t_0 + T(\eta)$ . In other words all solutions in the ball  $S(\delta_0) \subseteq C$  are uniformly asymptotically stable.

*Proof.* The hypotheses of the theorem imply there is a Liapunov functional  $V$  satisfying the conditions of Theorem 2.2. Choose  $\delta_0$  as in Theorem 2.2. Let  $\dot{V}_{(8)}$  denote the derivative of  $V$  along solutions of (8). If  $y = y(t_0, \phi)$ ,  $x = x(t_0, \phi)$  are the solutions of (8) and (6) respectively, then, as before,

$$\dot{V}_{(8)}(t, \phi) \leq \dot{V}_{(6)}(t, \phi) + \frac{K}{1-l(h_0)}|bf(\sigma)|,$$

provided  $\|\phi\| \leq \delta_0$ . On using

$$W = \frac{V^2}{2} + \int_0^\sigma f(s) ds,$$

one easily verifies that the conditions of Theorem 4.1 of [3] are satisfied for  $W$ , provided  $\|\phi\| \leq \delta_0$ . By the cited theorem the trivial solution of (8) is uniformly asymptotically stable when confined to the ball  $S(\delta_0) \subseteq C$ .



Consider the direct control case:

$$(12) \quad \begin{aligned} \frac{d}{dt}(D(t)x_t) &= A(t, x_t) + bf(\sigma), & \sigma &= c^T D(t)x_t, \\ \frac{d}{dt}(D(t)x_t) &= B(t, x_t) + bf(\sigma), & x_{t_0} &= \sigma, \end{aligned}$$

where the stable atomic operator  $\bar{D}$  is atomic at 0, and where the letters are there defined above and  $c^T b = -r < 0$ .

**THEOREM 3.3.** *Assume that  $D(t)$  and  $A(t, \cdot)$  are bounded linear operators from  $C$  into  $E^n$  such that*

$$|D(t)\phi| \leq L\|\phi\|$$

for all  $t \geq t_0$ ,  $\phi \in C$ , and

$$(13) \quad |B(t, \phi)| \leq \beta |D(t)\phi|, \quad \beta > 0.$$

Suppose (6) is uniformly asymptotically stable and

(i)  $f(0) = 0$ ,  $\sigma f(\sigma) > 0$ ,  $\sigma \neq 0$ ,  $f$  continuous and

$$\int_0^\sigma f(s) ds \rightarrow \infty, \quad \text{as } |\sigma| \rightarrow \infty.$$

(ii) Let  $\alpha$  and  $K$  be given by Theorem 2.1 and let the relation

$$(14) \quad 5\alpha r > \left( \frac{4|b|}{1-l(s)} + |c^T \beta| \right)^2,$$

hold for all  $s \in [0, h]$ , where  $l$  is defined in (5).

Then (12) is uniformly asymptotically stable.

*Proof.* Proceed as before, using Theorem 2.1 to obtain a Liapunov functional  $V$  for the system (6). Differentiating  $V$  along solutions of (12) yields

$$\dot{V}_{(12)}(t, \phi) \leq \dot{V}_{(6)}(t, \phi) + \frac{K}{1-l(h_0)} |b| |f(\sigma)|.$$

Set

$$W = \frac{V^2}{2} + \int_0^\sigma d(s) ds.$$

Then

$$\begin{aligned} W_{(12)} &\leq -\alpha V^2 + \frac{Vk|bf(\sigma)|}{1-l(h_0)} + f(\sigma)\dot{\sigma} \\ &\leq -\alpha V^2 - r|f(\sigma)|^2 + V \left[ \frac{k|b|}{1-l(h_0)} + \beta |c^T| \right] \end{aligned}$$

where we have used (13) and the property of  $V$ . We now use (14) to deduce the result as before.

## REFERENCES

- [1] S. LEFSCHETZ, *Stability of nonlinear control systems*, Mathematics in Science and Engineering, vol. 13, Academic Press, New York, 1965.
- [2] A. SOMOLINOS, *A generalization of the problem of Lurie to functional equations*, to appear.
- [3] M. A. CRUZ AND J. K. HALE, *Stability of functional differential equations of neutral type*, J. Differential Equations, 7 (1970), pp. 332–355.
- [4] J. K. HALE, *Ordinary Differential Equations*, Wiley–Interscience, New York, 1969.

## A NOTE ON MULTIPLE ASYMPTOTIC SERIES\*

R. D. GREGORY†

**Abstract.** There has appeared in the literature [K. D. Shere, *Introduction to multiple asymptotic series with an application to elastic scattering* an attempt to extend the concept of asymptotic series to “multiple asymptotic series” of the form

$$\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} \frac{e^{-\lambda mx}}{x^n}.$$

Applications referring to this work have also appeared in W. Bühring [J. Mathematical Phys., 18 (1977), pp. 1121–1136], W. Bühring [Angew. Math. Mech., 57 (1977), T226–T227], K. D. Shere [SIAM J. Math. Anal., 3 (1972), pp. 263–271], and K. D. Shere [SIAM J. Math. Anal., 3 (1972), pp. 272–284]. It is the purpose of this note to show that the definition of a “multiple asymptotic series” given in K. D. Shere [J. Mathematical Phys., 12 (1971), pp. 78–82] is unsound in the sense that it fails to satisfy certain basic criteria (for instance when the series involved are convergent then they are not necessarily asymptotic, according to this definition); also even though uncountably many alternative definitions exist which overcome these difficulties, each is quite arbitrary and has little practical value.

**1. The definition of multiple asymptotic series given in [5].** Suppose that we wish to give a meaning to the formal expansion

$$(1.1) \quad F(x) \approx \sum_{n=0}^{\infty} \frac{a_n}{x^n} + e^{-x} \sum_{n=0}^{\infty} \frac{b_n}{x^n} \quad \text{as } x \rightarrow \infty.$$

(Actually the work in [5] is carried through in greater generality, but there is no need for this in the present note.)

If

$$F(x) \sim \sum_{n=0}^{\infty} \frac{a_n}{x^n} \quad \text{as } x \rightarrow \infty,$$

(where here and elsewhere the  $\sim$  symbol is used in the normal Poincaré sense) and this series is *convergent* for  $x$  sufficiently large, then there is no difficulty. We would merely regard (1.1) as meaning

$$(1.2) \quad e^x \left\{ F(x) - \sum_{n=0}^{\infty} \frac{a_n}{x^n} \right\} \sim \sum_{n=0}^{\infty} \frac{b_n}{x^n} \quad \text{as } x \rightarrow \infty.$$

However, if the series  $\sum_{n=0}^{\infty} a_n x^{-n}$  is not convergent but only *asymptotic* as  $x \rightarrow \infty$ , the expression on the left in (1.2) is without meaning. The device suggested in [5] is to replace the divergent expression  $\sum_{n=0}^{\infty} a_n x^{-n}$  in (1.2) by a constructed function  $F^*(x)$  which is known to be asymptotic to

$$\sum_{n=0}^{\infty} \frac{a_n}{x^n} \quad \text{as } x \rightarrow \infty.$$

To be precise (for the case in which  $|a_n| \geq 1, \forall n$ ), the choice of  $F^*(x)$  given in [5] is

$$(1.3) \quad F^*(x) = \sum_{n=0}^{\infty} \frac{a_n}{x^n} \left[ 1 - \exp \left\{ -\frac{1}{2} \frac{x^2}{|a_n| n!} \right\} \right],$$

a construction due to Ritt [3].

The series in (1.3) is convergent and

$$(1.4) \quad F^*(x) \sim \sum_{n=0}^{\infty} \frac{a_n}{x^n} \quad \text{as } x \rightarrow \infty.$$

\* Received by the editors August 23, 1978, and in revised form January 4, 1979.

† Department of Mathematics, University of Manchester, Manchester, England M13 9PL. On leave of absence at the University of British Columbia, Vancouver, B.C. V6T 1W5 Canada, during 1977–79. This work was supported in part by N.R.C. under Grants A9259 and A9117.

In terms of this  $F^*$ , the definition of the multiple asymptotic series (M.A.S.) (1.1), as given in [5], is that

$$(1.5) \quad e^x[F(x) - F^*(x)] \sim \sum_{n=0}^{\infty} \frac{b_n}{x^n} \quad \text{as } x \rightarrow \infty.$$

This definition does however have very serious disadvantages, namely:

- (i) if  $F(x)$  is such that the series in (1.1) are both convergent, then  $F(x)$  may *not* possess an M.A.S. in the above sense;
- (ii) the functions which do possess an M.A.S. in this sense do not form a linear space.

To show (i), consider the function

$$(1.6) \quad F(x) = \frac{x}{x-1} + e^{-x} = \sum_{n=0}^{\infty} \frac{1}{x^n} + e^{-x} \quad (x > 1).$$

Then  $F^*$ , as given by (1.3) is

$$(1.7) \quad F^*(x) = \sum_{n=0}^{\infty} \frac{1}{x^n} \left[ 1 - \exp \left\{ -\frac{1}{2} \frac{x^2}{n!} \right\} \right]$$

and so

$$(1.8) \quad e^x \{F(x) - F^*(x)\} = \sum_{n=0}^{\infty} \frac{e^x}{x^n} \exp \left\{ -\frac{1}{2} \frac{x^2}{n!} \right\} + 1.$$

Now let  $x \rightarrow \infty$  through the sequence of values  $X \equiv (x_m)$ , where  $x_m = (m!)^{1/2}$ ,  $m \geq 0$ . Then

$$(1.9) \quad \begin{aligned} A_m &\equiv e^{x_m} \{F(x_m) - F^*(x_m)\} = 1 + \sum_{n=0}^{\infty} \exp \left( -\frac{1}{2} \frac{x_m^2}{n!} \right) \frac{e^{x_m}}{x_m^n} \\ &> \exp \left( -\frac{1}{2} \frac{x_m^2}{m!} \right) \frac{e^{x_m}}{x_m^m} = e^{-1/2} \frac{e^{(m!)^{1/2}}}{(m!)^{m/2}}. \end{aligned}$$

So

$$(1.10) \quad \log A_m > (m!)^{1/2} - \frac{1}{2} m \log(m!) - \frac{1}{2} \rightarrow \infty \quad \text{as } m \rightarrow \infty.$$

Hence

$$(1.11) \quad e^x \{F(x) - F^*(x)\} \rightarrow \infty$$

as  $x \rightarrow \infty$  through  $X$ , and so no choice of the  $(b_n)$  may be made to satisfy (1.5). Thus no M.A.S. exists for the function (1.6).

To show (ii), consider the function

$$(1.12) \quad G(x) = \sum_{n=0}^{\infty} \frac{n!}{x^n} \left[ 1 - \exp \left\{ -\frac{x^2}{2(n!)^2} \right\} \right] + e^{-x}.$$

Since

$$(1.13) \quad G(x) \sim \sum_{n=0}^{\infty} \frac{n!}{x^n} \quad \text{as } x \rightarrow \infty,$$

it follows from the construction (1.3) that

$$(1.14) \quad G^*(x) = \sum_{n=0}^{\infty} \frac{n!}{x^n} \left[ 1 - \exp \left\{ -\frac{x^2}{2(n!)^2} \right\} \right]$$

and hence from the definition (1.5) that  $G(x)$  has the M.A.S.

$$(1.15) \quad G(x) \approx \sum_{n=0}^{\infty} \frac{n!}{x^n} + e^{-x}.$$

But the function  $H(x) \equiv 2G(x)$  has no M.A.S., since

$$(1.16) \quad e^x \{H(x) - H^*(x)\} = \sum_{n=0}^{\infty} \frac{2n!e^x}{x^n} \left[ \exp \left\{ -\frac{x^2}{4(n!)^2} \right\} - \exp \left\{ -\frac{x^2}{2(n!)^2} \right\} \right] + 2$$

$$(1.17) \quad \rightarrow \infty \quad \text{as } x \rightarrow \infty \text{ through } X,$$

as with counterexample (i).

These two counterexamples contradict Theorems 2.1 and 2.2 in [5]. The error in these theorems is an (apparent) assumption that relations such as

$$(1.18) \quad \exp \left\{ -\frac{1}{2} \frac{x^2}{|a_n|n!} \right\} = o(e^{-x}) \quad \text{as } x \rightarrow \infty,$$

are *uniformly* valid in  $n$ . In particular the statements  $f_m^* \in \mathcal{F}_m^*$  in Theorem 2.1, and  $f_0(x)g_0(x) \in \mathcal{H}_0^*$  in Theorem 2.2 are false.

**2. Alternative definitions of multiple asymptotic series.** It could be argued that the preceding difficulties arise from the fact that (1.3) is a “wrong” choice for  $F^*$ , and that by a “correct” choice of  $F^*$  these difficulties would disappear. This is in fact so, and indeed there are uncountably many such definitions which will avoid the difficulties (i), (ii).

To show that such choices for  $F^*$  exist, proceed as follows:

Let

$$\mathcal{F} = \left\{ F(x); F(x) \sim \sum_{n=0}^{\infty} \frac{a_n}{x^n} \quad \text{as } x \rightarrow \infty, \text{ for some } (a_n) \right\}$$

and define on  $\mathcal{F}$  the equivalence relation

$$F_1(x) \equiv F_2(x) \quad \text{iff} \quad F_1(x) - F_2(x) = O(x^{-n}) \quad \text{as } x \rightarrow \infty, \quad \forall n \geq 0.$$

Let  $T$  be the linear space of equivalence classes generated on  $\mathcal{F}$  by  $\equiv$  (with the linear operations defined in the obvious way, as with the  $\mathcal{L}_p$ -spaces). Then  $T$  is clearly isomorphic to the linear space  $S$  of all sequences  $(a_n)$ . Thus any definition for  $F^*(x)$  merely consists of selecting a single representative from each of the above equivalence classes. Let  $\hat{S}$  be the subspace of  $S$  whose sequences  $(a_n)$  are such that the series  $\sum_{n=0}^{\infty} a_n x^{-n}$  is convergent for sufficiently large  $x$ , and let  $\hat{T}$  be the corresponding subspace of  $T$ . Take a Hamel basis  $\hat{B}$  of  $\hat{T}$ , and extend it to be a Hamel basis  $B$  of  $T$ . [See Rudin [4], p. 52 for a definition of Hamel basis.]

Now select  $F^*(x)$  as follows:

(a) For the elements of  $\hat{B}$  let

$$F^*(x) = \sum_{n=0}^{\infty} \frac{a_n}{x^n},$$

which is defined for  $x$  sufficiently large.

(b) For the elements of  $B$  not in  $\hat{B}$ , assign  $F^*$  arbitrarily.

(c) For all other elements of  $T$ , construct  $F^*$  by finite linear combinations from (a), (b).

Any of these uncountably many definitions of  $F^*$  will overcome the difficulties (i), (ii).

Unfortunately, each of these definitions is quite arbitrary so that in a problem where  $F(x)$  is unknown, and where we are seeking to determine its asymptotic behavior as  $x \rightarrow \infty$ , one would not know the correct  $F^*$  to choose unless perhaps there were some physical motivation to guide this choice. This is best illustrated by an example. Suppose we have shown that an (unknown)  $F(x)$  is such that

$$(2.1) \quad F(x) \sim \sum_{n=0}^{\infty} \frac{n!}{x^n} \quad \text{as } x \rightarrow \infty,$$

and we wish to proceed on to find an M.A.S. for  $F(x)$ . Suppose we have made the selection of  $F^*$ , corresponding to  $(n!)$ , to be (say)

$$(2.2) \quad F^*(x) = \sum_{n=0}^{\infty} \frac{n!}{x^n} \left[ 1 - \exp \left\{ -\frac{x^2}{(n!)^2} \right\} \right].$$

Then if the unknown function  $F$  were actually

$$(2.3) \quad F(x) = \sum_{n=0}^{\infty} \frac{n!}{x^n} \left[ 1 - \exp \left\{ -\frac{2x^2}{(n!)^2} \right\} \right] + e^{-x},$$

it follows that  $F(x)$  would possess no M.A.S. In short, one only knows the "correct" choice for  $F^*$  when  $F(x)$  is already known correct to  $O(e^{-x})$ .

**3. Comparison with uniform asymptotic expansions.** For functions of the form  $F(x, \varepsilon)$ , where  $\varepsilon \geq 0$  is an additional parameter, the statement

$$(3.1) \quad F(x, \varepsilon) \sim \sum_{n=0}^{\infty} \frac{a_n(\varepsilon)}{x^n} + e^{-\varepsilon x} \sum_{n=0}^{\infty} \frac{b_n(\varepsilon)}{x^n}$$

as  $x \rightarrow \infty$ , uniformly for  $\varepsilon \geq 0$ , certainly has a meaning and implies more information than the nonuniform asymptotic series

$$(3.2) \quad F(x, \varepsilon) \sim \sum_{n=0}^{\infty} \frac{a_n(\varepsilon)}{x^n},$$

valid for each  $\varepsilon > 0$ . The point of the previous sections is that a uniform formula such as (3.1) cannot be deduced by seeking further precision from the nonuniform formula (3.2), unless the series in (3.2) is actually convergent.

**Acknowledgments.** The author wishes to acknowledge his gratitude to Professors Don Ludwig and Fred Wan who made his visit to the University of British Columbia possible, and to the University of Manchester for generously granting leave of absence. Thanks are also due to Professor Frank Olver, who encouraged the author to publish this note and made helpful suggestions to improve the presentation.

#### REFERENCES

- [1] W. BÜHRING, *Schrödinger equation with Yukawa potential, a differential equation with two singular points*, J. Mathematical Phys., 18 (1977), pp. 1121–1136.
- [2] ———, *Das Zusammenhangsproblem bei der Schrödinger-Gleichung mit Yukawa-Potential*, Z. Angew. Math. Mech., 57 (1977), T226–T227.
- [3] J. F. RITT, *On the derivatives of a function at a point*, Ann. of Math., 18, (1916), pp. 18–23.
- [4] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [5] K. D. SHERE, *Introduction to multiple asymptotic series with an application to elastic scattering*, J. Mathematical Phys. 12 (1971), pp. 78–82.
- [6] ———, *Multiple asymptotic expansions and singular problems*, this Journal, 3 (1972), pp. 263–271.
- [7] ———, *On multiple asymptotic expansions*, this Journal, 3, (1972), pp. 272–284.

## STRUCTURE OF RESOLVENTS OF VOLTERRA INTEGRAL AND INTEGRODIFFERENTIAL SYSTEMS\*

G. S. JORDAN† AND ROBERT L. WHEELER‡

**Abstract.** Conditions are given which ensure that the resolvent of a linear Volterra integral or integrodifferential system whose kernel belongs to a weighted  $L^1$  space may be written as a matrix whose entries are finite sums of products of polynomials and exponentials, plus a matrix which belongs to the same weighted  $L^1$  space. These results are obtained from theorems of the same type which we prove for more general linear Volterra–Stieltjes equations. The results are stated in terms of Laplace transform hypotheses and moment conditions.

**1. Introduction.** The integral resolvent  $r(t)$  and differential resolvent  $R(t)$  determined by the equations

$$(1.1) \quad r(t) = B(t) - \int_0^t r(t-s)B(s) ds \quad (t \in \mathbf{R}^+ \equiv [0, \infty)),$$

$$(1.2) \quad R'(t) = R(t)\mathcal{A} + \int_0^t R(t-s)B(s) ds \quad (R(0) = I, t \in \mathbf{R}^+)$$

are associated with the linear Volterra integral and integrodifferential systems

$$(1.3) \quad x(t) = f(t) - \int_0^t x(t-s)B(s) ds \quad (t \in \mathbf{R}^+),$$

$$(1.4) \quad x'(t) = x(t)\mathcal{A} + \int_0^t x(t-s)B(s) ds + f(t) \quad (x(0) = x_0, t \in \mathbf{R}^+),$$

respectively. Here  $r(t)$ ,  $R(t)$ ,  $\mathcal{A}$  and  $B(t)$  are  $n \times n$  matrices,  $I$  is the identity matrix, and  $x(t)$  and  $f(t)$  are row vectors with  $n$  components. Under mild assumptions on  $B(t)$  and  $f(t)$  (see [10, Chap. 4] and [4]), (1.3) and (1.4) are solved by

$$(1.5) \quad x(t) = f(t) - \int_0^t f(t-s)r(s) ds \quad (t \in \mathbf{R}^+),$$

$$(1.6) \quad x(t) = x_0R(t) + \int_0^t f(t-s)R(s) ds \quad (t \in \mathbf{R}^+),$$

respectively.

Let  $B(t) \in L^1(\mathbf{R}^+)$  and let  $\tilde{B}(z) \equiv \int_0^\infty e^{-zt}B(t) dt$  denote the Laplace transform of  $B(t)$ . Then a classical result due to Paley and Wiener [13] is that  $r(t) \in L^1(\mathbf{R}^+)$  if and only if

$$(1.7) \quad \det [I + \tilde{B}(z)] \neq 0 \quad (\operatorname{Re} z \geq 0),$$

and a more recent result of Grossman and Miller [5] is that  $R(t) \in L^1(\mathbf{R}^+)$  if and only if

$$(1.8) \quad \det [zI - \mathcal{A} - \tilde{B}(z)] \neq 0 \quad (\operatorname{Re} z \geq 0).$$

For results on the integrability of resolvents when the kernel  $B(t) \notin L^1(\mathbf{R}^+)$ , see [14], [7], [9], [3].

\* Received by the editors June 21, 1978, and in revised form March 13, 1979.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916.

‡ Department of Mathematics, University of Missouri, Columbia, Missouri 65211. The work of this author was supported in part by the National Science Foundation under Grant MCS 78-01330.

Analogues of the results of Paley and Wiener and of Grossman and Miller when  $B(t)$  belongs to a weighted  $L^1$  space have been proved by Gelfand, Raikov and Shilov [2, p. 116] (see also [1]) and by Shea and Wainger [14], respectively. Our purpose here is to describe the structure of  $r(t)$  and  $R(t)$  when  $B(t)$  belongs to a weighted  $L^1$  space and the determinants in (1.7) and (1.8) have finitely many zeros in the associated closed half-plane of convergence of  $\tilde{B}(z)$ .

We consider a positive continuous weight function  $\rho(t)$  on  $R^+$  such that  $\rho(0) = 1$ ,

$$(1.9) \quad \rho(s+t) \leq \rho(s)\rho(t) \quad (0 \leq s, t < \infty),$$

and if

$$\rho_0 \equiv \lim_{t \rightarrow \infty} \frac{\log \rho(t)}{-t} \quad \text{with } -\infty < \rho_0 < \infty$$

then

$$(1.10) \quad \rho(t)e^{\rho_0 t} \text{ is nondecreasing on } R^+.$$

(The existence of the limit  $\rho_0$  follows from (1.9); see [2, p. 113]. The regularity condition (1.10) is used to estimate certain integrals and is a crucial hypothesis of the proposition of [8] which we use in § 3; no regularity condition is assumed in [1], [2], [14].) The space  $L^1(R^+, \rho)$  consists of all  $n \times n$  matrix functions  $B(t)$  for which each component  $B_{ij}$  is Borel measurable and satisfies

$$\int_0^\infty \rho(t)|B_{ij}(t)| dt < \infty \quad (1 \leq i, j \leq n).$$

Some of the many interesting and important special choices of  $\rho(t)$  satisfying our conditions (1.9) and (1.10) are

$$\begin{aligned} \rho_1(t) &= e^{-\rho_0 t} && (t \in R^+, -\infty < \rho_0 < \infty), \\ \rho_2(t) &= (1+t)^k \rho_1(t) && (t \in R^+, k \geq 0), \\ \rho_3(t) &= [1 + \log(1+t)]^p \rho_2(t) && (t \in R^+, \rho \geq 0). \end{aligned}$$

For  $B(t) \in L^1(R^+, \rho)$  the determinants in (1.7) and (1.8) exist for  $\text{Re } z \geq \rho_0$  and are analytic in  $\text{Re } z > \rho_0$ . Thus, the meaning of a zero of order  $m$  ( $1 \leq m < \infty$ ) in this open half-plane is clear. If  $z_0$  on  $\text{Re } z = \rho_0$  is a zero of one of the determinants, then we say that  $z_0$  is a zero of order  $m$  if  $t^m B(t) \in L^1(R^+, \rho)$  and the determinant and its first  $m - 1$  derivatives vanish at  $z_0$ , but its  $m$ th derivative is nonzero at  $z_0$ .

**THEOREM 1.1.** *Let  $B(t) \in L^1(R^+, \rho)$  and assume that the only zeros of  $\det [I + \tilde{B}(z)]$  in  $\text{Re } z \geq \rho_0$  occur at  $z = z_j$ ,  $1 \leq j \leq M$ . Let  $m_j$  be the order of the zero  $z_j$  and assume that either (i)  $\text{Re } z_j > \rho_0$ ,  $1 \leq j \leq M$ , or (ii)  $\text{Re } z_j = \rho_0$ ,  $1 \leq j \leq N$ , and  $\text{Re } z_j > \rho_0$ ,  $N < j \leq M$ . In case (i) put  $m = 0$  and in case (ii) put  $m = \max \{m_1, \dots, m_N\}$ .*

*If  $t^{2m} B(t) \in L^1(R^+, \rho)$ , then the solution  $r(t)$  of the integral resolvent equation (1.1) may be expressed as*

$$(1.11) \quad r(t) = \sum_{j=1}^M P_j(t) e^{z_j t} + r_1(t) \quad (t \in R^+),$$

where, for each  $j$ ,  $P_j(t)$  is a matrix of polynomials of degree at most  $m_j - 1$  which depend only on  $B(t)$ , and  $r_1(t) \in L^1(R^+, \rho)$ .

(In Theorem 1.1 and in similar situations later the requirement  $\text{Re } z_j > \rho_0$ ,  $N < j \leq M$ , is to be ignored when  $N = M$ .)



The analogous result for the differential resolvent  $R(t)$  is

**THEOREM 1.2.** *Let  $B(t) \in L^1(\mathbb{R}^+, \rho)$  and assume that the only zeros of  $\det [zI - \mathcal{A} - \tilde{B}(z)]$  in  $\text{Re } z \geq \rho_0$  occur at  $z = z_j, 1 \leq j \leq M$ . Let  $m_j$  be the order of the zero  $z_j$  and assume that either (i)  $\text{Re } z_j > \rho_0, 1 \leq j \leq M$ , or (ii)  $\text{Re } z_j = \rho_0, 1 \leq j \leq N$ , and  $\text{Re } z_j > \rho_0, N < j \leq M$ . In case (i) put  $m = 0$  and in case (ii) put  $m = \max \{m_1, \dots, m_N\}$ .*

*If  $t^{2m}B(t) \in L^1(\mathbb{R}^+, \rho)$ , then the solution  $R(t)$  of (1.2) satisfies (1.11) with  $r(t)$  and  $r_1(t)$  replaced by  $R(t)$  and  $R_1(t)$ , respectively, and with  $R_1(t)$  and  $R'_1(t)$  both in  $L^1(\mathbb{R}^+, \rho)$ .*

The scalar case of Theorem 1.2 (with  $\rho(t) \equiv 1$ ) was proved by Miller [11, Thm. 6]. If (1.2) and (1.4) are not scalar equations (i.e., if  $n > 1$ ), then Theorem 1.2 is new (even for  $\rho(t) \equiv 1$ ) and sharpens Theorem 5 of [11] in which the kernel  $B(t)$  is required to have  $2M_1$  moments where  $M_1 = m_1 + \dots + m_N$  is the total multiplicity of the zeros on  $\text{Re } z = \rho_0$ . Theorem 1.1 is new since a study of the structure of the integral resolvent  $r(t)$  in (1.1) has not previously been made when the determinant in (1.7) vanishes at a finite number of points in the closed half-plane  $\text{Re } z \geq \rho_0$ .

We remark that it is easy to show (see the discussion on pp. 613–614 of [8]) that the moment condition assumed in Theorems 1.1 and 1.2 is best possible even in the scalar case.

Hannsgen [6] has recently obtained the above decomposition of the differential resolvent  $R(t)$  in the scalar case when  $\mathcal{A} = 0$  and when the kernel is piecewise linear and in  $L^1(\mathbb{R}^+)$ . The assumption that the kernel is piecewise linear enables Hannsgen to avoid the moment hypothesis of Theorem 1.2.

As formulae (1.5) and (1.6) show, knowledge that the resolvents  $r(t)$  and  $R(t)$  have the form (1.11) with  $r_1(t)$  and  $R_1(t)$  in  $L^1(\mathbb{R}^+, \rho)$  is clearly useful in analyzing the solutions of the linear Volterra equations (1.3) and (1.4), respectively. Moreover,  $r(t)$  and  $R(t)$  also occur in “variation of constants” formulae (see [10, Chap. 4] and [4]) which solve certain nonlinear perturbed forms of (1.3) and (1.4). An examination of the results in [11, § 6] shows that formula (1.6) and the variation of constants formula may be combined with the fact that the remainder term  $R_1(t)$  is absolutely integrable to investigate the behavior of solutions of certain forced linear and nonlinear perturbed integrodifferential equations. For another application, see the paper [12] by Miller and Nohel. Also, results similar to those in [11, § 6] hold in the case of (1.3).

Theorems 1.1 and 1.2 are consequences of results in § 2 for more general Volterra–Stieltjes systems of convolution type; see § 5 for their proofs.

**2. Linear Volterra–Stieltjes systems.** In this section we consider the linear Volterra–Stieltjes systems.

$$(2.1) \quad u \star A(t) \equiv \int_0^t u(t-s) dA(s) = f(t) \quad (t \in \mathbb{R}^+)$$

and

$$(2.2) \quad u'(t) + u \star A(t) = f(t) \quad (u(0) = u_0, t \in \mathbb{R}^+)$$

where  $f = (f_1, \dots, f_n)$  and  $u$  are complex vector functions with  $n$  components, and  $A = [A_{ij}]$  is an  $n \times n$  matrix of complex-valued functions.

Our setting is similar to that of [8]. Namely, if  $\rho(t)$  is a weight function as defined in the Introduction, the weighted space  $V_+[\rho]$  consists of all  $n \times n$  matrix functions  $A(t)$  for which each component  $A_{ij}$  is of bounded variation on every finite interval  $[0, T]$ , is normalized to be left-continuous and vanish at 0, and satisfies

$$\|A_{ij}\| \equiv \int_0^\infty \rho(t) |dA_{ij}(t)| < \infty \quad (1 \leq i, j \leq n).$$

For  $A(t) \in V_+[\rho]$  the Laplace–Stieltjes transform  $\tilde{A}(z) \equiv \int_0^\infty e^{-zt} dA(t)$  converges absolutely for  $\operatorname{Re} z \geq \rho_0$ . Moreover,  $\tilde{A}(z)$  is bounded and continuous in  $\operatorname{Re} z \geq \rho_0$  and analytic in  $\operatorname{Re} z > \rho_0$ . Also,  $A(t)$  may be decomposed as

$$(2.3) \quad A(t) = h_A(t) + g_A(t) + s_A(t),$$

where  $h_A = [h_{A_{ij}}]$  is a matrix of discrete functions,  $g_A = [g_{A_{ij}}]$  is a matrix of functions absolutely continuous on each finite interval, and  $s_A = [s_{A_{ij}}]$  is a matrix of singular functions. See [8] and [2, p. 166] for a more complete discussion of these ideas.

If  $\rho(t)$  is a weight function,  $m$  is a nonnegative integer,  $A \in V_+[\rho]$  and  $f \in L^1(\mathbb{R}^+, \rho)$ , then we denote by  $H(A, m, \rho)$  and  $H(f, m, \rho)$  the (absolute) moment conditions

$$H(A, m, \rho): \int_0^\infty \rho(t)t^m |dA_{ij}(t)| < \infty \quad (1 \leq i, j \leq n),$$

$$H(f, m, \rho): \int_0^\infty \rho(t)t^m |f_i(t)| dt < \infty \quad (1 \leq i \leq n).$$

We remark that the definition of  $H(f, m, \rho)$  used here differs from the one used in [8].

Associate with  $A(t) \in V_+[\rho]$  the scalar function

$$D(t) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) A_{1\sigma(1)} \star \cdots \star A_{n\sigma(n)},$$

where  $S_n$  is the symmetric group on  $\{1, \dots, n\}$  and  $\operatorname{sgn}(\sigma) = \pm 1$  according as the permutation  $\sigma$  is even or odd. Equivalently,  $D(t)$  is the scalar function which satisfies

$$\tilde{D}(z) = \det \tilde{A}(z) \quad (\operatorname{Re} z \geq \rho_0).$$

Since  $\det \tilde{A}(z)$  is analytic in  $\operatorname{Re} z > \rho_0$ , the meaning of a zero of order  $m$  in  $\operatorname{Re} z > \rho_0$  is clear. If  $\operatorname{Re} z_0 = \rho_0$ , then we say that  $z_0$  is a zero of  $\det \tilde{A}(z)$  of order  $m$  if  $H(A, m, \rho)$  holds and

$$\int_0^\infty e^{-z_0 t} t^j dD(t) = 0 \quad (0 \leq j \leq m-1),$$

but

$$\int_0^\infty e^{-z_0 t} t^m dD(t) \neq 0.$$

Note that if we decompose  $D(t)$  as  $D(t) = h_D(t) + g_D(t) + s_D(t)$  (as in (2.3)), then, since the discrete part of the convolution of two functions in  $V_+[\rho]$  is the convolution of their discrete parts [2, p. 179], we have

$$\tilde{h}_D(z) = \det \tilde{h}_A(z) \quad (\operatorname{Re} z \geq \rho_0).$$

Finally, in Theorem 2.1 we assume that the solution  $u(t)$  of (2.1) exists and is Borel measurable on  $\mathbb{R}^+$  and that  $\int_0^T \rho(t)|u(t)| dt < \infty$  for all  $T > 0$ .

**THEOREM 2.1.** *Let  $A \in V_+[\rho]$  and assume that in  $\operatorname{Re} z \geq \rho_0$   $\det \tilde{A}(z)$  has zeros only at  $z = z_j$ ,  $1 \leq j \leq M$ . Suppose that*

$$(2.4) \quad \begin{aligned} &\text{except near the points } z_j, 1 \leq j \leq M, \\ &1/\tilde{D}(z) \text{ is bounded in } \operatorname{Re} z \geq \rho_0 \end{aligned}$$

and

$$(2.5) \quad \inf_{-\infty < \sigma < \infty} |\det \tilde{h}_A(\rho_0 + i\sigma)| > \|s_D\|.$$

Let  $m_j$  be the order of the zero  $z_j$  and assume that either (i)  $\operatorname{Re} z_j > \rho_0$ ,  $1 \leq j \leq M$  or (ii)  $\operatorname{Re} z_j = \rho_0$ ,  $1 \leq j \leq N$ , and  $\operatorname{Re} z_j > \rho_0$ ,  $N < j \leq M$ . In case (i) put  $m = 0$  and in case (ii) put  $m = \max \{m_1, \dots, m_N\}$ .

If  $f \in L^1(\mathbb{R}^+, \rho)$  and  $H(A, 2m, \rho)$ ,  $H(f, m, \rho)$  hold, then the solution  $u(t)$  of (2.1) satisfies

$$(2.6) \quad u(t) = \sum_{j=1}^M p_j(t) e^{z_j t} + u_1(t) \quad (t \in \mathbb{R}^+),$$

where, for each  $j$ ,  $p_j(t) = (p_{j1}(t), \dots, p_{jn}(t))$  with each  $p_{jk}(t)$  a polynomial of degree at most  $m_j - 1$  which depends on  $A$  and  $f$ , and  $u_1 \in L^1(\mathbb{R}^+, \rho)$ .

The proof of Theorem 2.1 is given in § 3.

We next consider (2.2). By a solution of (2.2) we mean a vector  $u(t)$  absolutely continuous on bounded intervals  $[0, T]$ , and such that  $u(0) = u_0$  and (2.2) holds a.e. on  $\mathbb{R}^+$ . To describe the solution of (2.2) we consider the zeros of

$$(2.7) \quad \det [zI + \tilde{A}(z)] \quad (\operatorname{Re} z \geq \rho_0),$$

where  $I$  is the  $n \times n$  identity matrix. Since this determinant is analytic in  $\operatorname{Re} z > \rho_0$ , the meaning of a zero of order  $m$  in  $\operatorname{Re} z > \rho_0$  is clear. If  $z_0$  on  $\operatorname{Re} z = \rho_0$  is a zero of the determinant, then  $z_0$  is a zero of order  $m$  if  $H(A, m, \rho)$  holds and

$$\frac{d^j}{dz^j} (\det [zI + \tilde{A}(z)]) = 0 \quad (z = z_0, 0 \leq j \leq m - 1),$$

but

$$\frac{d^m}{dz^m} (\det [zI + \tilde{A}(z)]) \neq 0 \quad (z = z_0).$$

We then have

**THEOREM 2.2.** Let  $A \in V_+[\rho]$  and assume that in  $\operatorname{Re} z \geq \rho_0$  the determinant in (2.7) has zeros only at  $z = z_j$ ,  $1 \leq j \leq M$ . Let  $m_j$  be the order of the zero  $z_j$  and assume that either (i)  $\operatorname{Re} z_j > \rho_0$ ,  $1 \leq j \leq M$ , or (ii)  $\operatorname{Re} z_j = \rho_0$ ,  $1 \leq j \leq N$ , and  $\operatorname{Re} z_j > \rho_0$ ,  $N < j \leq M$ . In case (i) put  $m = 0$  and in case (ii) put  $m = \max \{m_1, \dots, m_N\}$ .

If  $f \in L^1(\mathbb{R}^+, \rho)$  and  $H(A, 2m, \rho)$ ,  $H(f, m, \rho)$  hold, then the solution of (2.2) has the form (2.6) with  $u_1(t)$  and  $u'_1(t)$  both in  $L^1(\mathbb{R}^+, \rho)$ .

We note that in the case of Theorem 2.2, unlike Theorem 2.1, we do not require hypotheses such as (2.4) and (2.5). The reason for this will be apparent in the proof of Theorem 2.2 which is given in § 4.

Theorems 2.1 and 2.2 are somewhat similar to the results in [8] where it is proved (under different moment hypotheses) that if  $f \in L^\infty(0, T)$  for all  $T > 0$  and  $f(t) = o(1/\rho(t))$  as  $t \rightarrow \infty$ , then the solutions of (2.1) and (2.2) have the form (2.6) with  $u_1(t) = o(1/\rho(t))$  as  $t \rightarrow \infty$ . Theorems 2.1 and 2.2 are more useful than the results in [8] in view of the fact that they may be used to obtain results of the type in [11, § 6] (see the discussion at the end of § 1) which are more useful than the results one obtains from  $R_1(t) = o(1/\rho(t))$  as  $t \rightarrow \infty$ . For example, properties of  $\rho(t)f(t)$  such as boundedness, convergence at  $+\infty$ , and integrability on  $\mathbb{R}^+$  propagate the same behavior to the convolution  $\int_0^t f(t-s)R_1(s) ds$  when  $R_1(t)$  is integrable with respect to  $\rho(t)$ , but not necessarily when  $R_1(t) = o(1/\rho(t))$  as  $t \rightarrow \infty$ .

It would be of interest to analyze the behavior of solutions of (2.1) (respectively, (2.2)) when  $\det \tilde{A}(z)$  (respectively,  $\det [zI + \tilde{A}(z)]$ ) has an infinite number of zeros in  $\operatorname{Re} z \geq \rho_0$ . However, our technique of stripping zeros which comprises the heart of the

proof of Theorem 2.1 leads to convergence problems when there are infinitely many zeros. In particular, it is not clear what happens to the first term on the right in (2.6) as  $M \rightarrow \infty$ .

**3. Proof of Theorem 2.1.** We break the proof into three parts: (A) the scalar case when condition (i) holds, (B) the scalar case when condition (ii) holds, and (C) the vector case. We remark that in the two scalar cases  $D(t) = A(t)$ .

*Proof of part A.* The proof of this case is similar to the proof of Theorem 1 of [8]. Let

$$V_j(t) = \sum_{l=1}^{m_j} c_{l,j} \int_0^t \tau^{l-1} e^{z_j \tau} d\tau / (l-1)! \quad (-\infty < t < \infty),$$

where  $\tilde{V}_j(z) = \sum_{l=1}^{m_j} c_{l,j} (z - z_j)^{-l}$  is the principal part of  $1/\tilde{A}(z)$  at  $z = z_j$ . The proposition in [8] yields the existence of  $C_0(t) \in V_+[\rho]$  such that

$$(3.1) \quad u(t) = f \star C_0(t) + \sum_{j=1}^M f \star V_j(t) \quad (t \in \mathbb{R}^+).$$

A simple argument using (1.9), an interchange of the order of integration,  $f \in L^1(\mathbb{R}^+, \rho)$  and  $C_0 \in V_+[\rho]$  shows that  $f \star C_0 \in L^1(\mathbb{R}^+, \rho)$ . Also, the Laplace transform  $\tilde{f}(z)$  exists for  $\text{Re } z \geq \rho_0$ . Thus, as in [8],

$$\begin{aligned} f \star V_j(t) &= \sum_{l=1}^{m_j} c_{l,j} \left\{ \int_0^t f(s) (t-s)^{l-1} e^{-z_j s} ds \right\} e^{z_j t} / (l-1)! \\ &= \sum_{l=1}^{m_j} c_{l,j} \left\{ \sum_{p=0}^{l-1} \binom{l-1}{p} t^p \int_0^t f(s) (-s)^{l-1-p} e^{-z_j s} ds \right\} e^{z_j t} / (l-1)! \\ &= \sum_{l=1}^{m_j} c_{l,j} \left\{ \sum_{p=0}^{l-1} \binom{l-1}{p} \tilde{f}^{(l-1-p)}(z_j) t^p e^{z_j t} - \int_t^\infty f(s) (t-s)^{l-1} e^{z_j(t-s)} ds \right\} / (l-1)! \\ &\equiv p_j(t) e^{z_j t} - \sum_{l=1}^{m_j} c_{l,j} \int_t^\infty f(s) (t-s)^{l-1} e^{z_j(t-s)} ds / (l-1)!, \end{aligned}$$

where clearly  $p_j(t)$  is a polynomial of degree at most  $m_j - 1$  which depends only on  $A$  and  $f$ . Furthermore, if we write  $\text{Re } z_j = \rho_0 + \delta_j$ ,  $\delta_j > 0$ , then using (1.10) and  $f \in L^1(\mathbb{R}^+, \rho)$ , we find that

$$\begin{aligned} \int_0^\infty \rho(t) \left| \int_t^\infty f(s) (t-s)^{l-1} e^{z_j(t-s)} ds \right| dt &\leq \int_0^\infty \rho(t) \int_t^\infty |f(s)| |t-s|^{l-1} e^{(\rho_0 + \delta_j)(t-s)} ds dt \\ &\leq \int_0^\infty \int_t^\infty \rho(s) |f(s)| |t-s|^{l-1} e^{\delta_j(t-s)} ds dt \\ &\leq \int_0^\infty \rho(s) |f(s)| \int_0^s |t-s|^{l-1} e^{\delta_j(t-s)} dt ds \\ &\leq K \int_0^\infty \rho(s) |f(s)| ds < \infty. \end{aligned}$$

Combining these results with (3.1) yields (2.6) and completes the proof of Part A.

*Proof of part B.* The proof of this case is similar to the proof of Theorem 2 of [8]; however, to avoid numerous references to the argument in [8], we give the proof in some detail. The idea of the proof is to reduce this case to the situation covered by Part

A. First, let  $N_k$ ,  $1 \leq k \leq m$ , be the number of zeros of  $\tilde{A}(z)$  of order  $k$  on  $\text{Re } z = \rho_0$ , and put  $M_0 = 0$ ,  $M_k = \sum_{i=1}^k N_i$ ,  $1 \leq k \leq m$ . Furthermore, let the zeros be labeled so that  $z_j$  is a zero of order  $k$  for  $M_{k-1} < j \leq M_k$ .

The proof is by induction on  $m$ . Thus, suppose  $m = 1$ , so that all the zeros of  $\tilde{A}(z)$  on  $\text{Re } z = \rho_0$  are simple. Fix  $z_0$ ,  $\text{Re } z_0 < \rho_0$ , and define

$$(3.2) \quad S_j(t) = (z_j - z_0) \int_0^t e^{z_j s} ds \quad (1 \leq j \leq N, t \in \mathbb{R}^+)$$

and

$$B_1(t) = (J + S_1) \star \cdots \star (J + S_N)(t) \quad (t \in \mathbb{R}^+),$$

where  $J$  is the unit step function

$$(3.3) \quad J(0) = 0, \quad J(t) = 1 \quad (t > 0).$$

Finally, put

$$(3.4) \quad C_1(t) = A \star B_1(t) \quad (t \in \mathbb{R}^+).$$

Observe that  $B_1(t)$  may be written as

$$(3.5) \quad B_1(t) = J(t) + \sum_{j=1}^N \alpha_j S_j(t) \quad (t \in \mathbb{R}^+),$$

where

$$\alpha_j = \prod_{\substack{k=1 \\ k \neq j}}^N \frac{z_j - z_0}{z_j - z_k} \quad (j = 1, 2, \dots, N).$$

This representation may be obtained by a simple Laplace transform argument; see [8, p. 604] for details.

Combining (3.4), (3.5) and (3.2) with  $\tilde{A}(z_j) = 0$ ,  $1 \leq j \leq N$ , yields

$$C_1(t) = A(t) - \sum_{j=1}^N \alpha_j (z_j - z_0) E_j(t),$$

where

$$E_j(t) = \begin{cases} \left( e^{z_j t} \int_t^\infty e^{-z_j s} dA(s) + A(t) \right) / z_j, & (z_j \neq 0), \\ t \int_t^\infty dA(s) + \int_0^t s dA(s), & (z_j = 0). \end{cases}$$

Consequently,

$$dC_1(t) = dA(t) - \sum_{j=1}^N \alpha_j (z_j - z_0) \int_t^\infty e^{-z_j s} dA(s) e^{z_j t} dt.$$

(We remark that the expression (3.7) for  $C_1(t)$  and the expression for  $dC_1(t)$  in [8, p. 604] are inaccurate; however, the inaccuracies do not affect the results there. The

expressions given above are correct.) By (1.10) and  $H(A, 2, \rho)$

$$\begin{aligned} \int_0^\infty \rho(t) \left| e^{z_1 t} \int_t^\infty e^{-z_1 s} dA(s) \right| dt &\leq \int_0^\infty \rho(t) e^{\rho_0 t} \int_t^\infty e^{-\rho_0 s} |dA(s)| dt \\ &= \int_0^\infty e^{-\rho_0 s} \int_0^s \rho(t) e^{\rho_0 t} dt |dA(s)| \\ &\leq \int_0^\infty \rho(s) (s^2/2) |dA(s)| < \infty \end{aligned}$$

for  $1 \leq j \leq N$ ; hence,  $H(C_1, 1, \rho)$  clearly holds.

Now define

$$(3.6) \quad u^1(t) = u(t) - \sum_{j=1}^N \beta_j e^{z_j t} \quad (t \in \mathbb{R}^+),$$

where

$$(3.7) \quad \beta_j = \tilde{f}(z_j) / \tilde{A}'(z_j) \quad (1 \leq j \leq N).$$

Then it follows from (3.6), (3.4), (2.1), (3.5), (3.2),  $H(f, 1, \rho)$  and  $C_1 \in V_+[\rho]$  that

$$\begin{aligned} u^1 \star C_1(t) &= u \star C_1(t) - \sum_{j=1}^N \beta_j e^{z_j t} \int_0^t e^{-z_j s} dC_1(s) \\ &= f \star B_1(t) - \sum_{j=1}^N \beta_j \left\{ \tilde{C}_1(z_j) - \int_t^\infty e^{-z_j s} dC_1(s) \right\} e^{z_j t} \\ &= f(t) + \sum_{j=1}^N \alpha_j (z_j - z_0) \left\{ \tilde{f}(z_j) - \int_t^\infty e^{-z_j s} f(s) ds \right\} e^{z_j t} \\ &\quad - \sum_{j=1}^N \beta_j \left\{ \tilde{C}_1(z_j) - \int_t^\infty e^{-z_j s} dC_1(s) \right\} e^{z_j t} \\ &\equiv f_1(t) \quad (t \in \mathbb{R}^+). \end{aligned}$$

To see that the result established in part A applies to the equation

$$(3.8) \quad u^1 \star C_1(t) = f_1(t) \quad (t \in \mathbb{R}^+),$$

note first that by (3.4) and the definition of  $B_1(t)$ ,

$$\tilde{C}_1(z) = \tilde{A}(z) \tilde{B}_1(z) = \tilde{A}(z) \prod_{k=1}^N \frac{z - z_0}{z - z_k}$$

for  $\text{Re } z \geq \rho_0, z \neq z_j, 1 \leq j \leq N$ . It follows from  $m_j = 1$ , the definition of  $\alpha_j$ , and the continuity of  $\tilde{A}'(z)$  and  $\tilde{C}_1(z)$  that

$$(3.9) \quad \tilde{C}_1(z_j) = \alpha_j \tilde{A}'(z_j)(z_j - z_0) \neq 0$$

for  $1 \leq j \leq N$ . Thus,  $\tilde{C}_1(z)$  has the same zeros (including order) as  $\tilde{A}(z)$  in  $\text{Re } z > \rho_0$  and has no zeros on  $\text{Re } z = \rho_0$ . Furthermore, except near the points  $z_j, N < j \leq M, 1/\tilde{C}_1(z)$  is bounded in  $\text{Re } z \geq \rho_0$ .

Now, let the discrete and singular parts of  $C_1(t)$  be  $h_1(t)$  and  $s_1(t)$ , respectively. Then  $h_1 = h_A$  and  $s_1 = s_A$ , for by (3.5)  $C_1 = A \star B_1$  can be written as  $A \star J = A$  plus a linear combination of convolutions of  $A$  with the functions  $S_j, 1 \leq j \leq N$ , and it is easy to

see that these convolutions are absolutely continuous. Hence, (2.5) holds with the subscripts  $A$  and  $D$  replaced by 1 (recall  $D = A$  since  $n = 1$ ).

It remains to show that  $f_1(t) \in L^1(\mathbb{R}^+, \rho)$ . By (3.7) and (3.9),  $\beta_j \tilde{C}_1(z_j) = \alpha_j(z_j - z_0) \tilde{f}(z_j)$ , so that the expression for  $f_1(t)$  simplifies to

$$(3.10) \quad f_1(t) = f(t) + \sum_{j=1}^N \left\{ \beta_j \int_t^\infty e^{-z_j s} dC_1(s) - \alpha_j(z_j - z_0) \int_t^\infty e^{-z_j s} f(s) ds \right\} e^{z_j t}.$$

Since  $H(C_1, 1, \rho)$  and  $H(f, 1, \rho)$  hold, computations similar to those above establishing  $H(C_1, 1, \rho)$  now show that the integral terms in (3.10) belong to  $L^1(\mathbb{R}^+, \rho)$ . By assumption,  $f \in L^1(\mathbb{R}^+, \rho)$  and we may apply Part A of the proof of Theorem 2.1 to (3.8) to obtain

$$u_1(t) = \sum_{j=N+1}^M p_j(t) e^{z_j t} + u_1(t)$$

or, by (3.6),

$$u(t) = \sum_{j=1}^M p_j(t) e^{z_j t} + u_1(t),$$

where  $p_j(t)$ ,  $1 \leq j \leq M$ , is a polynomial of degree at most  $m_j - 1$  which depends only on  $A$  and  $f$ , and  $u_1 \in L^1(\mathbb{R}^+, \rho)$ . The proof of part B when  $m = 1$  is complete.

Now assume the theorem is true for  $1 \leq m \leq n$  and consider the case  $m = n + 1$ . For  $M_n < j \leq M_{n+1} = N$ , let  $S_j(t)$  be defined as in (3.2) and put

$$(3.11) \quad \begin{aligned} B_{n+1}(t) &= (J + S_{M_{n+1}}) \star \cdots \star (J + S_N)(t) \\ &= J(t) + \sum_{j=M_n+1}^N \alpha_j S_j(t) \quad (t \in \mathbb{R}^+); \end{aligned}$$

here, as before,  $J(t)$  is the unit step function defined in (3.3) and

$$\alpha_j = \prod_{\substack{k=M_n+1 \\ k \neq j}}^N \frac{z_j - z_0}{z_j - z_k} \quad (j = M_n + 1, \dots, N).$$

Finally, put

$$(3.12) \quad C_{n+1}(t) = A \star B_{n+1}(t) \quad (t \in \mathbb{R}^+).$$

An argument similar to the one yielding  $H(C_1, 1, \rho)$  shows that  $H(A, 2m, \rho)$  implies  $H(C_{n+1}, 2m - 1, \rho)$  where  $m = n + 1$ . Moreover, as was the case with  $\tilde{C}_1(z)$ , in  $\text{Re } z > \rho_0$   $\tilde{C}_{n+1}(z)$  has the same zeros (including order) as  $\tilde{A}(z)$ . On  $\text{Re } z = \rho_0$  the only zeros of  $\tilde{C}_{n+1}(z)$  are the  $z_j$ ,  $1 \leq j \leq N$ ; if  $1 \leq j \leq M_n$ , then  $z_j$  is a zero of order  $m_j$  for  $\tilde{C}_{n+1}(z)$ , but if  $M_n < j \leq N$ , then  $z_j$  is a zero of order  $m_j - 1$  for  $\tilde{C}_{n+1}(z)$ . Furthermore, for  $M_n < j \leq N$ , one may use (3.12), the expression for  $\tilde{B}_{n+1}(z)$ , the definition of  $\alpha_j$ ,  $n = m_j - 1$ , and Taylor's formula with remainder to obtain

$$(3.13) \quad \tilde{C}_{n+1}^{(n)}(z_j) = \alpha_j(z_j - z_0) \tilde{A}^{(n+1)}(z_j)/(n + 1).$$

Also, Taylor's formula with remainder yields

$$(3.14) \quad \tilde{A}(z) = \tilde{A}^{(n+1)}(w)(z - z_j)^{n+1}/(n + 1)! \quad (\text{Re } z \geq \rho_0)$$

for some  $w$  on the line from  $z_j$  to  $z$ . If at  $z_j$   $1/\tilde{A}(z)$  has "principal part"

$$(3.15) \quad \frac{c_{n+1,j}}{(z - z_j)^{n+1}} + \frac{c_{n,j}}{(z - z_j)^n} + \cdots + \frac{c_{1,j}}{(z - z_j)};$$

then, by writing  $1 = \tilde{A}(z)(1/\tilde{A}(z))$  and using (3.14) and (3.15), we find upon letting  $z$  tend to  $z_j$  that

$$(3.16) \quad c_{n+1,j} = (n+1)!/\tilde{A}^{(n+1)}(z_j).$$

Now define

$$u^{n+1}(t) = u(t) - g(t) \quad (t \in R^+),$$

where

$$g(t) = \sum_{j=M_{n+1}}^N (c_{n+1,j}/n!) \tilde{f}(z_j) t^n e^{z_j t} \quad (t \in R^+).$$

Then, by (3.12),

$$\begin{aligned} u^{n+1} \star C_{n+1}(t) &= u \star C_{n+1}(t) - g \star C_{n+1}(t) \\ &= f \star B_{n+1}(t) - g \star C_{n+1}(t) \\ &\equiv f_{n+1}(t) \quad (t \in R^+). \end{aligned}$$

In order to apply the inductive hypothesis to the equation

$$(3.17) \quad u^{n+1} \star C_{n+1}(t) = f_{n+1}(t) \quad (t \in R^+)$$

we must show that  $f_{n+1}(t) \in L^1(R^+, \rho)$  and that  $H(f_{n+1}, n, \rho)$  holds. First, using (3.11) in the definition of  $f_{n+1}(t)$  yields

$$(3.18) \quad f_{n+1}(t) = f(t) + \sum_{j=M_{n+1}}^N \alpha_j (z_j - z_0) \left\{ \int_0^t f(s) e^{-z_j s} ds \right\} e^{z_j t} - g \star C_{n+1}(t),$$

where

$$(3.19) \quad g \star C_{n+1}(t) = \sum_{j=M_{n+1}}^N (c_{n+1,j}/n!) \tilde{f}(z_j) \int_0^t (t-s)^n e^{-z_j s} dC_{n+1}(s) e^{z_j t}.$$

For  $M_n < j \leq N$ , put  $\beta_j = c_{n+1,j}/n!$  and observe from (3.13) and (3.16) that

$$\beta_j \tilde{C}_{n+1}^{(n)}(z_j) = \alpha_j (z_j - z_0).$$

Thus, by applying the binomial theorem to  $(t-s)^n$  in (3.19) and rewriting the integral, we have

$$\begin{aligned} g \star C_{n+1}(t) &= \sum_{j=M_{n+1}}^N \beta_j \tilde{f}(z_j) \sum_{p=0}^n \binom{n}{p} t^p \left\{ \tilde{C}_{n+1}^{(n-p)}(z_j) - \int_t^\infty (-s)^{n-p} e^{-z_j s} dC_{n+1}(s) \right\} e^{z_j t} \\ &= \sum_{j=M_{n+1}}^N \beta_j \tilde{f}(z_j) \left\{ \tilde{C}_{n+1}^{(n)}(z_j) - \int_t^\infty (t-s)^n e^{-z_j s} dC_{n+1}(s) \right\} e^{z_j t} \\ &= \sum_{j=M_{n+1}}^N \alpha_j (z_j - z_0) \tilde{f}(z_j) e^{z_j t} - \sum_{j=M_{n+1}}^N \beta_j \tilde{f}(z_j) \int_t^\infty (t-s)^n e^{-z_j s} dC_{n+1}(s) e^{z_j t}. \end{aligned}$$

Substituting this expression in (3.18) and writing

$$\int_0^t f(s) e^{-z_j s} ds = \tilde{f}(z_j) - \int_t^\infty f(s) e^{-z_j s} ds$$



yield

$$(3.20) \quad \begin{aligned} f_{n+1}(t) = f(t) - \sum_{j=M_{n+1}}^N \alpha_j(z_j - z_0) & \left\{ \int_t^\infty f(s) e^{-z_j s} ds \right\} e^{z_j t} \\ & + \sum_{j=M_{n+1}}^N \beta_j \tilde{f}(z_j) \int_t^\infty (t-s)^n e^{-z_j s} dC_{n+1}(s) e^{z_j t}. \end{aligned}$$

That  $f_{n+1}(t) \in L^1(\mathbb{R}^+, \rho)$  follows easily from (3.20) since  $f(t) \in L^1(\mathbb{R}^+, \rho)$ ,  $\rho(t)$  satisfies (1.10), and  $H(f, n+1, \rho)$  and  $H(C_{n+1}, 2n+1, \rho)$  hold.

Observe next that  $H(C_{n+1}, 2n+1, \rho)$  and (1.10) imply

$$(3.21) \quad \begin{aligned} \int_0^\infty \rho(t) t^n \left| \int_t^\infty (t-s)^n e^{z_i(t-s)} dC_{n+1}(s) \right| dt & \leq \int_0^\infty \rho(t) t^n \int_t^\infty s^n e^{\rho_0(t-s)} |dC_{n+1}(s)| dt \\ & \leq \int_0^\infty \rho(s) s^n \int_0^s t^n dt |dC_{n+1}(s)| \\ & \leq \int_0^\infty \rho(s) s^{2n+1} |dC_{n+1}(s)| < \infty. \end{aligned}$$

Similarly, it follows from  $H(f, n+1, \rho)$  and (1.10) that

$$\int_0^\infty \rho(t) t^n \left| \int_t^\infty f(s) e^{z_i(t-s)} ds \right| dt \leq \int_0^\infty \rho(s) s^{n+1} |f(s)| ds < \infty.$$

Using these results in (3.20) and recalling the hypothesis  $H(f, n+1, \rho)$ , we clearly have that  $H(f_{n+1}, n, \rho)$  holds.

The inductive hypothesis implies that  $u^{n+1}(t)$  has the form

$$u^{n+1}(t) = \sum_{j=1}^M p_j^*(t) e^{z_j t} + u_1(t) \quad (t \in \mathbb{R}^+),$$

where  $u_1(t) \in L^1(\mathbb{R}^+, \rho)$  and the  $p_j^*(t)$  are polynomials of degree at most  $m_j - 1$  unless  $M_n < j \leq N$ , in which case the degree is at most  $m_j - 2$ . Thus,

$$u(t) = \sum_{j=1}^M p_j(t) e^{z_j t} + u_1(t) \quad (t \in \mathbb{R}^+),$$

where  $p_j(t) = p_j^*(t)$ ,  $1 \leq j \leq M_n$  and  $N < j \leq M$ , and  $p_j(t) = p_j^*(t) + (c_{n+1,j}/n!) \tilde{f}(z_j) t^n e^{z_j t}$ ,  $M_n < j \leq N$ . The proof of part B is now complete.

*Proof of part C.* The proof of this case is similar to the proof of Theorem 3 of [8]. Let  $\text{adj } A$  denote the  $n \times n$  matrix which is formally the adjoint of  $A$  but with convolution replacing multiplication. Then the equation

$$(3.22) \quad u \star A \star \text{adj } A(t) = f \star \text{adj } A(t) \quad (t \in \mathbb{R}^+)$$

is equivalent to the  $n$  scalar equations

$$(3.23) \quad u_k \star D(t) = \phi_k(t) \quad (1 \leq k \leq n, t \in \mathbb{R}^+),$$

where  $\phi = (\phi_1, \dots, \phi_n) = f \star \text{adj } A$ . (Here  $u_k$ ,  $k = 1$ , should not be confused with the remainder term  $u_1$  in (2.6).)

Conditions  $H(\phi, m, \rho)$  and  $H(D, 2m, \rho)$  follow easily from hypotheses  $H(f, m, \rho)$  and  $H(A, 2m, \rho)$ . Also, since  $A \in V_+[\rho]$  and  $f(t) \in L^1(\mathbb{R}^+, \rho)$ , we have  $\phi(t) \in L^1(\mathbb{R}^+, \rho)$ . Finally, we recall (see § 2) that  $\hat{h}_D(z) = \det \hat{h}_A(z)$ . Thus, part B of the proof of Theorem

2.1 may be applied to each of the scalar equations in (3.23) to obtain

$$u_k(t) = \sum_{j=1}^M p_{jk}(t) e^{z_j t} + u_{k1}(t) \quad (1 \leq k \leq n),$$

where  $u_{k1}(t) \in L^1(\mathbb{R}^+, \rho)$  and  $p_{jk}(t)$  is a polynomial of degree at most  $m_j - 1$  which depends only on  $A$  and  $f$ . Setting  $p_j(t) = (p_{j1}(t), \dots, p_{jn}(t))$  ( $1 \leq j \leq M, t \in \mathbb{R}^+$ ) and  $u_1(t) = (u_{11}(t), \dots, u_{1n}(t))$  ( $t \in \mathbb{R}^+$ ) completes part C of the proof of Theorem 2.1.

**4. Proof of Theorem 2.2.** The proof that  $u_1 \in L^1(\mathbb{R}^+, \rho)$  is similar to the proof of Theorem 4 of [8]. Namely, define  $G(t) = e^{-(1-\rho_0)t} I$  and convolve both sides of (2.2) with  $G(t)$  to obtain

$$(4.1) \quad u' * G(t) + u * G * A(t) = f * G(t) \quad (t \in \mathbb{R}^+).$$

Here,  $f * G(t) = \int_0^t f(t-s)G(s) ds$  for  $t \in \mathbb{R}^+$ . Integrating the first term in (4.1) by parts enables us to rewrite (4.1) as

$$(4.2) \quad u(t) + u * b(t) = k(t) \quad (t \in \mathbb{R}^+),$$

where  $b(t) \equiv G'(t) + G * A(t)$  and  $k(t) \equiv f * G(t) + u(0)G(t)$ . Equation (4.2) may be rewritten as

$$(4.3) \quad u * B(t) = k(t) \quad (t \in \mathbb{R}^+),$$

where  $B \in V_+[\rho]$  is defined by  $B(t) = J(t)I + \int_0^t b(s) ds$  with  $J(t)$  as defined in (3.3). Since  $H(G, j, \rho)$  and  $H(G', j, \rho)$  are satisfied for  $j = 1, 2, \dots$ , it is easy to check that  $H(B, 2m, \rho)$  and  $H(k, m, \rho)$  follow from  $H(A, 2m, \rho)$  and  $H(f, m, \rho)$ , respectively. Also, it follows from the definitions of  $B$  and  $b$  that  $\tilde{B}(z) = \tilde{G}(z)[zI + \tilde{A}(z)]$  for  $\text{Re } z \geq \rho_0$ . Since  $\tilde{G}(z) = (z + 1 - \rho_0)^{-1}I$ , we see that  $\det \tilde{B}(z)$  and  $\det [zI + \tilde{A}(z)]$  have the same zeros with identical multiplicities in  $\text{Re } z \geq \rho_0$ . In addition, the discrete and singular parts of  $B$  are  $h_B(t) = J(t)I$  and  $s_B(t) = 0$ , respectively. Hence,  $B(t)$  satisfies hypotheses (2.4) and (2.5) of Theorem 2.1. Thus, we may apply Theorem 2.1 to see that  $u(t)$  has the form (2.6) with  $u_1 \in L^1(\mathbb{R}^+, \rho)$ .

It remains to show that  $u'_1(t) \in L^1(\mathbb{R}^+, \rho)$ . To do this, set  $w(t) = \sum_{j=1}^M w_j(t)$  where the  $w_j(t) \equiv p_j(t) e^{z_j t}$  are the terms appearing in the sum in (2.6). We begin by showing that for  $1 \leq j \leq M$ ,

$$(4.4) \quad w'_j(t) + \int_0^\infty w_j(t-s) dA(s) = 0.$$

In order to verify (4.4), we first observe that an examination of the proof of Theorem 2.1 yields that the polynomials  $p_j(t)$  are given by

$$(4.5) \quad p_j(t) = \sum_{k=0}^{m_j-1} \sum_{r=k+1}^{m_j} \frac{c_{r,j}}{(r-k-1)! k!} [\tilde{k}(z) \text{adj } \tilde{B}(z)]_{z=z_j}^{(r-k-1)} t^k,$$

where

$$\sum_{r=1}^{m_j} c_{r,j} (z - z_j)^{-r}$$

is the principal part of  $[\det \tilde{B}(z)]^{-1}$  at  $z = z_j$ . An easy calculation using the binomial formula on  $(t-s)^k$ , the formula for the derivatives of  $\tilde{A}(z)$ , (4.5) and interchanges of the orders of summation yields that the left side of (4.4) is given by

$$\sum_{l=0}^{m_j-1} \sum_{r=l+1}^{m_j} (c_{r,j}/l!) t^l e^{z_j t} S(r, l),$$

where

$$S(r, l) = \sum_{k=l}^{r-1} [(r-1-k)!(k-l)!]^{-1} [\tilde{k}(z) \operatorname{adj} \tilde{B}(z)]^{(r-1-k)} [zI + \tilde{A}(z)]^{(k-1)}|_{z=z_j}$$

For each  $r$  and  $l$  satisfying  $l+1 \leq r \leq m_j$ ,  $0 \leq l \leq m_j$ , Leibnitz's formula may be used to rewrite  $S(r, l)$  as

$$\begin{aligned} & [(n-1-l)!]^{-1} [\tilde{k}(z) (\operatorname{adj} \tilde{B}(z)) (zI + \tilde{A}(z))]_{z=z_j}^{(r-1-l)} \\ &= [(n-1-l)!]^{-1} [\tilde{k}(z) \operatorname{adj} \tilde{G}(z) \det(zI + \tilde{A}(z))]_{z=z_j}^{(r-1-l)}, \end{aligned}$$

where the equality follows from the expression for  $\tilde{B}(z)$  and a basic property of adjoints. Since  $0 \leq r-1-l \leq m_j-1$  and  $z_j$  is a zero of  $\det[zI + \tilde{A}(z)]$  of order  $m_j$ , the last expression is zero for each  $r$  and  $l$ , and (4.4) holds.

To complete the proof that  $u' \in L^1(\mathbb{R}^+, \rho)$ , note that (4.4) and linearity yield

$$(4.6) \quad w'(t) + w \star A(t) = - \int_t^\infty w(t-s) dA(s).$$

Thus, if we substitute the expression (2.6) into (2.2), rearrange and use (4.6), we find that

$$u'_1(t) = f(t) - u_1 \star A(t) + \int_t^\infty w(t-s) dA(s)$$

a.e. on  $\mathbb{R}^+$ . A calculation similar to that in (3.21) shows that the third term on the right side of the last equation belongs to  $L^1(\mathbb{R}^+, \rho)$ . Since  $f(t)$  and  $u_1 \star A(t)$  also are in  $L^1(\mathbb{R}^+, \rho)$ , it follows that  $u'_1 \in L^1(\mathbb{R}^+, \rho)$ , and the proof of Theorem 2.2 is complete.

**5. Proofs of Theorems 1.1 and 1.2.** In this section Theorems 1.1 and 1.2 are deduced from the results of § 2.

*A. Proof of Theorem 1.1.* Let  $J(t)$  be the unit step function defined in (3.3) and put  $A(t) = J(t)I + \int_0^t B(s) ds$  ( $t \in \mathbb{R}^+$ ). Then (1.1) may be written as

$$r \star A(t) = B(t) \quad (t \in \mathbb{R}^+).$$

This matrix equation is equivalent to the  $n$  equations

$$r_i \star A(t) = B_i(t) \quad (i = 1, \dots, n, t \in \mathbb{R}^+),$$

where  $r_i$  and  $B_i$  denote the  $i$ th rows of the matrices  $r$  and  $B$ , respectively.

Since  $B(t)$ ,  $t^{2m}B(t) \in L^1(\mathbb{R}^+, \rho)$ , it follows that  $t^m B(t) \in L^1(\mathbb{R}^+, \rho)$  and that  $H(A, 2m, \rho)$  holds. Moreover,  $\det \tilde{A}(z) = \det [I + \tilde{B}(z)]$  for  $\operatorname{Re} z \geq 0$ , and both (2.4) and (2.5) hold. Thus, by Theorem 2.1, for  $i = 1, \dots, n$ ,

$$r_i(t) = \sum_{j=1}^M p_j^i(t) e^{z_j t} + r_{i1}(t) \quad (t \in \mathbb{R}^+),$$

where, for each  $j$ ,  $p_j^i(t) = (p_{j1}^i(t), \dots, p_{jm}^i(t))$  with each  $p_{jk}^i(t)$  a polynomial of degree at most  $m_j - 1$  which depends only on  $A$  and  $B$ , and the row vector  $r_{i1}(t) \in L^1(\mathbb{R}^+, \rho)$ . Theorem 1.1 follows upon taking  $P_j(t)$  and  $r_1(t)$  to be the matrices with  $i$ th rows  $p_j^i(t)$  and  $r_{i1}(t)$ , respectively.

*B. Proof of Theorem 1.2.* Equation (1.2) may be written as

$$(5.1) \quad R'(t) + R \star A(t) = 0 \quad (R(0) = I, t \in \mathbb{R}^+),$$

where  $A(t) = -J(t)\mathcal{A} - \int_0^t B(s) ds$  ( $t \in R^+$ ) with  $J(t)$  defined in (3.3). The matrix equation (5.1) is equivalent to the  $n$  equations

$$(5.2) \quad R_i'(t) + R_i \star A(t) = 0 \quad (R_i(0) = I_i, t \in R^+),$$

where, for each  $i = 1, \dots, n$ ,  $R_i$  and  $I_i$  denote the  $i$ th rows of the matrices  $R$  and  $I$ , respectively.

It follows from  $t^{2m}B(t) \in L^1(R^+, \rho)$  that  $H(A, 2m, \rho)$  holds. Furthermore,  $\det [zI + \tilde{A}(z)] = \det [zI - \mathcal{A} - \tilde{B}(z)]$  for  $\operatorname{Re} z \geq 0$ . Thus, we may apply Theorem 2.2 to each of the equations (5.2) to find that  $R(t)$  satisfies (1.11) with  $r(t)$  and  $r_1(t)$  replaced by  $R(t)$  and  $R_1(t)$ , respectively, and with  $R_1(t)$  and  $R_1'(t)$  both in  $L^1(R^+, \rho)$ .

**Acknowledgment.** The authors thank the referees for their useful suggestions.

#### REFERENCES

- [1] I. M. GELFAND, *Über absolut konvergente trigonometrische Reihen und Integrale*, Mat. Sb., 9 (1941), pp. 51–66.
- [2] I. M. GELFAND, D. A. RAIKOV AND G. E. SHILOV, *Commutative Normed Rings*, Chelsea, New York, 1964.
- [3] S. I. GROSSMAN, *Integrability of resolvents of certain Volterra equations*, J. Math. Anal. Appl., 48 (1974), pp. 785–793.
- [4] S. I. GROSSMAN AND R. K. MILLER, *Perturbation theory for Volterra integrodifferential systems*, J. Differential Equations, 8 (1970), pp. 457–474.
- [5] ———, *Nonlinear Volterra integrodifferential systems with  $L^1$ -kernels*, Ibid., 13 (1973), pp. 551–566.
- [6] K. B. HANNSGEN, *An  $L^1$  remainder theorem for an integrodifferential equation with asymptotically periodic solution*, Proc. Amer. Math. Soc., 73 (1979), pp. 331–337.
- [7] G. S. JORDAN AND R. L. WHEELER, *A generalization of the Wiener–Lévy theorem applicable to some Volterra equations*, Ibid., 57 (1976), pp. 109–114.
- [8] ———, *Asymptotic behavior of unbounded solutions of linear Volterra integral equations*, J. Math. Anal. Appl., 55 (1976), pp. 596–615.
- [9] R. K. MILLER, *On Volterra integral equations with nonnegative integrable resolvents*, Ibid., 22 (1968), pp. 319–340.
- [10] ———, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
- [11] ———, *Structure of solutions of unstable linear Volterra integrodifferential equations*, J. Differential Equations, 15 (1974), pp. 129–157.
- [12] R. K. MILLER AND J. A. NOHEL, *A stable manifold theorem for a system of Volterra integro-differential equations*, this Journal, 6 (1975), pp. 506–522.
- [13] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, Amer. Math. Soc. Colloq. Publ., vol. 19, American Mathematical Society, Providence, RI, 1934.
- [14] 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.

## THE PRODUCT FORMULA AND CONVOLUTION STRUCTURE FOR THE GENERALIZED CHEBYSHEV POLYNOMIALS\*

THOMAS P. LAINE†

**Abstract.** The generalized Chebyshev polynomials  $T_n^{(\alpha,\beta)}(x)$ ,  $\alpha, \beta > -1$ , are the polynomials orthogonal on  $(-1, 1)$  with respect to the weight function  $(1-x^2)^\alpha |x|^{2\beta+1}$  and normalized by  $T_n^{(\alpha,\beta)}(1) = 1$ . We show that for certain  $(\alpha, \beta)$  the product formula

$$T_n^{(\alpha,\beta)}(x)T_n^{(\alpha,\beta)}(y) = \int_{-1}^1 T_n^{(\alpha,\beta)}(z) d\mu_{x,y}(z), \quad -1 < x, y < 1, \quad n = 0, 1, 2, \dots,$$

holds, where  $\mu_{x,y}$  is a real Borel measure which is independent of  $n$ , and explicitly determine  $\mu_{x,y}$ . We also completely determine the set of  $(\alpha, \beta)$  for which the product formula holds with  $\mu_{x,y}$  quasipositive; i.e.,

$$\int_{-1}^1 |d\mu_{x,y}(z)| \leq M, \quad -1 < x, y < 1,$$

where  $M$  does not depend on  $x$  and  $y$ . For certain  $(\alpha, \beta)$ ,  $\mu_{x,y}$  is absolutely continuous and so can be expressed in terms of a kernel  $k(x, y, z; \alpha, \beta)$ ; in this case we further determine for which  $(\alpha, \beta)$   $k(x, y, z; \alpha, \beta)$  is nonnegative for  $-1 < x, y, z < 1$ .

As an application, we show that a positive or quasipositive product formula allows the construction of a convolution structure for expansions in generalized Chebyshev polynomials.

**1. Introduction.** The generalized Chebyshev polynomials  $T_n^{(\alpha,\beta)}(x)$ ,  $\alpha, \beta > -1$ , are those polynomials normalized by  $T_n^{(\alpha,\beta)}(1) = 1$  which are orthogonal on  $(-1, 1)$  with respect to the weight function  $(1-x^2)^\alpha |x|^{2\beta+1}$ ; that is, they satisfy

$$(1.1) \quad \int_{-1}^1 T_n^{(\alpha,\beta)}(x)T_m^{(\alpha,\beta)}(x)(1-x^2)^\alpha |x|^{2\beta+1} dx = 0, \quad n \neq m.$$

The results of this paper concern the product formula

$$(1.2) \quad T_n^{(\alpha,\beta)}(x)T_n^{(\alpha,\beta)}(y) = \int_{-1}^1 T_n^{(\alpha,\beta)}(z) d\mu_{x,y}(z),$$

$$-1 < x, y < 1, \quad n = 0, 1, 2, \dots,$$

where  $\mu_{x,y}$  is a real Borel measure (which depends on  $\alpha, \beta, x, y$  but not on  $n$ ). If  $\mu_{x,y}$  is absolutely continuous and

$$d\mu_{x,y}(z) = k(x, y, z; \alpha, \beta)(1-z^2)^\alpha |z|^{2\beta-1} dz,$$

then (1.2) becomes

$$(1.3) \quad T_n^{(\alpha,\beta)}(x)T_n^{(\alpha,\beta)}(y) = \int_{-1}^1 T_n^{(\alpha,\beta)}(z)k(x, y, z; \alpha, \beta)(1-z^2)^\alpha |z|^{2\beta+1} dz.$$

Notice that if there is an integral representation of the form (1.2) or (1.3), then by virtue of the completeness of orthogonal polynomial systems on finite intervals, the kernel or measure is unique ( $k(x, y, \dots; \alpha, \beta)$  to within a.e. equivalence).

We will show that (1.2) holds for certain  $(\alpha, \beta)$  and explicitly determine the measure  $\mu_{x,y}$ . Moreover, we completely determine the set of  $(\alpha, \beta)$  for which (1.2) holds

\* Received by the editors March 7, 1979.

† Department of Mathematics, University of Alabama, University, Alabama 35486.

with the measure quasipositive; i.e.,

$$(1.4) \quad \int_{-1}^1 |d\mu_{x,y}(z)| \leq M, \quad -1 < x, y < 1,$$

where  $M$  is independent of  $x$  and  $y$ . We also completely determine those  $(\alpha, \beta)$  for which (1.3) holds with a positive kernel; i.e.

$$(1.5) \quad k(x, y, z; \alpha, \beta) \geq 0, \quad -1 < x, y, z < 1.$$

Setting  $n = 0$  in (1.2) and using  $T_0(x) \equiv 1$  shows that positivity implies quasipositivity.

Our main results are given in Theorems 1 and 2 below.

**THEOREM 1.** *Let  $\alpha, \beta > -1$  and  $-1 < x, y, z < 1, z \neq 0$ .*

(i) *If  $xy \neq 0$  and  $\alpha \geq \beta, \alpha + \beta > -1$  or if  $xy \neq 0, x^2 + y^2 \neq 1$  and  $\alpha \geq \beta, \alpha > -\frac{1}{2}$ , then (1.3) holds with  $k(x, y, z; \alpha, \beta)$  defined by (3.1)–(3.3) below.*

(ii) *If  $xy = 0$  and  $\alpha > \beta$ , then (1.3) holds with  $k(x, y, z; \alpha, \beta)$  defined by (2.3), (2.4), (2.6) and (2.7) below, while if  $\alpha = \beta$  and  $x = 0$  or  $y = 0$ , (1.2) holds with  $\mu_{0,y}$  or  $\mu_{x,0}$  the discrete measure with half-unit masses concentrated at  $z = \sqrt{1 - y^2}$  and  $z = -\sqrt{1 - y^2}$  or  $z = \sqrt{1 - x^2}$  and  $z = -\sqrt{1 - x^2}$ , respectively.*

(iii) *Formula (1.2) holds with  $\mu_{x,y}$  quasipositive if  $\alpha \geq \beta, \alpha + \beta > -1$  or  $\alpha = \beta = -\frac{1}{2}$ .*

(iv) *Formula (1.3) holds with  $k(x, y, z; \alpha, \beta) \geq 0$  if*

$$\alpha > \beta \geq -\frac{1}{2},$$

or if

$$\alpha \geq \beta \geq -\frac{1}{2}, \quad \alpha \neq -\frac{1}{2}, \quad xy \neq 0.$$

(v) *If  $\alpha = \beta = -\frac{1}{2}$ ,  $\mu_{x,y}$  is the discrete measure such that (1.2) takes the form*

$$(1.6) \quad T_n(\cos \theta)T_n(\cos \Psi) = \frac{1}{2}T_n(\cos(\theta - \Psi)) + \frac{1}{2}T_n(\cos(\theta + \Psi)).$$

These results are the best possible in the sense of

**THEOREM 2.** *Let  $\alpha, \beta > -1$ .*

(i) *If  $\alpha < \beta$  or  $\alpha + \beta \leq -1$  and  $(\alpha, \beta) \neq (-\frac{1}{2}, -\frac{1}{2})$ , then there does not exist an  $M$  independent of  $x$  and  $y$  for which (1.2) and (1.4) hold.*

(ii) *If  $\alpha < \beta$  or  $\beta < -\frac{1}{2}$  and  $k(x, y, z; \alpha, \beta)$  is defined by (3.1)–(3.3), then  $k(x, y, z; \alpha, \beta) < 0$  for some  $-1 < x, y, z < 1$  with  $xyz \neq 0$ .*

In the concluding section, we show that for those  $(\alpha, \beta)$  for which (1.4) or (1.5) hold, formulas (1.2) and (1.3) give a convolution structure for expansions in generalized Chebyshev polynomials.

These results are related to and rely in part upon Gasper's [4], [5] results on the product formula for the Jacobi polynomials  $P_n^{(\alpha, \beta)}(x)$ , by virtue of a connection between the two sets of polynomials. If

$$(1.7) \quad R_k^{(\alpha, \beta)}(x) = \frac{P_k^{(\alpha, \beta)}(x)}{P_k^{(\alpha, \beta)}(1)} = \tilde{F}(-k, k + \alpha + \beta + 1; \alpha + 1; \frac{1}{2}(1 - x)),$$

then the  $T_n^{(\alpha, \beta)}(x)$  are given by

$$(1.8) \quad T_n^{(\alpha, \beta)}(x) = \begin{cases} R_k^{(\alpha, \beta)}(2x^2 - 1) & \text{if } n = 2k, \\ xR_k^{(\alpha, \beta+1)}(2x^2 - 1) & \text{if } n = 2k + 1. \end{cases}$$

This follows from the orthogonality relation of the Jacobi polynomials,

$$(1.9) \quad \int_{-1}^1 R_n^{(\alpha,\beta)}(x)R_m^{(\alpha,\beta)}(x)(1-x)^\alpha(1+x)^\beta dx = 0, \quad n \neq m,$$

by means of the change of variables  $x = 2z^2 - 1$ .

If  $\beta = -\frac{1}{2}$ , we have the additional relation

$$(1.10) \quad T_n^{(\alpha,-1/2)}(x) = \frac{C_n^{\alpha+1/2}(x)}{C_n^{\alpha+1/2}(1)} = \frac{n!}{(2\alpha+1)_n} C_n^{\alpha+1/2}(x),$$

where  $C_n^\lambda(x)$ ,  $\lambda > -\frac{1}{2}$ , is the Gegenbauer polynomial of degree  $n$ . This follows from the quadratic transformations [7, p. 59]

$$R_{2n}^{(\alpha,\alpha)}(x) = R_n^{(\alpha,-1/2)}(2x^2 - 1),$$

$$R_{2n+1}^{(\alpha,\alpha)}(x) = xR_n^{(\alpha,1/2)}(2x^2 - 1).$$

A special case of (1.10) is

$$(1.11) \quad T_n^{(-1/2,-1/2)}(\cos \theta) = T_n(\cos \theta) = \cos n\theta,$$

where  $T_n(x)$  is the Chebyshev polynomial of the first kind.

There are some important differences between the  $T_n^{(\alpha,\beta)}(x)$  and the Jacobi and Gegenbauer polynomials, despite these relations. Notice from (1.8) that  $T_n^{(\alpha,\beta)}(x)$  is even or odd according as  $n$  is even or odd, which is true for the Jacobi polynomials only in the Gegenbauer case  $\alpha = \beta$ . Also, unlike the classical orthogonal polynomials, the weight function for  $T_n^{(\alpha,\beta)}(x)$  has a zero (if  $\beta > -\frac{1}{2}$ ) or a singularity (if  $\beta < -\frac{1}{2}$ ) within the interval of orthogonality.

Moreover, Theorems 1 and 2 reflect some unexpected differences from the Jacobi case. For Gasper [5] has shown that the product formula for Jacobi polynomials holds with a positive kernel if and only if

$$\alpha \geq \beta \geq -\frac{1}{2} \quad \text{or} \quad \alpha + \beta \geq 0, \quad -1 < \beta < -\frac{1}{2},$$

and that quasipositivity holds if and only if  $\alpha + \beta \geq -1$ ,  $\alpha > \beta > -1$ ,  $\alpha > -\frac{1}{2}$ ; these regions are larger than in our case. Furthermore, the singularity or zero of the weight function for the  $T_n^{(\alpha,\beta)}(x)$  is manifested in a singularity in the kernel when  $x = 0$  or  $y = 0$ .

**2. A relation between kernels.** Gasper [5] has derived an explicit formula for the kernel  $K(x, y, z; \alpha, \beta)$  in the product formula

$$(2.1) \quad R_k^{(\alpha,\beta)}(x)R_k^{(\alpha,\beta)}(y) = \int_{-1}^1 R_k^{(\alpha,\beta)}(z)K(x, y, z; \alpha, \beta)(1-z)^\alpha(1+z)^\alpha dz$$

and has shown that (2.1) holds, for  $-1 < x, y, z < 1$ , if  $\alpha \geq \beta > -1$ ,  $\alpha + \beta > -1$  or if  $\alpha \geq -\frac{1}{2}$ ,  $\alpha + \beta = -1$ ,  $x \neq y$ . Using this result, we shall first show that for  $\alpha \geq \beta + 1$ , (1.3) holds with  $k(x, y, z; \alpha, \beta)$  given by

$$(2.2) \quad k(x, y, z; \alpha, \beta) = 2^{\alpha+\beta+1}[K(2x^2 - 1, 2y^2 - 1, 2z^2 - 1; \alpha, \beta) + 2xyzK(2x^2 - 1, 2y^2 - 1, 2z^2 - 1; \alpha, \beta + 1)]$$

if  $-1 < x, y, z < 1$ ,  $xyz \neq 0$ , and later show that (2.2) can be extended to a larger region in the  $(\alpha, \beta)$  plane.

To prove (2.2), replace  $x$  with  $2x^2 - 1$  and  $y$  with  $2y^2 - 1$  in (2.1), change variables by  $z = 2t^2 - 1$  and use (1.8) to obtain

$$\begin{aligned} T_{2k}^{(\alpha,\beta)}(x)T_{2k}^{(\alpha,\beta)}(y) &= R_k^{(\alpha,\beta)}(2x^2 - 1)R_k^{(\alpha,\beta)}(2y^2 - 1) \\ &= 2^{\alpha+\beta+2} \int_0^1 T_{2k}^{(\alpha,\beta)}(t)K(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta) \\ &\qquad \qquad \qquad \cdot (1 - t^2)^\alpha t^{2\beta+1} dt \\ &= 2^{\alpha+\beta+1} \int_{-1}^1 T_{2k}^{(\alpha,\beta)}(t)K(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta) \\ &\qquad \qquad \qquad \cdot (1 - t^2)^\alpha |t|^{2\beta+1} dt. \end{aligned}$$

Clearly,

$$2xy \int_{-1}^1 T_{2k}^{(\alpha,\beta)}(t)tK(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta + 1) \cdot (1 - t^2)^\alpha |t|^{2\beta+1} dt = 0,$$

since the integrand is odd. Hence when  $n$  is even and  $\alpha \geq \beta > -1$ , (1.3) holds with  $k(x, y, z; \alpha, \beta)$  defined by (2.2).

Similarly,

$$\begin{aligned} T_{2k+1}^{(\alpha,\beta)}(x)T_{2k+1}^{(\alpha,\beta)}(y) &= xyR_k^{(\alpha,\beta+1)}(2x^2 - 1)R_k^{(\alpha,\beta+1)}(2y^2 - 1) \\ &= 2^{\alpha+\beta+3} xy \int_0^1 R_k^{(\alpha,\beta+1)}(2t^2 - 1)K(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta + 1) \\ &\qquad \qquad \qquad \cdot (1 - t^2)^\alpha |t|^{2\beta+3} dt \\ &= 2^{\alpha+\beta+2} \int_{-1}^1 T_{2k+1}^{(\alpha,\beta)}(t)xytK(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta + 1) \\ &\qquad \qquad \qquad \cdot (1 - t^2)^\alpha |t|^{2\beta+1} dt, \end{aligned}$$

and again by the oddness of the integrand

$$\int_{-1}^1 T_{2k+1}^{(\alpha,\beta)}(t)K(2x^2 - 1, 2y^2 - 1, 2t^2 - 1; \alpha, \beta)(1 - t^2)^\alpha |t|^{2\beta+1} = 0.$$

Hence (1.3) holds for all  $n$  with  $k(x, y, z; \alpha, \beta)$  given for  $-1 < x, y, z < 1$  and  $xyz \neq 0$  by (2.2) if  $\alpha > \beta + 1 > 0$ . In § 5 we will show that (2.2) can be analytically continued so as to hold for other  $(\alpha, \beta)$ .

The restriction  $xy \neq 0$  in (2.2) is necessary because the case  $x = 0$  or  $y = 0$  in  $k(x, y, z; \alpha, \beta)$  corresponds by (2.2) to the case  $x = -1$  or  $y = -1$  respectively in  $K(x, y, z; \alpha, \beta)$ . However, although this was not mentioned in [5], for  $\alpha > \beta > -1$ , (2.1) holds if for example  $x = -1$  with

$$(2.3) \quad K(-1, y, z; \alpha, \beta) = \frac{\Gamma(\alpha + 1)}{\Gamma(\beta + 1)\Gamma(\alpha - \beta)} \frac{(-y - z)^{\alpha - \beta - 1}}{(1 - y)^\alpha (1 - z)^\alpha} \quad \text{if } -1 < z < -y$$

and

$$(2.4) \quad K(-1, y, z; \alpha, \beta) = 0 \quad \text{if } -y < z < 1.$$



This follows, as in [1, p. 31], from

$$(1+y)^{\beta+\mu} \frac{P_n^{(\alpha-\mu, \beta+\mu)}(y)}{P_n^{(\alpha-\mu, \beta+\mu)}(-1)}$$

$$= \frac{\Gamma(\beta+\mu+1)}{\Gamma(\beta+1)\Gamma(\mu)} \cdot \int_{-1}^y (1+z)^\beta \frac{P_n^{(\alpha, \beta)}(z)}{P_n^{(\alpha, \beta)}(-1)} (y-z)^{\mu-1} dz, \quad \mu > 0, \quad -1 \leq y < 1$$

which is (3.4) of [2]. Just set  $\mu = \alpha - \beta$ , use

$$P_n^{(\beta, \alpha)}(y) = P_n^{(\beta, \alpha)}(-1)R_n^{(\alpha, \beta)}(-y)$$

and then replace  $y$  with  $-y$  to get

$$(2.5) \quad R_n^{(\alpha, \beta)}(-1)R_n^{(\alpha, \beta)}(y) = \int_{-1}^1 K(-1, y, z)R_n^{(\alpha, \beta)}(z)(1-z)^\alpha(1+z)^\beta,$$

$$-1 \leq y < 1, \quad \alpha > \beta > -1,$$

with  $K(-1, y, z)$  given by (2.3) and (2.4). The case  $y = 0$ , of course, can be dealt with by symmetry.

Hence, by an argument similar to that used to prove (2.2), if  $xy = 0$  and  $-1 < x, y < 1$ , (1.3) holds with

$$(2.6) \quad k(0, y, z; \alpha, \beta) = 2^{\alpha+\beta+1}K(-1, 2y^2-1, 2z^2-1; \alpha, \beta), \quad -1 < y, z < 1,$$

$$(2.7) \quad k(x, 0, z; \alpha, \beta) = 2^{\alpha+\beta+1}K(-1, 2x^2-1, 2z^2-1; \alpha, \beta), \quad -1 < x, z < 1,$$

where  $K(-1, y, z; \alpha, \beta)$  is given by (2.3) and (2.4) and  $\alpha > \beta > -1$ .

On the other hand, since  $T_n^{(\alpha, \beta)}(-y) = (-1)^n T_n^{(\alpha, \beta)}(y)$  and

$$R_k^{(\alpha, \alpha)}(-1)R_k^{(\alpha, \alpha)}(2y^2-1) = R_k^{(\alpha, \alpha)}(1-2y^2),$$

it follows that if  $\alpha = \beta$  and  $x = 0$ , then (1.7) holds if  $\mu_{0,y}$  is the discrete measure with half-unit masses concentrated at  $\sqrt{1-y^2}$  and  $-\sqrt{1-y^2}$ . Similarly if  $y = 0$ .

This completes the proof of Theorem 1 (ii). Since the statements about positivity and quasipositivity in Theorem 1 (iii) and (iv) for the case  $xy = 0$  are now obvious from (2.6), (2.7), (2.3), and (2.4) and the remark above, we will henceforth assume that  $xy \neq 0$  (as well as  $-1 < x, y, z < 1, z \neq 0$ ).

Because of (1.11), (1.6) is just the identity

$$\cos n\phi \cos n\Psi = \frac{1}{2} \cos n(\phi - \Psi) + \frac{1}{2} \cos n(\phi + \Psi)$$

and so we may also assume that  $\alpha$  and  $\beta$  are not both  $-\frac{1}{2}$ .

**3. The kernel  $k(x, y, z; \alpha, \beta)$ .** Supposing, as we now may, that  $-1 < x, y, z < 1$  and  $xyz \neq 0$ , let  $0 < \phi, \Psi, \theta < \pi/2$  and set  $\cos \phi = |x|, \cos \Psi = |y|, \cos \theta = |z|, a = \sin \phi \sin \Psi, b = \cos \phi \cos \Psi, c = \cos \theta$ ,

$$B = \frac{b^2 + c^2 - a^2}{2bc}.$$

Then  $0 < a, b, c < 1$  and  $2x^2-1 = \cos 2\phi, 2y^2-1 = \cos 2\Psi, 2z^2-1 = \cos 2\theta$  and  $|xyz| = bc$ . Thus, using the expressions [5, (3.3)–(3.5)] for  $K(x, y, z; \alpha, \beta)$ , it follows from (1.10)

that if  $|a - b| < c < a + b$ , then

$$\begin{aligned}
 k(x, y, z; \alpha, \beta) &= \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(bc)^{\alpha - \beta - 1}(1 - B^2)^{\alpha - 1/2}}{2\Gamma(\alpha + \frac{1}{2})\Gamma(\frac{1}{2})}, \\
 & [F(\alpha - \beta, \alpha + \beta; \alpha + \frac{1}{2}; \frac{1}{2}(1 - B)) \\
 & \quad + \operatorname{sgn}(xyz)F(\alpha - \beta - 1, \alpha + \beta + 1; \alpha + \frac{1}{2}; \frac{1}{2}(1 - B))] \\
 (3.1a) \quad & = \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(bc)^{\alpha - \beta - 1}(1 - B)^{\alpha - 1/2}}{2^{(3/2 - \alpha)}\Gamma(\alpha + \frac{1}{2})\Gamma(\frac{1}{2})},
 \end{aligned}$$

$$(3.1b) \quad [F(\frac{1}{2} + \beta, \frac{1}{2} - \beta; \alpha + \frac{1}{2}; \frac{1}{2}(1 - B)) + \operatorname{sgn}(xyz)F(\frac{3}{2} + \beta, -\frac{1}{2} - \beta; \alpha + \frac{1}{2}; \frac{1}{2}(1 - B))]$$

if  $c < a - b$ , then

$$\begin{aligned}
 k(x, y, z; \alpha, \beta) &= \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(a^2 - b^2 - c^2)^{\alpha - \beta - 1}(1 - B^{-2})^{\alpha - (1/2)}}{\Gamma(\alpha - \beta)\Gamma(\beta + 2)}, \\
 & [(\beta + 1)F(\frac{1}{2}(\alpha + \beta), \frac{1}{2}(\alpha + \beta + 1); \beta + 1; B^{-2}) \\
 & \quad - \operatorname{sgn}(xyz)\frac{(\alpha - \beta - 1)}{2B}F(\frac{1}{2}(\alpha + \beta + 1), \frac{1}{2}(\alpha + \beta + 2); \beta + 2, B^{-2})] \\
 (3.2a) \quad & = \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(a^2 - b^2 - c^2)^{\alpha - \beta - 1}}{\Gamma(\alpha - \beta)\Gamma(\beta + 2)},
 \end{aligned}$$

$$\begin{aligned}
 (3.2b) \quad & [(\beta + 1)F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \beta + 1; B^{-2}) \\
 & \quad - \operatorname{sgn}(xyz)\frac{(\alpha - \beta - 1)}{2B}F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \beta + 2, B^{-2})]
 \end{aligned}$$

and if either  $c < b - a$  or  $c > a + b$  then

$$(3.3) \quad k(x, y, z; \alpha, \beta) = 0.$$

The kernel is defined to be zero if  $c = |a \pm b|$ .

The two expressions for the kernel in each of (3.1) and (3.2) are related to each other by means of the transformation ([3, p. 105])

$$(3.4) \quad F(\lambda, \gamma; \delta; x) = (1 - x)^{\delta - \lambda - \gamma}F(\delta - \lambda, \delta - \gamma; \delta).$$

For certain  $\alpha, \beta$ ,  $k(x, y, z; \alpha, \beta)$  reduces to an elementary function. In particular, if  $a - b < c < a + b$ , then

$$(3.5) \quad k(x, y, z; \alpha, -\alpha - 1) = \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(bc)^{2\alpha}(1 - B^2)^{\alpha - 1/2}(B + \operatorname{sgn}(xyz))}{2\Gamma(\alpha + \frac{1}{2})\Gamma(\frac{1}{2})},$$

$$(3.6) \quad k(x, y, z; \alpha, \alpha) = \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(bc)^{-1}(1 - B^2)^{\alpha - 1/2}(1 + \operatorname{sgn}(xyz)B)}{2\Gamma(\alpha + \frac{1}{2})\Gamma(\frac{1}{2})},$$

and if  $c < a - b$ , then

$$\begin{aligned}
 (3.7) \quad & k(x, y, z; \alpha, -\alpha - 1) \\
 & = \frac{\Gamma(\alpha + 1)a^{-2\alpha}(1 - c^2)^{-\alpha}(a^2 - b^2 - c^2)^{2\alpha}(1 - B^2)^{\alpha - 1/2}(1 + \operatorname{sgn}(xyz)B^{-1})}{\Gamma(2\alpha + 1)\Gamma(-\alpha)},
 \end{aligned}$$

$$(3.8) \quad k(x, y, z; \alpha, \alpha) = 0.$$

Formulas (3.5)–(3.8) follow from (3.1a) and (3.2a) by virtue of  $F(0, \gamma; \delta; x) = 1$ ,

$F(-1, \gamma; \delta; x) = 1 - (\gamma/\delta)x$  and

$$\frac{1}{\Gamma(\alpha - \beta)} = \frac{\alpha - \beta}{\Gamma(\alpha - \beta + 1)}.$$

**4. Positivity of  $k(x, y, z; \alpha, \beta)$ .** In this section we prove that  $k(x, y, z; \alpha, \beta)$  as given for  $\alpha \neq -\frac{1}{2}$  and  $xyz \neq 0$  by (3.1) to (3.3) is nonnegative if and only if  $\alpha \geq \beta \geq -\frac{1}{2}$ ,  $\alpha \neq -\frac{1}{2}$ .

We first show that  $k(x, y, z; \alpha, \beta) \geq 0$  if  $\alpha \geq \beta \geq -\frac{1}{2}$ ,  $\alpha \neq -\frac{1}{2}$ . Since this is clear for the case  $\alpha = \beta$  from (3.6), (3.8), and

$$(4.1) \quad -1 < B < 1 \quad \text{if } |a - b| < c < a + b,$$

$$(4.2) \quad B < -1 \quad \text{if } c < a - b$$

which are easily verified, we can assume  $\alpha > \beta$ . We consider the cases  $|a - b| < c < a + b$  and  $c < a - b$  separately.

**I.  $|a - b| < c < a + b$ .** By (4.1) and  $0 < a, b, c < 1$ , the function multiplying the bracketed hypergeometric functions in (3.1) is positive, provided  $\alpha > -\frac{1}{2}$ . So for  $\alpha > \beta \geq -\frac{1}{2}$ ,  $k(x, y, z; \alpha, \beta)$  will be nonnegative provided

$$(4.3) \quad F(\alpha - \beta, \alpha + \beta; \alpha + \frac{1}{2}; W) + \text{sgn}(xyz)F(\alpha - \beta - 1, \alpha + \beta + 1; \alpha + \frac{1}{2}; W)$$

is nonnegative for  $0 \leq W = \frac{1}{2}(1 - B) < 1$ .

Taking the cases  $\text{sgn}(xyz) = \pm 1$  of (4.3) separately, we have

$$(4.4) \quad F(\alpha - \beta, \alpha + \beta; \alpha + \frac{1}{2}; W) + F(\alpha - \beta - 1, \alpha + \beta + 1; \alpha + \frac{1}{2}; W) \\ = 2(1 - W)F(\alpha - \beta, \alpha + \beta + 1; \alpha + \frac{1}{2}; W),$$

$$(4.5) \quad F(\alpha - \beta, \alpha + \beta; \alpha + \frac{1}{2}; W) - F(\alpha - \beta - 1, \alpha + \beta + 1; \alpha + \frac{1}{2}; W) \\ = \frac{2\beta + 1}{\alpha + 1/2} W F(\alpha - \beta, \alpha + \beta + 1; \alpha + \frac{3}{2}; W).$$

Formula (4.4) follows from [3, 2.8, (37)] and (4.5) may be proved by comparing the coefficients of powers of  $W$ .

Since the nonnegativity of the right-hand sides of (4.4) and (4.5) is obvious for  $\alpha > \beta \geq -\frac{1}{2}$ ,  $0 \leq W < 1$ , we are done.

**II.  $c < a - b$ .** By (3.2b) and (4.2),  $k(x, y, z; \alpha, \beta)$  is nonnegative for these  $x, y, z$  if

$$(4.6) \quad h(w; \alpha, \beta) \equiv (\beta + 1)F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \beta + 1; w^2) \\ + \frac{1}{2}(\beta - \alpha + 1)wF(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \beta + 2; w^2)$$

is nonnegative for  $-1 < w < 1$ . We divide the region  $\alpha > \beta \geq -\frac{1}{2}$  into two subsets.

(A)  $0 \leq \beta - \alpha + 1 < 1$ ,  $\beta > -\frac{1}{2}$ . Notice that the second hypergeometric function on the right-hand side of (4.6) is positive. Hence for  $-1 < w < 1$  and  $\alpha$  and  $\beta$  in this region,

$$h(w; \alpha, \beta) \geq (\beta + 1)F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \beta + 1; w^2) \\ - \frac{1}{2}(\beta - \alpha + 1)F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \beta + 2; w^2) \\ = \frac{1}{2}(\alpha + \beta + 1)F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \beta + 2; w^2) \\ \geq 0,$$

where we have used the contiguous function relation [3, (2.8), (3.5)].

(B)  $\beta - \alpha + 1 < 0$ ,  $\beta > -\frac{1}{2}$ . To show  $h(w; \alpha, \beta) \geq 0$  in this region, it suffices to

consider the case  $0 < w < 1$  since  $\beta - \alpha + 1 < 0$  and, by (3.4), both hypergeometric functions in (4.6) are positive. In this case we have to use a different method, which relies on Bateman's integral ([3, 2.4, (2)]),

$$(4.7) \quad F(a, b, c + \mu; x) = \frac{\Gamma(c + \mu)}{\Gamma(c)\Gamma(\mu)} \int_0^1 y^{c-1}(1-y)^{\mu-1} F(a, b; c; xy) dy$$

if  $\mu > 0$ ,  $c > 0$ , and  $-1 < x < 1$ . Also, we need the quadratic transformation

$$(4.8) \quad F(\gamma, \gamma + \frac{1}{2}; \frac{1}{2}; x^2) + 2\gamma x F(\gamma + \frac{1}{2}, \gamma + 1; \frac{3}{2}; x^2) = (1-x)^{-2\gamma}$$

which follows from [3, 2.11, (3) and 2.3, (4)]. If  $\beta = -\frac{1}{2}$ , then by means of (4.8), (4.6) becomes

$$h(w; \alpha, -\frac{1}{2}) = \frac{1}{2}(1-w)^{\alpha-1/2}, \quad -1 < w < 1,$$

so we may assume  $\beta > -\frac{1}{2}$ .

Setting  $a = \frac{1}{2}(\beta - \alpha + 1)$ ,  $b = \frac{1}{2}(\beta - \alpha + 2)$ ,  $c = \frac{1}{2}$ ,  $\mu = \beta + \frac{1}{2}$ , and  $x = w^2$  in (4.7) and making the change of variables  $y = z^2$  gives

$$(4.9) \quad F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \beta + 1; w^2) = \frac{2\Gamma(\beta + 1)}{\Gamma(\beta + \frac{1}{2})\Gamma(\frac{1}{2})} \cdot \int_0^1 (1-z^2)^{\beta-1/2} F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \frac{1}{2}; w^2 z^2) dz.$$

Similarly, setting  $a = \frac{1}{2}(\beta - \alpha + 2)$ ,  $b = \frac{1}{2}(\beta - \alpha + 3)$ ,  $c = \frac{3}{2}$ ,  $\mu = \beta + \frac{1}{2}$ , and  $x = w^2$  and making the same change of variables gives

$$(4.10) \quad F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \beta + 2; w^2) = \frac{2\Gamma(\beta + 2)}{\Gamma(\beta + \frac{1}{2})\Gamma(\frac{3}{2})} \cdot \int_0^1 z^2 (1-z^2)^{\beta-1/2} F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \frac{3}{2}; w^2 z^2) dz.$$

Hence, by (4.6), (4.9), and (4.10),

$$(4.11) \quad h(w; \alpha, \beta) = \frac{2\Gamma(\beta + 2)}{\Gamma(\beta + \frac{1}{2})\Gamma(\frac{1}{2})} \int_0^1 (1-z^2)^{\beta-1/2} \cdot [F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \frac{1}{2}; w^2 z^2) + (\beta - \alpha + 1)wz^2 F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \frac{3}{2}; w^2 z^2)] dz \\ \cong \frac{2\Gamma(\beta + 2)}{\Gamma(\beta + \frac{1}{2})\Gamma(\frac{1}{2})} \int_0^1 (1-z^2)^{\beta-1/2} \cdot [F(\frac{1}{2}(\beta - \alpha + 1), \frac{1}{2}(\beta - \alpha + 2); \frac{1}{2}; w^2 z^2) + (\beta - \alpha + 1)wz F(\frac{1}{2}(\beta - \alpha + 2), \frac{1}{2}(\beta - \alpha + 3); \frac{3}{2}; w^2 z^2)] dz,$$

since  $z^2 < z$  for  $0 < z < 1$ ,  $\beta - \alpha + 1 < 0$ , and, by (3.4), the second hypergeometric function in the integrand is positive. Then (4.11) and (4.8) with  $\gamma = \frac{1}{2}(\beta - \alpha + 1)$  give

$$h(w; \alpha, \beta) \cong \frac{2\Gamma(\beta + 2)}{\Gamma(\beta + \frac{1}{2})\Gamma(\frac{1}{2})} \int_0^1 (1-z^2)^{\beta-1/2} (1-wz)^{\alpha-\beta-1} dz \cong 0$$

for  $\alpha > \beta > -\frac{1}{2}$ , as required.

This completes the proof of Theorem 1(iv). Since we will show in § 5 that quasipositivity, and thus positivity, fails if  $\beta > \alpha$  or  $\alpha + \beta < -1$ , to prove Theorem 2(ii) it suffices to show that positivity fails for  $-1 < \beta < -\frac{1}{2}$ ,  $\alpha > -\frac{1}{2}$ ,  $\alpha + \beta \cong -1$ . But this is clear from (4.5).

**5. Analytic continuation.** Here we show that the integral representation (1.3) holds with  $k(x, y, z; \alpha, \beta)$  given by (3.1)–(3.3) for  $\alpha, \beta > -1$  and  $-1 < x, y, z < 1$ ,  $xyz \neq 0$  if

$$\alpha \cong \beta, \quad \alpha + \beta > -1,$$

or if

$$\alpha \cong \beta, \quad \alpha > -\frac{1}{2} \quad \text{and} \quad x^2 + y^2 \neq -1$$

and show that it fails on the part of the boundary of these sets excluded above.

If  $s$  is fixed and  $s < 1$ , then  $F(\lambda, \gamma; \nu; s)/\Gamma(\nu)$  is an entire analytic function of  $\gamma, \lambda$ , and  $\nu$ . It therefore follows from (1.7), (1.8), and (3.1)–(3.3) that, for fixed  $x, y, z$ ,  $T_n^{(\alpha, \beta)}(x)$  and  $k(x, y, z; \alpha, \beta)$  are analytic functions of the complex parameters  $\alpha, \beta$  for  $\text{Re}(\alpha) > -1$ . Hence the extension of (1.3) to the desired regions can be accomplished by an analytic continuation argument. Since the required argument is a simple modification of the analytic continuation proof of [5], we omit the details.

The integral representation (1.3) fails, however, on the line segment  $\alpha + \beta = -1$ ,  $-\frac{1}{2} < \alpha < 0$  if  $x^2 + y^2 = 1$ . In fact, on this segment,

$$(5.1) \quad T_n^{(\alpha, -\alpha-1)}(x) T_n^{(\alpha, -\alpha-1)}(\pm\sqrt{1-x^2}) = \frac{1}{2} T_n^{(\alpha, -\alpha-1)}(0) \cdot \int_{-1}^1 T_n^{(\alpha, -\alpha-1)}(z) k(x, \pm\sqrt{1-x^2}, z; \alpha, -\alpha-1) (1-z^2)^\alpha |z|^{-2\alpha-1} dz.$$

To prove (5.1), note that if  $n = 2k$ , (5.1) becomes

$$R_k^{(\alpha, -\alpha-1)}(2x^2-1) R_k^{(\alpha, -\alpha-1)}(1-2x^2) = \frac{1}{2} R_k^{(\alpha, \alpha-1)}(1) \cdot \int_{-1}^1 R_k^{(\alpha, \alpha-1)}(2z^2-1) K(2x^2-1, 1-2x^2, 2z^2-1; \alpha, -\alpha-1) \cdot (1-z^2)^\alpha |z|^{-2\alpha-1} dz$$

which, after a change of variables, is just [5, (6.4)]. If  $n = 2k + 1$ , then (5.1) becomes

$$R_k^{(\alpha, -\alpha)}(2x^2-1) R_k^{(\alpha, -\alpha)}(1-2x^2) = 2 \int_{-1}^1 R_k^{(\alpha, -\alpha)}(2z^2-1) K(2x^2-1, 1-2x^2, 2z^2-1; \alpha, -\alpha) (1-z^2)^\alpha |z|^{1-2\alpha} dz$$

or, after a change of variables,

$$(5.2) \quad R_k^{(\alpha, -\alpha)}(x) R_k^{(\alpha, -\alpha)}(-x) = \int_{-1}^1 R_k^{(\alpha, -\alpha)}(z) K(x, -x, z; \alpha, -\alpha) (1-z)^\alpha (1+z)^{-\alpha} dz$$

for  $-\frac{1}{2} < \alpha < 0$ . Now (5.2) is (2.1) when  $\alpha + \beta = 0$ , which is stated in [5] for  $\alpha > 0$ . Since by [5, (3.3)–(3.5)], the integral on the right-hand side of (5.2) is

$$\frac{\Gamma(\alpha+1)a^{-1}}{\Gamma(\alpha+\frac{1}{2})\Gamma(\frac{1}{2})} \int_0^{2a} R_k^{(\alpha, -\alpha)}(2c^2-1) \left[1 - \frac{c^2}{4a^2}\right]^{\alpha-1/2} dc,$$

(5.2) can be analytically continued to  $\alpha > -\frac{1}{2}$ , completing the verification of (5.2).

Finally, we note that  $\mu_{x,y}$  is also not absolutely continuous on the half-line  $\alpha = -\frac{1}{2}$ ,  $\beta > -1$ , even if  $x^2 + y^2 \neq 1$ . This follows from [5, (6.13)] which in terms of our definitions of  $a$ ,  $b$ , and  $c$  is

$$\begin{aligned}
 & R_n^{(-1/2,\beta)}(2x^2-1)R_n^{(-1/2,\beta)}(2y^2-1) \\
 &= \frac{1}{2} \left(\frac{b-a}{b}\right)^{\beta+1/2} R_n^{(-1/2,\beta)}(2[b-a]^2-1) \\
 (5.3) \quad & + \frac{1}{2} \left(\frac{a+b}{b}\right)^{\beta+1/2} R_n^{(-1/2,\beta)}(2[a+b]^2-1) \\
 & + \frac{1}{2}(\frac{1}{4}-\beta^2)ab^{-\beta-3/2} \int_{b-a}^{a+b} c^{\beta-1/2}(1+B)^{-1} R_n^{(-1/2,\beta)}(2c^2-1) \\
 & \quad \cdot F(\frac{1}{2}-\beta, \frac{1}{2}+\beta; 2; \frac{1}{2}(1-B)) dc
 \end{aligned}$$

for  $-1 < \beta < -\frac{1}{2}$ ,  $x^2 + y^2 > 1$  (i.e.,  $b > a$ ). For then by analytic continuation, (5.3) can be extended to  $\beta > -1$ , and so by an argument similar to that used to prove (2.2), we have that for  $-1 < \beta < -\frac{1}{2}$  and  $x^2 + y^2 > 1$ ,

$$\begin{aligned}
 & T_n^{(-1/2,\beta)}(x)T_n^{(-1/2,\beta)}(y) \\
 &= \frac{1}{2} \left(\frac{b-a}{b}\right)^{\beta+1/2} T_n^{(-1/2,\beta)}(\operatorname{sgn}(xy)[b-a]) \\
 & + \frac{1}{2} \left(\frac{b+a}{b}\right)^{\beta+1/2} T_n^{(-1/2,\beta)}(\operatorname{sgn}(xy)[b+a]) \\
 & + \frac{1}{4}ab^{-\beta-3/2} \int_E |c|^{\beta-1/2}(1+|B|)^{-1} T_n^{(-1/2,\beta)}(c) \\
 & \quad \cdot \{(\frac{1}{4}-\beta^2)F(\frac{1}{2}-\beta, \frac{1}{2}+\beta; 2; \frac{1}{2}(1-|B|)) \\
 & \quad + \operatorname{sgn}(xyc)[\frac{1}{4}-(\beta+1)^2]F(-\frac{1}{2}-\beta, \frac{3}{2}+\beta; 2; \frac{1}{2}(1-|B|))\} dc
 \end{aligned}$$

where  $E = (-a-b, a-b) \cup (b-a, a+b)$ . If  $\beta = -\frac{1}{2}$ , (5.4) reduces of course to (1.6).

**6. Estimates for  $k(x, y, z; \alpha, \beta)$ .** Here we show that  $\mu_{x,y}$  satisfies (1.4) if and only if  $\alpha = \beta = -\frac{1}{2}$  (which follows from (1.6)) or

$$(6.1) \quad \alpha \geq \beta > -1, \quad \alpha + \beta > -1.$$

Except for the exclusion of the line segment  $\alpha + \beta = -1$ ,  $-\frac{1}{2} < \alpha < 0$ , (6.1) is the same region for which quasipositivity holds for the Jacobi polynomials, as in [5]. In view of the results of [5],  $\mu_{x,y}$  is absolutely continuous in the region (6.1), with  $d\mu_{x,y} = k(x, y, z; \alpha, \beta)$  as given by (3.1) to (3.3). Hence, to show that  $\mu_{x,y}$  is quasipositive in the region (6.1) it is enough to show that

$$(6.2) \quad V_1 \equiv \int_{|a-b|}^{a+b} |k(x, y, z; \alpha, \beta)|(1-c^2)^\alpha c^{2\beta+1} dc \leq M,$$

$$(6.3) \quad V_2 \equiv \int_0^{a-b} |k(x, y, z; \alpha, \beta)|(1-c^2)^\alpha c^{2\beta+1} dc \leq M$$

with  $z = c$  or  $-c$ .

The proof of quasipositivity for much of (6.1) is virtually identical to that in [5] for Jacobi polynomials. In fact, in (3.1) and (3.2) the functions multiplying the quantities in

brackets are the same as the functions multiplying  $F$ 's in the expressions for the Jacobi kernel in [5, (3.3) and (3.4)]; the weight function expressed in terms of  $c$  is also the same, i.e.,  $(1 - c^2)^\alpha c^{2\beta+1}$ . Moreover, using (4.1) and (4.2) and the fact that  $F(\lambda, \gamma; \nu; \omega)$  is a continuous function of  $\omega$  for  $0 \leq \omega \leq 1$  if  $\text{Re}(\nu - \lambda - \gamma) > 0$ , it is easy to check that the quantities in brackets in (3.1) and (3.2) are bounded by a constant independent of  $a, b$  and  $c$  for exactly the same  $\alpha, \beta$  as the  $F$ 's in (3.3) and (3.4) of [5]. Thus as this is the only property of the  $F$ 's that is used there to establish quasipositivity for the regions

$$(6.4) \quad \alpha + \beta > -1, \quad -\frac{1}{2} < \alpha < \frac{1}{2}, \quad -1 < \beta < -\frac{1}{2},$$

$$(6.5) \quad \alpha > +\frac{1}{2}, \quad -1 < \beta < -\frac{1}{2},$$

quasipositivity follows in our case in exactly the same way, and we do not repeat the proof. Also, the estimate used in [5] to deal with the case

$$(6.6) \quad \alpha = \frac{1}{2}, \quad -1 < \beta < -\frac{1}{2},$$

namely

$$\lim_{s \rightarrow 1} \frac{F(\lambda, \gamma; \lambda + \gamma; s)}{\log [1/(1-s)]} = \frac{\Gamma(\lambda + \gamma)}{\Gamma(\lambda)\Gamma(\gamma)}$$

applies in exactly the same way to our case for the set (6.6).

The remaining subset of (6.1),

$$(6.7) \quad \alpha \geq \beta \geq -\frac{1}{2}, \quad \alpha > -\frac{1}{2}$$

is easily disposed of by virtue of the positivity of  $k(x, y, z; \alpha, \beta)$  for  $\alpha, \beta$  in (6.7). For, as  $T_0^{(\alpha, \beta)}(z) \equiv 1$ , setting  $n = 0$  in (1.3) gives

$$\int_{-1}^1 |k(x, y, z; \alpha, \beta)|(1 - z^2)^\alpha |z|^{2\beta+1} dz = 1.$$

To verify that (1.4) fails outside the region (6.1), we shall consider

$$(6.8) \quad \alpha < \beta,$$

$$(6.9) \quad \beta \leq \alpha < -\frac{1}{2},$$

$$(6.10) \quad \alpha + \beta = -1, \quad -\frac{1}{2} < \alpha < 0,$$

$$(6.11) \quad \alpha + \beta < -1, \quad -\frac{1}{2} < \alpha < 0, \quad \beta > -1,$$

$$(6.12) \quad \alpha = -\frac{1}{2}, \quad -1 < \beta < -\frac{1}{2}.$$

For (6.8) and (6.9), observe that if (1.2) and (1.4) hold, then setting

$$r_n = \max \{|T_n^{(\alpha, \beta)}(x)| : -1 \leq x \leq 1\}$$

it follows that  $r_n^2 \leq M r_n$ ; i.e.,

$$(6.13) \quad r_n \leq M.$$

But from (1.7),

$$r_{2n} = \max \{|R_n^{(\alpha, \beta)}(x)| : -1 \leq x \leq 1\}$$

and so, as in [5],

$$r_{2n} \geq |R_n^{(\alpha, \beta)}(-1)| = \frac{(\beta + 1)_n}{(\alpha + 1)_n} \sim n^{\beta - \alpha}$$

and

$$r_{2n} \cong Mn^{-\alpha-1/2}, \quad \beta \leq \alpha < -\frac{1}{2}$$

which contradicts (6.13).

In the case (6.10), suppose that (1.4) held. Then, since the integral representation (1.3) holds in this case if  $a \neq b$ , we would have  $V_1 \leq M$  for  $a \neq b$  and so, by Fatou's lemma, for  $a = b$  as well. But  $B = c(2a)^{-1}$  when  $a = b$  and so by (6.2) and (3.5)

$$(6.14) \quad V_1 = A \int_0^{2a} c^{-1} \left( \frac{c}{2a} \pm 1 \right) \left( 1 - \frac{c^2}{4a^2} \right)^{\alpha-1/2} dc, \quad a = b.$$

But the integral in (6.14) obviously diverges, which contradicts  $V_1 \leq M$ . Therefore (1.4) fails.

The argument is similar in the case (6.11). From (3.1) we have, much as in [5],

$$(6.15) \quad \begin{aligned} V_1 = & A(\Gamma(\alpha + \frac{1}{2}))^{-1} a^{-2\alpha} b^{\alpha-\beta-1} \int_{|a-b|}^{a+b} c^{\alpha+\beta} (1-B^2)^{\alpha-1/2} \\ & \cdot \{F(\alpha-\beta, \alpha+\beta; \alpha+\frac{1}{2}, \frac{1}{2}(1-B)) \\ & \pm F(\alpha-\beta-1, \alpha+\beta+1; \alpha+\frac{1}{2}, \frac{1}{2}(1-B))\} dc. \end{aligned}$$

Also ([3, 2.8, (50)]),

$$F(2\lambda, 2\gamma; \lambda + \gamma + \frac{1}{2}; \frac{1}{2}) = \frac{\Gamma(\lambda + \gamma + \frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(\lambda + \frac{1}{2})\Gamma(\gamma + \frac{1}{2})}$$

if  $\lambda + \gamma + \frac{1}{2}$  is not a negative integer or zero, so the sum of  $F$ 's in (6.15) is bounded away from zero as  $c \rightarrow 0$ . Hence the integral in (6.15) diverges when  $a = b$ . Since the integral representation holds for (6.11) if  $a \neq b$ , Fatou's lemma then again shows that (1.4) fails.

Finally, failure of (1.4) in the case (6.12) follows from (5.4), since if  $\beta < -\frac{1}{2}$ ,

$$\left( \frac{b-a}{b} \right)^{\beta+1/2} \rightarrow \infty \quad \text{as } b \rightarrow a+.$$

**7. Applications.** In this section, we show that Theorem 1 gives a convolution structure for the generalized Chebyshev polynomials. This convolution structure allows the extension to generalized Chebyshev expansions parts of Fourier analysis which cannot be extended to orthogonal polynomial expansions in general. As a related application we also prove the positivity of the generalized translation operator. For other applications, see [4], [5].

For  $\alpha, \beta > -1$ , let  $L_1^{(\alpha, \beta)}$  denote the class of measurable functions  $f(x)$  on  $(-1, 1)$  for which the norm

$$\|f\|_1 = \int_{-1}^1 |f(x)|(1-x^2)^\alpha |x|^{2\beta+1} dx$$

is finite. The transform  $\hat{f}$  of a function in  $L_1^{(\alpha, \beta)}$  is defined by

$$\hat{f}(n) = \int_{-1}^1 f(x) T_n^{(\alpha, \beta)}(x) (1-x^2)^\alpha |x|^{2\beta+1} dx.$$

Then  $f$  has the expansion

$$f(x) \sim \sum_{n=0}^{\infty} t_n^{(\alpha, \beta)} \hat{f}(n) T_n^{(\alpha, \beta)}(x),$$



where

$$t_n^{(\alpha, \beta)} = \left( \int_{-1}^1 [T_n^{(\alpha, \beta)}(x)]^2 (1-x^2)^\alpha |x|^{2\beta+1} dx \right)^{-1}$$

$$\cdot \begin{cases} \frac{(2k + \alpha + \beta + 1)\Gamma(k + \alpha + \beta + 1)\Gamma(k + \alpha + 1)}{\Gamma(k + \beta + 1)\Gamma(k + 1)\Gamma(\alpha + 1)\Gamma(\alpha + 1)} & \text{if } n = 2k, \\ \frac{(2k + \alpha + \beta + 2)\Gamma(k + \alpha + \beta + 2)\Gamma(k + \alpha + 1)}{\Gamma(k + \beta + 2)\Gamma(k + 1)\Gamma(\alpha + 1)\Gamma(\alpha + 1)} & \text{if } n = 2k + 1. \end{cases}$$

If  $\alpha \geq \beta, \alpha + \beta > -1$ , we define the convolution  $f * g$  of two functions  $f, g \in L_1^{(\alpha, \beta)}$  by

$$(7.1) \quad (f * g)(x) = \int_{-1}^1 \int_{-1}^1 f(y)g(z)(1-y^2)^\alpha |y|^{2\beta+1} d\mu_{x,y}(z) dy.$$

Also, let  $\|f\|_\infty$  be the sup norm. Then, as in [6] in the Gegenbauer case  $\beta = -\frac{1}{2}$ , we have the following corollary of Theorem 1(iii) and (iv):

**COROLLARY 3.** *Let  $\alpha \geq \beta, \alpha + \beta > -1$  and  $f, g, h \in L_1^{(\alpha, \beta)}$ . Then  $f * g \in L_1^{(\alpha, \beta)}$  and*

- (i)  $\|f * g\|_1 \leq M \|f\|_1 \|g\|_1$ ;
- (ii)  $\|f * g\|_\infty < M \|f\|_\infty \|g\|_1$ ;
- (iii)  $f * g = g * f$ ;
- (iv)  $(f * g) * h = f * (g * h)$ ;
- (v)  $(f * g)^\wedge(n) = \hat{f}(n) \hat{g}(n), \quad n = 0, 1, 2, \dots$ ;

with  $M = 1$  if  $\alpha \geq \beta > -\frac{1}{2}$ . Moreover,  $L_1$  is a commutative semisimple regular Banach algebra (with the norm  $\|f\| = M \|f\|_1$ ) whose maximal ideal space is isomorphic to the space  $\{0, 1, 2, \dots\}$  endowed with its discrete topology.

If  $\alpha \geq \beta > -\frac{1}{2}$  and  $f \in L_1^{(\alpha, \beta)}$ , then following [4], we define the generalized translate  $f(x, y)$  of  $f(x)$  by

$$f(x, y) = \int_{-1}^1 f(z)k(x, y, z; \alpha, \beta)(1-z^2)^\alpha |z|^{2\beta-1} dz, \quad -1 < x, y < 1.$$

(Note that for  $-1 < y < 1, f(\cdot, y) \in L_1^{(\alpha, \beta)}$ .) Then by Theorem 1(iv) we immediately obtain

**COROLLARY 4.** *Let  $\alpha \geq \beta > -\frac{1}{2}$ . Then the operator which takes  $f \in L_1^{(\alpha, \beta)}$  into  $f(x, y)$  is a positive operator in the sense that if  $f(x) \geq 0, -1 < x < 1$ , then  $f(x, y) \geq 0, -1 < x, y < 1$ .*

Note that if  $\alpha \geq \beta > -\frac{1}{2}$ , (7.1) takes the form

$$(f * g)(x) = \int_{-1}^1 f(x, y)g(y)(1-y^2)^\alpha |y|^{2\beta+1} dy.$$

**Acknowledgment.** This paper is part of the author’s doctoral dissertation written at the Northwestern University under the guidance of Professor George Gasper. The author wishes to express his gratitude for the patient help and concern of his advisor.

REFERENCES

[1] R. ASKEY, *Orthogonal polynomials and positivity*, Studies in Applied Mathematics 6, Wave Propagation and Special Functions, D. Ludwig and F. W. J. Olver, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1970, pp. 64-85.

- [2] R. ASKEY AND J. FITCH, *Integral representations for Jacobi polynomials and some applications*, J. Math. Anal. Appl., 26 (1969), pp. 82–86.
- [3] A. ERDÉLYI, *Higher Transcendental Functions*, vol. I, McGraw-Hill, New York, 1953.
- [4] G. GASPER, *Positivity and the convolution structure for Jacobi series*, Ann. of Math., 93 (1971), pp. 112–18.
- [5] ———, *Banach algebras for Jacobi series and positivity of a kernel*, Ibid., 95 (1972), pp. 261–80.
- [6] I. I. HIRSCHMAN, JR., *Harmonic analysis and ultraspherical polynomials*, Symposium of the Conference on Harmonic Analysis, Cornell, 1956.
- [7] G. SZEGŐ, *Orthogonal Polynomials*, Colloquium Publications, vol. 23, 4th ed., American Mathematical Society, Providence, RI, 1975.

## DIMENSIONALITY REDUCTION METHODS FOR EFFICIENT NUMERICAL SOLUTION, BACKWARD IN TIME, OF PARABOLIC EQUATIONS WITH VARIABLE COEFFICIENTS\*

PAOLO MANSELLI† AND KEITH MILLER‡

**Abstract.** We review several general purpose numerical methods for the ill-posed problem of solving a parabolic equation backward in time. Most of those methods are applicable only to the case of constant coefficients or else suffer from greatly excessive computational and storage requirements. For the general problem with variable coefficients we instead propose certain modifications of known least squares methods and eigenfunction expansion methods. Numerical trials show, as expected, a dramatic reduction in the number of elements required in our “approximate basis” for the space of initial functions.

**1. Some previous methods.** We wish to develop efficient methods for the problem of approximately determining the solution  $u(x, t)$  of a linear parabolic equation when data  $g$  for  $u$  is given not at the initial time  $t = 0$ , but at a later time  $t = T > 0$ . That is, let  $u$  be an exact solution of

$$(1) \quad \begin{aligned} (a) \quad & u_t = \sum_{ij} a_{ij} u_{x_i x_j} + \sum_j b_j u_{x_j} + cu \quad \text{in } \Omega \times [0, \infty), \\ (b) \quad & u = 0 \quad \text{on } \partial\Omega \times [0, \infty), \\ (c) \quad & u(x, T) \approx g(x), \quad \text{a given data function.} \end{aligned}$$

We assume here that  $\Omega$  is a bounded domain in  $R^n$  with sufficiently smooth boundary and that the coefficients  $a_{ij}(x, t)$ ,  $b_i(x, t)$ ,  $c(x, t)$  are uniformly parabolic and fairly smooth.

This ill-posed problem can be stabilized for times  $t > 0$  if it is known that the initial function  $u(x, 0)$  satisfies a prescribed bound. Writing (1a) as an ordinary differential equation on the Hilbert space  $L^2(\Omega)$ , and the data accuracy and the prescribed bound in terms of the  $L^2$  norm, one has

$$(2) \quad \begin{aligned} u' &= -L(t)u, \quad \text{for } t > 0, \\ \|u(T) - g\| &\leq \varepsilon, \\ \|u(0)\| &\leq E, \end{aligned}$$

with  $g$  in  $L^2(\Omega)$ , and  $\varepsilon, E$  given. Let  $A$  denote the evolution operator which maps the unknown initial value  $f = u(0)$  into the final value  $u(T) \equiv Af$ . The conditions (2) can then be written:

$$(3) \quad \|Af - g\| \leq \varepsilon, \quad \|Bf - 0\| \leq E,$$

where  $B$  will denote the identity operator except where otherwise stated. It can be shown now (for example by log convexity type arguments, see [1] and [9]) that the problem of determining  $u(t)$  among all solutions satisfying the constraints (3) is stable.

\* Received by the editors January 13, 1977, and in revised form September 18, 1978.

† Department of Mathematics, University of Florence, Italy. The work of this author was supported by a C.N.R. grant at the University of California, Berkeley.

‡ Department of Mathematics, University of California, Berkeley, California 94720. The work of this author was supported by the National Science Foundation under Grants MPS 73-08593 and MCS 76-06967 and in part under INT 76-04031.

That is, as  $E$  is fixed and  $\varepsilon$  gets small, then the difference at time  $t > 0$  between any two solutions  $u_1$  and  $u_2$  satisfying (3) is also guaranteed to be small (in the  $L^2$  norm, the uniform norm, or any other decent norm; the usual type bound is of the Hölder form  $O(\varepsilon^{\lambda(t)} E^{1-\lambda(t)})$  with  $0 < \lambda(t) < 1$ ).

In practice the parabolic equation (1) will usually be replaced by a finite difference or finite element approximation on a discretization  $\Omega_h$  of  $\Omega$  and  $\bar{A}$  will then denote the matrix mapping the discrete initial function  $\xi = u(0)$  into the discrete final solution  $\bar{A}\xi = u(T)$ .

This problem is susceptible to application of certain general purpose numerical methods for ill-posed problems devised by the second author and others. The problem is that most of these methods, without modification, lead to computations of huge dimensionality. For example, if  $\Omega$  is the square in two dimensions, and  $\Omega_h$  is a  $60 \times 60$  discretization of  $\Omega$ , then  $\bar{A}$  will be a  $3,600 \times 3,600$  nonsparse matrix.

Let us mention very briefly some of these methods and their difficulties; for a more extensive discussion see the symposium notes [11]. These notes also announce portions of the present joint work.

*Partial eigenfunction expansion.* (See Miller [6].) Let  $\phi_1, \phi_2, \dots$ , be a complete system of “orthonormal eigenfunctions” which are simultaneously orthogonal with respect to both  $A$  and  $B$  (here  $B$  is not necessarily the identity operator); that is

$$(4) \quad \begin{aligned} (A\phi_i, A\phi_j) &= (A_j)^2 \delta_{ij}, \\ (B\phi_i, B\phi_j) &= (B_j)^2 \delta_{ij}. \end{aligned}$$

For example, if  $B$  is the identity, then the weights  $(B_j)^2$  will all be 1's and the functions  $\phi_j$  and the weights  $(A_j)^2$  will be the orthonormal eigenfunctions and corresponding eigenvalues of the compact self-adjoint operator  $A^T A$ . If

$$(5) \quad f = \sum_1^\infty f_j \phi_j, \quad g = \sum_1^\infty g_j A \phi_j,$$

then (3) can be written

$$(6) \quad \begin{aligned} \|Af - g\|^2 &= \left\| \sum_1^\infty (f_j - g_j) A \phi_j \right\|^2 = \sum_1^\infty |(f_j - g_j) A_j|^2 \leq \varepsilon^2, \\ \|Bf - 0\|^2 &= \left\| \sum_1^\infty (f_j - 0) B \phi_j \right\|^2 = \sum_1^\infty |(f_j - 0) B_j|^2 \leq E^2. \end{aligned}$$

Assume now that the eigenfunctions  $\phi_1, \phi_2, \dots$  have been so ordered that the ratios  $A_j/B_j$  are nonincreasing with respect to  $j$ . Then truncate our expansion of  $g$  at exactly that order  $\alpha$  just previous to where  $A_j/\varepsilon$  becomes  $< B_j/E$ . Let  $\xi^\alpha = \sum_1^\alpha g_j \phi_j$  denote that initial function obtained by this  $\alpha$ th order eigenfunction expansion of the data function  $g$ . Then it can be shown that:

$$(7) \quad \|A\xi^\alpha - g\| \leq 2\varepsilon, \quad \|B\xi^\alpha - 0\| \leq 2E,$$

and hence  $\xi^\alpha$  is a “nearly best possible” approximation to  $f$ , in the sense that  $\xi^\alpha$  satisfies nearly the same “fit to the data” and “prescribed bound” as does  $f$  itself.

This is a pretty specialized method, because, except in certain special cases, we just aren't usually given the eigenfunctions of  $A^T A$ , and they can be very difficult to obtain. However, when they're available, it is a *very* good method and should be used.

This method is computationally equivalent to the singular value decomposition method introduced by Golub and Kahan [5] for the solution of ill conditioned matrix equations  $Ax = b$ .

*Least squares.* These and similar methods seem to have been discovered independently by several authors; we mention Morozov [13], Backus [2], Miller [7], and also Tihonov [14] and Bellman [3]; see [11] for a fuller discussion. See also the paper by Miller and Viano [10] for an exposition of both expansion and least square methods.

Notice that the unknown initial function  $f$  from (3) satisfies:

$$(8) \quad \|Af - g\|^2 + (\epsilon/E)^2 \|Bf\|^2 \leq 2\epsilon^2.$$

Thus, let our approximation  $\xi$  be such that

$$(9) \quad \|A\xi - g\|^2 + (\epsilon/E)^2 \|B\xi\|^2$$

is minimized, i.e., the solution of the least squares equation

$$(10) \quad (A^T A + (\epsilon/E)^2 B^T B)\xi = A^T g.$$

Since  $\xi$  will also satisfy the claimed fit to the data and prescribed bounds (3) (except for a factor of at most  $\sqrt{2}$ ) this is also a nearly best possible method. Moreover, one can compute exactly the best possible error bound for any linear functional of the solution. The problem is that  $A$ , and hence  $A^T A$  in (10) is horribly *nonsparse*. In this form, therefore, the method seems totally impractical for multidimensional parabolic problems.

*Stabilized quasi-reversibility.* See Miller [8]. Suppose  $L$  in (2) is self adjoint,  $\geq 0$  and constant with respect to  $t$ . This method involves perturbing the equation (2) a bit, replacing  $L$  in (2) by  $F(L)$ , where  $F(\lambda)$  is  $\approx \lambda$  for small  $\lambda$ , but is bounded above for large  $\lambda$ . One then solves the perturbed equation backward, to get an approximation  $v(t)$  to  $u(t)$ :

$$(11) \quad \begin{aligned} v' &= -F(L)v, & t \leq T, \\ v(T) &= g. \end{aligned}$$

Then, if desired, solve the unperturbed equation forward with the initial value  $\xi = v(0)$  so obtained to yield a solution  $w(t)$ . In this way we get some very efficient methods which yield the best possible error bound  $\|u(t) - v(t)\| \leq \epsilon^{t/T} E^{1-t/T}$ . The advantage of this method is that  $F(L)$  can be taken to be a rational function that factors into its linear (complex) or quadratic (real) factors above and below; hence each factor had a sparseness pattern only little worse than that of  $L$  itself. The shortcoming of this method is that it doesn't extend well to very general  $L$ , and definitely not well to  $L(t)$ .

*The backward beam equation approach.* See Buzbee-Carasso [4]. Once again, let  $L$  be self-adjoint, constant with respect to  $t$ , and let  $T = 1$ . Then  $y(t) = e^{\alpha t} u(t)$ , with  $\alpha = \log(E/\epsilon)$ , satisfies:

$$(12) \quad \begin{aligned} (a) \quad & y'' = (L - \alpha)^2 y, \\ (b) \quad & \|y(1) - e^\alpha g\| \leq \epsilon e^\alpha, \\ (c) \quad & \|y(0) - 0\| \leq E. \end{aligned}$$

One then lets our approximation be  $v(t) = e^{-\alpha t} w(t)$ , where  $w(t)$  is the solution of the two-point boundary value problem for (12) with  $w(1) = e^\alpha g$  and  $w(0) = 0$ . Because the norm of any solution of (12) must be convex with respect to  $t$ , one gets the best possible error bound  $\epsilon^t E^{1-t}$  once again. The shortcoming here is that we have to simultaneously solve for all time levels at once; it thus introduces one higher dimension to the storage and computational difficulties. It does seem, however, that the method extends readily to variable  $L(t)$  (not with best possible stability) and perhaps even to nonlinear equations.

We would now like to propose the previously mentioned least squares methods or eigenfunction expansion methods, but with some modifications, for the general problem with variable coefficients.

**2. Reduced dimensionality for  $\xi$ .** In a typical problem with smooth coefficients, most highly oscillatory initial functions  $\xi$  will damp out drastically by  $t = T$ , so the space of  $\xi$  we need to deal with should be quite small.

Let  $\phi_1, \dots, \phi_N$  be an “approximate basis” for our space  $L^2(\Omega)$  of initial functions  $\xi$ ; let  $P_N$  denote the orthogonal projection onto their linear span; let  $Q_N = I - P_N$  denote the projection onto their orthogonal complement, and suppose that

$$(13) \quad \|AQ_N\| \leq .1(\varepsilon/E).$$

Instead of (3), we have

$$(14) \quad \begin{aligned} \|AP_N f - g\| &\leq \|A(P_N - I)f\| + \|Af - g\| \leq 1.1\varepsilon, \\ \|B(P_N f)\| &\leq \|Bf\| \leq E; \end{aligned}$$

i.e., the projection  $P_N f$  satisfies nearly the same constraints as  $f$  itself; its high order part  $Q_N f$  just hardly enters into the “fit to the data” of  $f$ . Therefore, we can do the least squares approach of (8)–(10), but with  $\xi$  a linear combination of the  $\phi_1, \dots, \phi_N$  only, and with  $\varepsilon$  replaced everywhere by  $1.1\varepsilon$ . The matrix in (10) is then only  $N \times N$  and involves only computing the solutions  $A\phi_1, \dots, A\phi_N$  and their inner products.

Alternatively, if  $N$  is not too large, one can apply the eigenfunction expansion methods to the operator  $AP_N$ . This involves computing the eigenvalue and eigenvectors of the matrix  $b_{ij} = (A\phi_i, A\phi_j)$ .

Notice that once we’ve guessed at a good “approximate basis” it is possible to check computationally whether  $\|AQ_N\|$  is sufficiently small, since  $\|AQ\|^2$  is the spectral radius of  $QA^T A Q$ , which can be computed by the power method. This involves computing high powers  $(QA^T A Q)^n \phi$ , where  $\phi$  is any initial function (say  $\phi_{N+1}$ ) which has a nonzero component of the dominant eigenfunction of  $QA^T A Q$ . Recall that  $A^T$  is itself an evolution operator; it carries the initial value  $u(0) = \xi$  into the final value  $u(T) = A^T \xi$  for the parabolic equation:  $u' = -(L(T-t))^T u$ ,  $0 \leq t \leq T$ .

**3. Conjecture and counterexample.** Let’s consider several possible choices of the approximate basis. Suppose, for example that our equation is  $u_t = (a(x, t)u_x)_x$  on the one-dimensional interval  $[0, \pi]$ , with  $u(0, t) = u(\pi, t) = 0$  and with  $a(x, t)$  uniformly elliptic (i.e. uniformly bounded above and below by positive constants).

One good choice for  $\phi_1, \dots, \phi_N$  might then be the Fourier basis  $\sqrt{2} \sin(1x), \sqrt{2} \sin(2x), \dots, \sqrt{2} \sin(Nx)$ . If the coefficients  $a(x, t)$  (when  $a$  and  $u$  are extended periodically across the end points by symmetric and antisymmetric reflection) are sufficiently smooth (say  $C^q$ ), then the kernel function  $k$  for the evolution operator  $A$ ,

$$(15) \quad A\phi(x) = \int_0^\pi k(x, y)\phi(y) dy,$$

is known to be a smooth  $C^{q+2}$  function of  $y$ , whose  $C^{q+2}$  norm is bounded in terms of the  $C^q$  norm of the coefficient function; hence integration by parts  $q + 2$  times in (15) yields that:

$$(16) \quad \|A\phi_{N+1}\| = O(N^{-q-2}).$$

The same result would hold for the Fourier basis  $\{2 \sin(nx) \sin(my)\}$  in the case of the square  $[0, \pi] \times [0, \pi]$  in two space dimensions.

A better choice might be to let  $\phi_1, \dots, \phi_N; \phi_{N+1}, \dots$  be the eigenfunctions of  $L(0)$  (if these are easily available), since  $u' = L(t)u \approx L(0)u$  for small  $t$ . If  $u(0)$  is a high order eigenfunction  $\phi_{N+1}$  of  $L(0)$  then the solution ought to die out exponentially so fast (initially at an  $e^{-CN_2t}$  rate) that it becomes very small before  $L(t)$  can change much from  $L(0)$ . (We wish to assume here that the coefficients are not necessarily  $C^\infty$  but merely  $C^2$  or so.) The authors originally had the following *loose conjecture*: that  $\|A\phi_{N+1}\| = O(e^{-CN})$  at least, instead of merely  $O(N^{-q-2})$ . We had a bit of computational experience with a test program and our intuition appeared to be justified; the solution, starting out in a high order eigencomponent (with respect to  $L(0) = \partial^2/\partial x^2$ ) did not seem to diffuse very quickly into the low order components.

However, our intuition let us down, and analysis of the results for some anomalous computer runs led us to the following *counterexample*.

Let  $u$  be the solution of the problem

$$(17) \quad u_t = (a(x, t)u_x)_x \quad \text{on } 0 \leq x \leq \pi, \quad 0 \leq t \leq 1,$$

$$(18) \quad u(0, t) = u(\pi, t) = 0,$$

$$(19) \quad u(x, 0) = \sin((N+1)x) = \phi_{N+1},$$

with:

$$(20) \quad a(x, t) = 1 - KtN^{-q} \cos(Nx)$$

where  $N \geq 1$  and  $q \geq 1$  are integers and  $0 < K < 1$  is a sufficiently small positive constant. It will be shown that

$$\|A\phi_{N+1}\| \equiv \|u(\cdot, 1)\| \geq O(N^{-q-3}).$$

Notice that the coefficient  $a(x, t)$  has uniformly bounded derivatives of order  $q$ , independently of  $N$ , and that the coefficient stays uniformly elliptic.

We treat the nonconstant part of the coefficient as a small perturbation, transfer it to the right-hand side, and apply the method of successive approximations with the initial approximant  $u^{(0)}$  being the unperturbed solution  $e^{-(N+1)^2t} \sin((N+1)x)$ , and with  $u^{(\nu+1)}$  solving

$$u_t^{(\nu+1)} - u_{xx}^{(\nu+1)} = -\left(\frac{Kt}{N^q} \cos(Nx)u_x^{(\nu)}\right)_x$$

on  $0 \leq x \leq \pi, 0 \leq t$  with boundary conditions (18), (19). Notice that the 1st approximant  $u^{(1)}$  feeds a substantial increment of itself *from* the quickly decaying  $(N+1)$ st Fourier component *into* the slowly decaying 1st Fourier component. We have  $u^{(1)}$  satisfying (18), (19) and the equation

$$u_t^{(1)} - u_{xx}^{(1)} = f_N(t)(\sin(1 \cdot x) + (2N+1) \sin((2N+1)x)),$$

where  $f_N(t) = Kt(N+1) e^{-(N+1)^2t}/(2N^q)$ . The solution is

$$u^{(1)}(x, t) = e^{-(N+1)^2t} \sin((N+1)x) + \int_0^t f_N(\tau)[e^{-(t-\tau)} \sin x + (2N+1) e^{-(2N+1)^2(t-\tau)} \sin((2N+1)x)] d\tau,$$

which has norm  $\|u^{(1)}(1)\| \geq (\pi/2) \int_0^1 f_N(\tau) e^{-1+\tau} d\tau = KN^{-q-3}(\pi/2) e^{-1}(1 - e^{-N^2} - N^2 e^{-N^2}) \geq O(N^{-q-3})$ .

We now show rigorously that the norm of the true solution  $u(1)$  is even larger than the norm of its 1st order approximant  $u^{(1)}(1)$ . For this purpose we transfer to an

argument in terms of the Fourier coefficients  $\{u_j(t)\}$  of  $u(t)$ . Let  $X$  be the Banach space of functions whose Fourier coefficients satisfy:

$$\|v\|_X = \sum_1^\infty \sup_{[0, 1]} |v_j(t)| < \infty.$$

Notice that the solution  $u$  of (17)–(20) is  $C^\infty$  so certainly belongs to this space. One sees easily that the Fourier coefficients of this solution  $u$  must satisfy the integral equation

$$(21) \quad u_j(t) = e^{-j^2 t} \delta_j^{N+1} + \left\{ \frac{Kj}{2N^q} \int_0^t \tau e^{-(t-\tau)j^2} [(j+N)u_{j+N}(\tau) + |j-N|u_{|j-N|}(\tau)] d\tau \right\}.$$

Moreover one can show that the integral operator  $\kappa$  corresponding to  $\{\cdot\}$  on the right-hand side of (21) is norm bounded on  $X$  by  $2K/N^{q-1}$ . Therefore if  $K < \frac{1}{2}$ , then this is a contraction mapping and hence the Fourier coefficients  $u_j^{(\nu)}(t)$  of the method of successive linear approximation converge to the Fourier coefficients  $u_j(t)$  of the unique solution  $u(x, t)$ . On the other hand, the operator  $\kappa$  is “positive”, i.e. it maps positive coefficients into positive coefficients. Thus coefficients of approximants are increasing with  $\nu$ , i.e.:

$$0 \leq u_j^{(0)}(t) \leq u_j^{(1)}(t) \leq \dots \leq u_j^{(\nu)}(t) \leq u_j^{(\nu+1)}(t) \leq \dots \leq u_j(t).$$

Hence  $u_j$  has even larger (positive) Fourier coefficients than its first approximant  $u_j^{(1)}(t)$ . Thus as claimed,

$$\|A\phi^{N+1}\| = \|u(1)\| \geq \|u^{(1)}(1)\| \geq K\pi N^{-q-3} (e^{-1} - e^{-1-N^2} - N^2 e^{-1-N^2})/2 \geq O(N^{-q-3}).$$

This example certainly fails the  $O(e^{-CN})$  behavior conjectured earlier. Incidentally, notice that this example (with  $C^q$  bounded coefficients) nearly attains the  $O(N^{-q-2})$  upper bound proved in (15), (16).

**4. Refinement of the approximate basis.** The best choice of our orthonormal basis, to make  $\|AQ_N\|$  as small as possible for each given dimension  $N$ , is the eigenfunctions of  $A^T A$  of course. This follows from the Courant minimax principle. With this in mind we could start with an initial set of functions  $\psi_1, \dots, \psi_N$  then do a *refinement* of them, rotating span  $(\psi_1, \dots, \psi_N)$  approximately into span (first  $N$  eigenfunctions of  $A^T A$ ) by the *block power method*, applying powers of the parabolic evolution operator  $A^T A$ . Because  $A^T A$  can be expected to have quickly decaying eigenvalues, and because of the fast rate of convergence of the block power method in this case, we would hope to see a big improvement in  $\|AQ_N\|$  with only a few iterates of the refinement process.

We now describe a computer test program based on the parabolic equation (17), (18) (but on the interval  $[0, 1]$  rather than  $[0, \pi]$ ). Let  $A$  be the evolution operator which carries the initial function  $u(\cdot, 0)$  at time 0 into the final function  $u(\cdot, T)$  at time  $T$ . We approximate  $A$  by a finite-difference equation discretization  $\bar{A}$ , with  $NX$  equal subdivisions of the space interval  $[0, 1]$  and  $NT$  equal subdivisions of the time interval  $[0, T]$ . For the space discretization of  $L(t) = \partial/\partial x(a(x, t)\partial/\partial x)$  we use the usual centered second difference operator (with the coefficient function of course evaluated at the subinterval center points). This leaves us with a system of ODE’s for the  $NX - 1$  dimensional discrete function  $\bar{u}(t)$ :

$$(22) \quad \bar{u}^1 = \bar{L}(t)\bar{u}, \quad 0 \leq t \leq T.$$

We discretize this system in time, using the second order diagonally implicit Runge–Kutta method of Miller [12]. Such stiffly stable methods are absolutely necessary in the present case in order to accurately damp out the high order components of



(22) as does the true ODE. Our approximant  $\bar{v}(t)$  to the solution  $\bar{u}(t)$  of (22) is given by  $\bar{v}(T) = \bar{A}\bar{v}(0) = \bar{A}_{NT} \cdots \bar{A}_{j+1} \cdots \bar{A}_2 \bar{A}_1 \bar{v}(0)$  where  $\bar{A}_{j+1}$  carries us from time  $t_j$  to time  $t_{j+1} = t_j + \Delta t$  by means of the two successive linear tridiagonal implicit equations:

$$(23) \quad \bar{v}_{j+1/3} = \bar{v}_j + (\Delta t/3) \bar{L}_{j+1/3} \bar{v}_{j+1/3},$$

$$(24) \quad \bar{v}_{j+1} = (\frac{9}{4} \bar{v}_{j+1/3} - \frac{5}{4} \bar{v}_j) + (\Delta t/4) \bar{L}_{j+1} \bar{v}_{j+1},$$

where, of course  $\bar{v}_j$  denotes  $\bar{v}(t_j)$ ,  $\bar{L}_{j+1/3}$  denotes  $\bar{L}(t_j + \Delta t/3)$ , etc.

The exact transpose  $(\bar{A})^T$  is then computable by a similar finite difference sequence whose solution  $\bar{w}(t)$  is given by  $\bar{w}(0) = (\bar{A})^T \bar{w}(T) = (\bar{A}_1)^T \cdots (\bar{A}_{j+1})^T \cdots (\bar{A}_{NT})^T \bar{w}(T)$ , where  $(\bar{A}_{j+1})^T$  carries us from time  $t_{j+1}$  to  $t_j$  by means of the two implicit equations

$$(25) \quad \bar{w}_{j+3/4} = \bar{w}_{j+1} + (\Delta t/4) (\bar{L}_{j+1})^T \bar{w}_{j+3/4},$$

$$(26) \quad \bar{w}_{j+5/12} = \bar{w}_{j+3/4} + (\Delta t/3) (\bar{L}_{j+1/3})^T \bar{w}_{j+5/12},$$

and

$$(27) \quad \bar{w}_j = \frac{9}{4} \bar{w}_{j+5/12} - \frac{5}{12} \bar{w}_{j+3/4}.$$

We begin with functions  $\psi_1, \dots, \psi_M$  (which in the beginning stage for us are always the discrete functions  $\sin(1\pi x), \dots, \sin(M\pi x)$ ), where  $M \leq NX$ . The 1st step is an orthonormalization of these (repeated twice to correct some serious roundoff difficulties in later stages when the  $\psi_j$  for larger  $j$  may all be nearly zero) to yield discrete functions  $\varphi_1, \dots, \varphi_M$ . The 2nd step is to compute  $\bar{A}\varphi_1, \dots, \bar{A}\varphi_M$  (the implicit equations in (23)–(24), and later in (25)–(27), can be solved in block form, of course), and print out their norms,  $\|\bar{A}\varphi_j\|$ . The third step (explanations for this step later) is to compute the  $M1 \times M1$  matrix  $b_{ij} = (\bar{A}\varphi_i, \bar{A}\varphi_j)$ ,  $i, j = 1, \dots, M1$  with  $M1 < M$  and compute its eigenvalues and eigenvectors with a standard computer center routine (EISPACK). Using the computed eigenvectors we easily rotate the basis elements  $\varphi_1, \dots, \varphi_{M1}$  within their own span to new elements  $\chi_1, \dots, \chi_{M1}$  which are the eigenvectors (corresponding to  $\lambda_1, \dots, \lambda_{M1}$ ) of  $P_{M1}(\bar{A})^T \bar{A} P_{M1}$ , where  $P_{M1}$  is the projection onto the span of these elements. This yields a new orthonormal basis  $\chi_1, \dots, \chi_{M1}, \varphi_{M1+1}, \dots, \varphi_M$  and their corresponding  $\bar{A}$  values  $\bar{A}\chi_1, \dots, \bar{A}\chi_{M1}, \bar{A}\varphi_{M1+1}, \dots, \bar{A}\varphi_M$ . The 4th step is to compute  $(\bar{A})^T \bar{A}\chi_1, \dots, (\bar{A})^T \bar{A}\varphi_M$ . These functions then become the initial functions for the next refinement stage. At the beginning, before the refinement stages, a check is performed to see if  $\Delta t$  is sufficiently small in our time discretization, by printing out  $\|\bar{A}\psi_i - \bar{A}\bar{\psi}_i\|$ ,  $i = 1, \dots, M$ , where  $\bar{A}$  represents the discretization of  $A$  with  $\Delta t$  replaced by  $\Delta t/2$ . In the last refinement stage, to save computer time, only the first and second step are performed.

The resulting sets of orthonormal functions  $\{\varphi_1, \dots, \varphi_M\}$  in the zeroth, 1st, 2nd,  $\dots$  refinements stages should converge toward the first  $M$  eigenfunctions  $\beta_1, \dots, \beta_M$  of  $(\bar{A})^T \bar{A}$ . The rate of convergence depends in a rather complicated way upon the ratios of the eigenvalues  $\lambda_j$  (these are  $\lambda_j = e^{-\pi^2 j^2 T}$  when  $a(x, t) \equiv 1$ , so large  $T$  should yield larger ratios). Since we expect these ratios to be large for large  $j$  and rather small for small  $j$ , we suspect that the block power method by itself might be rather slow to align  $\{\varphi_1, \dots, \varphi_{M1}\}$  one by one with  $\{\beta_1, \dots, \beta_{M1}\}$  even though their spans may be nearly parallel. Therefore, for the sake of a quicker alignment we have introduced the 3rd step.

Recall that  $P_N$  denotes the projection onto span  $\{\varphi_1, \dots, \varphi_N\}$ , with  $N < M$ , and  $Q_N$  the projection onto the orthogonal complement span  $\{\varphi_{N+1}, \dots, \varphi_M, \varphi_{M+1}, \dots, \varphi_{NX}\}$ .

Therefore:

$$(28) \quad \|\bar{A}Q_N\| \cong \max \{ \|\bar{A}\varphi_{N+1}\|, \dots, \|\bar{A}\varphi_{NX}\| \}.$$

Likewise:

$$(29) \quad \|\bar{A}Q_N\| \cong \|\bar{A}\varphi_{N+1}\| + \dots + \|\bar{A}\varphi_{NX}\|.$$

The sequence  $\|\bar{A}\varphi_1\|, \dots, \|\bar{A}\varphi_N\|, \dots, \|\bar{A}\varphi_N\|$  is usually rapidly decreasing, with the elements  $\|\bar{A}\varphi_{M+1}\|, \dots, \|\bar{A}\varphi_{NX}\|$  much smaller than  $\|\bar{A}\varphi_{N+1}\|$ ; we merely assume that the “unseen” values  $\|\bar{A}\varphi_{M+1}\|, \dots, \|\bar{A}\varphi_{NX}\|$ , would also be much smaller than  $\|\bar{A}\varphi_{N+1}\|$ . Therefore the upper bound in (29) is approximately the observable

$$(30) \quad \|\bar{A}Q_N\| \cong \|\bar{A}\varphi_{N+1}\| + \dots + \|\bar{A}\varphi_M\|.$$

(In fact, both the upper and lower bounds in (28) and (30) are usually approximately equal  $\|\bar{A}\varphi_{N+1}\|$ ; this is especially true in the later refinement stages and sometimes false in the zeroth refinement stage.)

With this in mind we examine several computer run results; the tables for Examples 1 through 5 display the norms of  $\bar{A}\varphi_j$  for our “approximate basis” elements  $\varphi_j$  after successive refinements of this basis.

TABLE 1

EXAMPLE 1.  $a(x, t) = 1 - .5(1 - t/T) \cos(15\pi x)$ .  $NX = 60$ ;  $M = 20$ ;  $M1 = 10$ ;  $T = .04$ ;  $NT = 90$ . Initial basis:  $\varphi_j = \sqrt{2} \sin(\pi jx)$ ,  $j = 1, \dots, 20$ .  $\|\bar{A}\varphi_j - \bar{A}\varphi_j\| \cong .69 \times 10^{-5}$ .

0-th refinement: $\ \bar{A}\varphi_j\ $ , $j = 1, \dots, 20$				
.69	.23	$.36 \times 10^{-1}$	$.46 \times 10^{-2}$	$.42 \times 10^{-2}$
$.29 \times 10^{-2}$	$.36 \times 10^{-2}$	$.26 \times 10^{-2}$	$.36 \times 10^{-2}$	$.29 \times 10^{-2}$
$.45 \times 10^{-2}$	$.59 \times 10^{-2}$	$.11 \times 10^{-1}$	$.10 \times 10^{-1}$	$.27 \times 10^{-3}$
$.90 \times 10^{-2}$	$.84 \times 10^{-2}$	$.38 \times 10^{-2}$	$.23 \times 10^{-2}$	$.13 \times 10^{-2}$
1-st refinement: $\ \bar{A}\varphi_j\ $ , $j = 1, \dots, 20$				
.69	.23	$.36 \times 10^{-1}$	$.27 \times 10^{-2}$	$.11 \times 10^{-3}$
$.24 \times 10^{-5}$	$.49 \times 10^{-7}$	$.39 \times 10^{-11}$	$.13 \times 10^{-11}$	$.50 \times 10^{-11}$
$.12 \times 10^{-11}$	$.80 \times 10^{-12}$	$.73 \times 10^{-12}$	$.11 \times 10^{-11}$	$.16 \times 10^{-12}$
$.81 \times 10^{-12}$	$.10 \times 10^{-12}$	$.25 \times 10^{-12}$	$.32 \times 10^{-13}$	$.55 \times 10^{-12}$

Consider Example 1: here the coefficients begin with  $a = 1 - .5 \cos(15\pi x)$  at  $t = 0$  and end with  $a = 1$  at  $t = T = .04$ . We use  $NX = 60$  subdivisions on  $x$ , which should be sufficient for our purposes since we consider only  $M = 20$  basis elements for our block power method. We first check that our  $\Delta t = T/NT = .04/90$  is sufficiently fine; we see that  $A(\Delta t) = \bar{A}$  and  $A(\Delta t/2) = \bar{A}$  differ at most by  $.69 \times 10^{-5}$  when applied to all our normalized initial basis functions  $\varphi_j = \sqrt{2} \sin(\pi jx)$ ,  $j = 1, \dots, 20$ , which is more than adequate accuracy for our purposes.

Now notice that even with 19 elements of the initial basis,  $\varphi_1, \dots, \varphi_{19} \sim \sqrt{2} \sin(\pi x), \dots, \sqrt{2} \sin(19\pi x)$ , we have  $\|\bar{A}Q_{19}\| \cong \|\bar{A}\varphi_{20}\| = 1.3 \times 10^{-3}$ . On the other hand, with only one refinement step we obtain a refined basis  $\varphi_1, \dots$  such that only 4 elements are needed for  $\approx 10^{-4}$  accuracy, since  $\|\bar{A}Q_4\| \cong \|\bar{A}\varphi_5\| + \dots + \|\bar{A}\varphi_{60}\| \approx \|\bar{A}\varphi_5\| = 1.1 \times 10^{-4}$ . Only 5 elements are needed for  $\approx 10^{-5}$  accuracy since  $\|\bar{A}Q_5\| \cong \|\bar{A}\varphi_6\| + \dots + \|\bar{A}\varphi_{60}\| \approx \|\bar{A}\varphi_6\| = 2.4 \times 10^{-6}$ .

Example 2 is exactly the same but with a larger  $\Delta t = T/NT = .04/20$ . We see that our truncation error  $A(\Delta t) - A(\Delta t/2)$  is a bit larger ( $\cong 1.4 \times 10^{-4}$  when applied to the initial basis), but that the results remain essentially unchanged. Here we have computed

TABLE 2

EXAMPLE 2.  $a(x, t) = 1 - .5(1 - t/T) \cos(15\pi x)$ ;  $NX, M, M1, T$  as in Example 1;  $NT = 20$ . Initial basis: as in Example 1.  $\|\bar{A}\varphi_j - \bar{A}\varphi_j\| \leq .14 \times 10^{-3}$ .

0-th refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$				
.6885971517	.2253621125		.3562653288	$\times 10^{-1}$
.4614599385	$\times 10^{-2}$	.4167318435	$\times 10^{-2}$	
.2894939091	$\times 10^{-2}$	.3606328407	$\times 10^{-2}$	.2620407039
.3601174424	$\times 10^{-2}$		.2918928903	$\times 10^{-2}$
.4491688360	$\times 10^{-2}$	.5908097559	$\times 10^{-2}$	.1112936136
.1009673234	$\times 10^{-1}$		.2708656351	$\times 10^{-3}$
.9081274800	$\times 10^{-2}$	.8468108222	$\times 10^{-2}$	.3784955149
.2335640104	$\times 10^{-2}$		.1319870876	$\times 10^{-2}$
1-st refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$				
.6888836077	.2257455333		.3540456421	$\times 10^{-1}$
.2677930783	$\times 10^{-2}$	.9887008367	$\times 10^{-4}$	
.1880674611	$\times 10^{-5}$	.2700461045	$\times 10^{-7}$	.5046380266
.3078001887	$\times 10^{-12}$		.1292009366	$\times 10^{-11}$
.3535879306	$\times 10^{-13}$	.7620555335	$\times 10^{-13}$	.1241809583
.5230060660	$\times 10^{-13}$		.8213544145	$\times 10^{-13}$
.6715851716	$\times 10^{-13}$	.6691155044	$\times 10^{-13}$	.4142766621
.5443072151	$\times 10^{-13}$		.5571022312	$\times 10^{-13}$
2-nd refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$				
.6888836078	.2257455330		.3540456408	$\times 10^{-1}$
.2677930773	$\times 10^{-2}$	.9887008300	$\times 10^{-4}$	
.1880674593	$\times 10^{-5}$	.2700461040	$\times 10^{-7}$	.3658578281
.2482395905	$\times 10^{-13}$		.2970692916	$\times 10^{-13}$
.2702875322	$\times 10^{-13}$	.2771698023	$\times 10^{-13}$	.2049124567
.2220004413	$\times 10^{-13}$		.2536235882	$\times 10^{-13}$
.1297460422	$\times 10^{-13}$	.1647823106	$\times 10^{-13}$	.1682785794
.1317570645	$\times 10^{-13}$		.1168381198	$\times 10^{-13}$

a second refinement step; notice that the corresponding  $\|\bar{A}\varphi_j\|$  differ almost not at all between the first and second refinements, for  $j \leq 7$ . This indicates that  $\varphi_1, \dots, \varphi_7$ , after only the first refinement step have probably been almost exactly rotated into the eigenlements of  $\bar{A}^T \bar{A}$ .

Consider Example 3. Here the coefficients begin with  $a = 1$  at  $t = 0$  and end with  $a = 1 - (10/11) \cos(7\pi x)$  at  $t = T = .02$ . Once again we use 60 subdivisions on  $x$ . Notice

TABLE 3

EXAMPLE 3.  $a(x, t) = 1 - (1 - (1/(1 + 10t/T))) \cos(7\pi x)$ ;  $NX, M, M1$  as in Example 1;  $T = .02$ ;  $NT = 20$ . Initial basis: as in Example 1.  $\|\bar{A}\varphi_j - \bar{A}\varphi_j\| \leq .42 \times 10^{-3}, j = 1, \dots, 20$ .

0-th refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$				
.89	.64	.41	.18	.10
.42	$\times 10^{-1}$	.24	$\times 10^{-1}$	.23
.73	$\times 10^{-2}$	.21	$\times 10^{-3}$	.11
.13	$\times 10^{-2}$	.20	$\times 10^{-3}$	.14
1-st refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$				
.89	.65	.44	.51	$\times 10^{-1}$
.55	$\times 10^{-2}$	.16	$\times 10^{-3}$	.18
.89	$\times 10^{-7}$	.30	$\times 10^{-9}$	.14
.13	$\times 10^{-9}$	.90	$\times 10^{-9}$	.24
	.11		.18	$\times 10^{-4}$
	.85		.14	$\times 10^{-8}$
	.58		.24	$\times 10^{-9}$
			.99	$\times 10^{-9}$
			.25	$\times 10^{-9}$

TABLE 4

EXAMPLE 4.  $a(x, t) = 1 - .2(1 - t/T)(\cos(8\pi x) + \cos(15\pi x)) - .2(1 - (1 + (1 + 10t/T))) \cos(3\pi x)$ ;  $NX, M, M1, T$  as in Example 1;  $NT = 20$ . Initial basis: as in Example 1.  $\|\bar{A}\varphi_j - \bar{A}\varphi_j\| \leq .16 \times 10^{-3}$ .

0-th refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$ .				
.69	.22	$.34 \times 10^{-1}$	$.17 \times 10^{-1}$	$.11 \times 10^{-1}$
$.12 \times 10^{-2}$	$.11 \times 10^{-1}$	$.30 \times 10^{-2}$	$.92 \times 10^{-2}$	$.56 \times 10^{-2}$
$.28 \times 10^{-2}$	$.25 \times 10^{-2}$	$.47 \times 10^{-2}$	$.38 \times 10^{-2}$	$.47 \times 10^{-3}$
$.37 \times 10^{-2}$	$.28 \times 10^{-2}$	$.11 \times 10^{-2}$	$.42 \times 10^{-3}$	$.73 \times 10^{-4}$
1-st refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$ .				
.69	.22	$.33 \times 10^{-1}$	$.29 \times 10^{-2}$	$.67 \times 10^{-4}$
$.11 \times 10^{-5}$	$.63 \times 10^{-8}$	$.18 \times 10^{-8}$	$.49 \times 10^{-9}$	$.13 \times 10^{-8}$
$.73 \times 10^{-12}$	$.87 \times 10^{-13}$	$.61 \times 10^{-12}$	$.82 \times 10^{-12}$	$.70 \times 10^{-12}$
$.72 \times 10^{-12}$	$.21 \times 10^{-12}$	$.20 \times 10^{-12}$	$.34 \times 10^{-12}$	$.27 \times 10^{-12}$

that our initial basis  $\varphi_j = \sqrt{2} \sin(\pi j x)$  is the basis of eigenelements for the initial operator  $L(0) = \partial/\partial x(1 \cdot \partial/\partial x)$ . Nevertheless, even with 16 elements of the initial basis we have  $\|\bar{A}Q_{16}\| \cong \|\bar{A}\varphi_{17}\| = 1.1 \times 10^{-3}$ ; and even with 19 elements of the initial basis we have  $\|\bar{A}Q_{19}\| \cong \|\bar{A}\varphi_{20}\| = 1.4 \times 10^{-4}$ . Again, after only one refinement step we obtain a great improvement; only 6 elements of the refined basis are needed for  $\approx 10^{-3}$  accuracy since  $\|\bar{A}Q_6\| \leq \|\bar{A}\varphi_7\| + \dots + \|\bar{A}\varphi_{60}\| \approx \|\bar{A}\varphi_7\| = 1.1 \times 10^{-3}$ . Only 7 elements are needed for  $\approx 10^{-4}$  accuracy since  $\|\bar{A}Q_7\| \leq \|\bar{A}\varphi_8\| + \dots + \|\bar{A}\varphi_{60}\| \approx \|\bar{A}\varphi_8\| = 1.6 \times 10^{-4}$ .

Example 4 has a more complicated coefficient structure, but also shows a marked improvement with only one refinement step. The coefficients begin with  $a = 1 - .2 \cdot (\cos(15\pi x) - \cos(8\pi x))$  at  $t = 0$  and end with  $a = 1 - (2/11) \cos(3\pi x)$  at  $t = T = .04$ . Notice that at least 19 basis elements are needed for  $10^{-4}$  accuracy with the initial basis since  $\|\bar{A}Q_{18}\| \cong \|\bar{A}\varphi_{19}\| \cong 4.2 \times 10^{-4}$ ; however only 4 elements are needed for  $10^{-4}$  accuracy with the refined basis since  $\|\bar{A}Q_4\| \leq \|\bar{A}\varphi_5\| + \dots + \|\bar{A}\varphi_{10}\| \approx \|\bar{A}\varphi_5\| = .67 \times 10^{-4}$ .

Example 5 shows a shorter final time  $T = .01$ . Hence it is to be expected that the eigenvalues  $\lambda_j$  of  $A^T A$  will decrease less rapidly with  $j$  than in the previous cases.

TABLE 5

EXAMPLE 5.  $a(x, t) = 1 - (1 - (1/(1 + 10t/T))) \cos(12\pi x)$ ;  $NX, M, M1$  as in Example 1;  $T = .01$ ;  $NT = 20$ . Initial basis: as in Example 1.  $\|\bar{A}\varphi_j - \bar{A}\varphi_j\| \leq .42 \times 10^{-3}$ .

0-th refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$ .				
.94	.79	.60	.41	.27
.23	.12	$.84 \times 10^{-1}$	$.60 \times 10^{-1}$	$.39 \times 10^{-1}$
$.18 \times 10^{-1}$	$.21 \times 10^{-2}$	$.13 \times 10^{-1}$	$.19 \times 10^{-1}$	$.19 \times 10^{-1}$
$.16 \times 10^{-1}$	$.12 \times 10^{-1}$	$.71 \times 10^{-2}$	$.58 \times 10^{-2}$	$.44 \times 10^{-2}$
1-st refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$ .				
.94	.80	.60	.42	.29
.21	$.14 \times 10^{-1}$	$.68 \times 10^{-2}$	$.26 \times 10^{-2}$	$.82 \times 10^{-3}$
$.23 \times 10^{-3}$	$.56 \times 10^{-4}$	$.12 \times 10^{-4}$	$.23 \times 10^{-5}$	$.39 \times 10^{-6}$
$.63 \times 10^{-7}$	$.90 \times 10^{-8}$	$.25 \times 10^{-8}$	$.11 \times 10^{-8}$	$.16 \times 10^{-8}$
2-nd refinement: $\ \bar{A}\varphi_j\ , j = 1, \dots, 20$ .				
.94	.80	.60	.42	.29
.21	$.14 \times 10^{-1}$	$.68 \times 10^{-2}$	$.26 \times 10^{-2}$	$.82 \times 10^{-3}$
$.23 \times 10^{-3}$	$.56 \times 10^{-4}$	$.12 \times 10^{-4}$	$.23 \times 10^{-5}$	$.39 \times 10^{-6}$
$.63 \times 10^{-7}$	$.99 \times 10^{-8}$	$.21 \times 10^{-8}$	$.44 \times 10^{-10}$	$.48 \times 10^{-11}$

Nevertheless our refinement process shows a fair improvement. One needs at least 20 elements for  $10^{-3}$  accuracy and 16 elements for  $10^{-2}$  accuracy with the initial basis; with the refined basis one needs 9 elements for  $10^{-3}$  accuracy and 7 elements for  $10^{-2}$  accuracy.

All of our present dimensionality reduction methods are not really worth the trouble in one space variable. Consider for instance the Example 3 with  $\approx 10^{-4}$  accuracy. We have reduced from a least square method with a  $60 \times 60$  matrix  $\bar{A}^T \bar{A}$  using the full 60 dimensional basis, to a  $19 \times 19$  matrix  $(\bar{A}Q_{19})^T (\bar{A}Q_{19})$  using the sinusoidal initial basis  $\varphi_1, \dots, \varphi_{19} \sim \sqrt{2} \sin(1\pi x), \dots, \sqrt{2} \sin(19\pi x)$ , then finally to a  $4 \times 4$  matrix  $(\bar{A}Q_4)^T (\bar{A}Q_4)$  using the refined basis  $\varphi_1, \dots, \varphi_4$ . Of course, computing the full  $60 \times 60$  matrix  $\bar{A}^T \bar{A}$  would involve solving the parabolic equation with 60 different initial functions, which would be a bit or work, but the inversion of the  $60 \times 60$  least squares matrix  $\bar{A}^T \bar{A} + \varepsilon^2 I$  could then be easily accomplished.

The true advantages of these methods should arise for problems with several space variables. Let us concoct a multivariable example (admittedly artificial) in which a great reduction demonstrably does occur. We construct this example from our one dimensional Example 1, thus avoiding the much larger computing cost of having to apply our refinement method to examples with two space variables.

Consider the parabolic equation

$$(31) \quad u_t = (a(x, t)u_x)_x + (a(y, t)u_y)_y, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1, \quad 0 \leq t \leq T,$$

with the conditions:

$$(32) \quad u = 0 \text{ on the lateral boundary, } u = \xi_{ij}(x, y) = \varphi_i(x)\varphi_j(y) \text{ at } t = 0,$$

where  $a(x, t)$  is the coefficient function of Example 1 and the  $\varphi_j$  are the orthonormal basis functions of Example 1 (either the full basis, the sinusoidal basis or the refined basis). Now, because of the separability of this problem (i.e., the commutativity of the  $x$  and  $y$  differential operators in (28)), we find that the discrete solution at time  $T$  is given by  $\bar{A}\xi_{ij} = \bar{A}_1\varphi_i(x)\bar{A}_2\varphi_j(y)$  where  $\bar{A}_1$  is the solution operator (with respect to  $x$ ) of Example 1 and  $\bar{A}_2$  is the same solution operator (with respect to  $y$ ) of Example 1. Notice that the  $\xi_{ij}$  form an orthonormal basis on the square and that:

$$\|\bar{A}\xi_{ij}\| = \|\bar{A}_1\varphi_i\| \|\bar{A}_2\varphi_j\|.$$

Now, letting the  $\varphi_j$  be the sinusoidal basis we see that  $\|\bar{A}\xi_{ij}\| \geq 10^{-4}$  exactly for the 107 index pairs  $\{(1, 1)-(1, 20), (2, 2)-(2, 14), (2, 16)-(2, 18), (3, 3)-(3, 7), (3, 9)-(3, 14), (3, 16)-(3, 18), (13, 13), (13, 14), (13, 16), (14, 14) \text{ and the symmetric pairs}\}$  and that these norm values drop off rapidly for larger values of the indices. Let  $S_{107}$  denote this set of indices, let  $Q_{107}$  denote the projection onto the space spanned by the complementary basis elements  $\xi_{ij}, i, j \notin S_{107}$  and notice that:

$$\|\bar{A}Q_{107}\| \geq \max_{i,j \notin S_{107}} \|\bar{A}\xi_{ij}\| \geq 10^{-4}.$$

Thus it seems that at least 107 sinusoidal basis elements (and possibly quite a few more) are needed for  $10^{-4}$  accuracy.

Letting the  $\varphi_j$  be the refined basis of Example 1, however, we see that we can let  $S_{17}$  be the set of 17 index pairs  $(1, 1)-(1, 5); (2, 1)-(2, 4); (3, 1)-(3, 4), (4, 1)-(4, 3), (5, 1)$ . With only these 17 refined basis elements we then have at least  $10^{-4}$  accuracy. In fact, the numerical evidence from Example 2, where we did a second refinement, indicates that the first few elements  $\varphi_1, \dots, \varphi_7$  of our first refined basis are almost exactly equal to the first few eigenfunctions  $\beta_1, \dots, \beta_7$  of  $\bar{A}_1^T \bar{A}_1$  and  $\bar{A}_2^T \bar{A}_2$ . Notice, moreover,

because of the commutativity of  $\bar{A}_1$  and  $\bar{A}_2$ , that the eigenfunctions of  $\bar{A}^T \bar{A}$  are exactly of the form  $\beta_i(x)\beta_j(y) \approx \varphi_i(x)\varphi_j(y)$ . Hence  $\|\mathbf{A}\mathbf{Q}_{17}\| \cong \max_{(i,j) \in S_{17}} \|\bar{A}\xi_{ij}\| \cong .7 \times 10^{-4}$ .

Notice that the block power method with 17 elements would eventually give us exactly the 17 largest eigenlements

$$\{\bar{\xi}_{ij}(x, y) = \beta_i(x)\beta_j(y), \text{ for } (i, j) \in S_{17}\}.$$

(Probably for purposes of faster convergence for the block power method we would carry along a few more basis elements, say  $M = 30$ , then proceed with an eigenvalue routine such as `ESIPACK`, as we have done before, to pick out the dominant 17 or so eigenlements from this refined space of dimension 30.)

In this two variable example we have therefore reduced the required dimensionality of our approximate basis from  $60 \times 60 = 3,600$  for the original full basis, to 17 for a refined basis. Notice that computation of the parabolic solution  $\bar{A}\xi$  for a single initial function  $\xi(x, y)$  is a big job, but easily feasible. So also would be the computation of the parabolic solutions  $\bar{A}^T \bar{A}\xi_1, \dots, \bar{A}^T \bar{A}\xi_{30}$  for several steps of our block power refinement method. Computation of the parabolic solutions  $\bar{A}\xi_1, \dots, \bar{A}\xi_{107}$  for the sinusoidal approximate basis would be a much larger problem, and so would be the inversion of the corresponding  $3,600 \times 3,600$  nonsparse least square equations would capabilities of modern computers. However, computation of the parabolic solutions  $\bar{A}\varphi_1, \dots, \bar{A}\varphi_{3,600}$  for the full basis would require an enormous expenditure of time, and inversion of the corresponding  $3,600 \times 3,600$  non sparse least square equations would be beyond the memory and speed limitations of modern computers.

One final point should be called to attention. In some cases (such as when the final time  $t = T$  is rather large) the eigenvalues of  $\mathbf{A}^T \mathbf{A}$  die out so rapidly that it is rather easy to accurately compute (by a variety of methods) the required first few exact eigenvalues and eigenfunctions. (As we have said, the numerical evidence in our Example 2 seems to indicate that we have found almost exactly the first seven eigenvalues and eigenfunctions with only one application of our block power method; W. Kahan has pointed out to us that a Lanczos type method would probably give even faster convergence.) In such a case, having evidence that we have found nearly exact eigenvalues and eigenfunctions, it would probably be best to apply the method of partial eigenfunction expansion (4)–(7) rather than the least squares method (8)–(9).

#### REFERENCES

- [1] S. AGMON, *Unicité et convexité dans les problèmes différentiels*, Seminar Univ. de Montréal, Les Presses d l'Univ. Montréal, 1966.
- [2] G. BACKUS, *On inference from incomplete and inaccurate data I*, Proceedings Nat. Acad. Sci. U.S.A., 65 (1970), pp. 1–8.
- [3] R. BELLMAN, R. KALABA AND K. LOCKETT, *Numerical Inversion of the Laplace Transform*, American Elsevier, New York, 1966.
- [4] B. BUZBEE AND A. CARASSO, *On the numerical computation of parabolic problems for preceding times*, Technical Report No. 299, Dept. of Math. and Stat., Univ. of New Mexico, Albuquerque, November 1971.
- [5] G. GOLUB AND W. KAHAN, *Calculating the singular values and pseudo inverse of a matrix*, J. Soc. Indust. Appl. Math. ser. B, Number Anal., 2 (1965), pp. 205–224.
- [6] K. MILLER, *Three circle theorems in partial differential equations and applications to improperly posed problems*, Arch. Rational Mech. Anal., 16 (1964), pp. 126–154.
- [7] ———, *Least squares methods for ill-posed problems with a prescribed bound*, this Journal, 1 (1970), pp. 52–74.
- [8] ———, *Stabilized quasi-reversibility and other nearly-best-possible methods for non-well-posed problems*, Symp. on Non-well-posed Problems and Logarithmic Convexity, Springer Lecture Notes # 316, Springer-Verlag, New York, 1973, pp. 161–176.

- [9] K. MILLER, *Nonunique continuation for certain ODE's in Hilbert space and for uniformly parabolic and elliptic equations in self-adjoint divergence form*, Symp. on Non-well-posed Problems and logarithmic Convexity, Springer Lecture Notes #316, Springer-Verlag, New York, 1973, pp. 85–101.
- [10] K. MILLER AND G. A. VIANO, *On the necessity of nearly-best-possible methods for analytic continuation of scattering data*, J. Math. Phys., 14 (1973), pp. 1037–1048.
- [11] K. MILLER, *Efficient numerical methods for backward solution of parabolic equations with variable coefficients*, Improperly Posed Boundary Value Problems, Research Notes in Mathematics, Series No. 1, A Carasso and A. P. Stone, eds., Putnam Publishing, London, 1975.
- [12] ———, *Math 228A Notes, Part II on stiff equations*, UC Berkeley, Fall '73, available from author.
- [13] V. A. MOROZOV, *Choice of parameter for the solution of functional equations by the regularization method*, Dokl. Akad. Nauk SSSR, 175, no. 6 (1967). (Translated in Soviet Math. Dokl., 8, no. 4 (1967), pp. 1000–1003.)
- [14] A. N. TICHONOV, *On the solution of ill-posed problems and the method of regularization*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 501–504.

## CANONICAL FACTORIZATIONS OF DISCONJUGATE DIFFERENTIAL OPERATORS\*

ANTONIO GRANATA†

**Abstract.** In a previous paper W. F. Trench (1974) proved that it is always possible to factorize a linear ordinary differential operator  $L$ , disconjugate on  $(a, b)$ , in the form  $Lu \equiv p_n(p_{n-1}(\cdots(p_0u)'\cdots))'$  with  $\int_a^b (1/p_i) = +\infty$  or  $\int^b (1/p_i) = +\infty$  ( $i = 1, \dots, n-1$ ). Following on from this we consider factorizations with the conditions  $\int_a (1/p_i) < +\infty$  or  $\int^b (1/p_i) < +\infty$  ( $i = 1, \dots, n-1$ ). All circumstances where it is possible to obtain factorizations of such two types are characterized, taking into account the behavior at both endpoints. In doing so two subclasses of disconjugate operators on  $(a, b)$  are pointed out: the well-known one consisting of those operators which are also disconjugate on  $[a, b]$  and another one with properties opposite, so to say, to those of the first subclass, as far as the double asymptotic behavior at the two endpoints is concerned. In the first case some new characterizations are added to the many already known whereas the second subclass is interesting in itself. Some of the results are especially important if viewed as useful lemmas in studying global or asymptotic properties of solutions to perturbed disconjugate equations.

**1. Introduction.** A generic nonempty interval of  $\mathbb{R}$  will be denoted by  $\mathcal{I}; (a, b)$ ,  $-\infty \leq a < b \leq +\infty$ , is an open interval. All functions are real-valued.  $L_{loc}^1(\mathcal{I})$  denotes the set of functions which are integrable on every compact subset of  $\mathcal{I}$ ;  $C^k(\mathcal{I})$  and  $AC^k(\mathcal{I})$  denote respectively the set of functions with continuous or absolutely continuous  $k$ th derivatives on  $\mathcal{I}$ .  $L_n$  ( $n \in \mathbb{N}$ ) will stand for an  $n$ th order linear ordinary differential operator represented by

$$(1.1) \quad L_n u \equiv u^{(n)} + a_1(t)u^{(n-1)} + \cdots + a_n(t)u, \quad \forall u \in AC^{n-1}(\mathcal{I}).$$

When there is no ambiguity we will write  $L$  instead of  $L_n$ ; it is tacitly assumed that  $n \geq 2$ . Such an operator will be called type (\*) on  $\mathcal{I}$  if it can be represented by (1.1) with  $a_i \in L_{loc}^1(\mathcal{I}), \forall i$ . An operator  $L_n$  of type (\*) on  $\mathcal{I}$  (or the equation  $L_n u = 0$ ) is termed *disconjugate* on  $\mathcal{I}$  if every nontrivial solution of  $L_n u = 0$  has at most  $n - 1$  zeros on  $\mathcal{I}$ , counting multiplicities. Fundamental papers concerning disconjugate operators are those by Polya [25], Hartman [6], [7], [8] and Levin [17] whereas important monographs and textbooks on the subject are those by Karlin [9], Karlin–Studden [10], Coppel [2], Willett [38].

A fundamental property of an  $n$ th order disconjugate operator is that it can be represented as a symbolic product of  $n$  first-order operators. The proof of the following theorem may be found in Polya [25] and Mammana [21] for particular cases and in Levin [17, Cor. 2.2, p. 62] and Rosati [26] in its full generality.

**THEOREM 1.1.** *Let  $L_n$  be an operator of type (\*) on an open interval  $\mathcal{I}$  (bounded or not); then the following properties are equivalent:*

- i)  $L_n$  is disconjugate on  $\mathcal{I}$ .
- ii) Equation  $L_n u = 0$  has a fundamental system of solutions,  $u_1, \dots, u_n$ , such that

$$(1.2) \quad W(u_1, \dots, u_k) > 0 \quad \text{on } \mathcal{I}, \quad k = 1, \dots, n,$$

where  $W(u_1, \dots, u_k) \equiv W(u_1(t), \dots, u_k(t))$  is the Wronskian determinant of  $u_1, \dots, u_k$  and  $W(u) \equiv u$ .

- iii)  $L_n$  has a factorization of the type

$$(1.3) \quad L_n u \equiv p_n[p_{n-1}(\cdots(p_1(p_0u)')\cdots)]', \quad \forall u \in AC^{n-1}(\mathcal{I}),$$

\* Received by the editors June 6, 1978 and in final revised form January 23, 1979.

† Dipartimento di Matematica, Università della Calabria, C.P. 9-87030 Roges (Cosenza)-Italy.



where the  $p_i$ 's are suitable functions such that

$$(1.4) \quad p_i(t) > 0 \quad \forall t \in \mathcal{T}, \quad \forall i; \quad p_i \in AC^{n-1-i}(\mathcal{T}), \quad 0 \leq i \leq n-1; \quad p_n \in AC^0(\mathcal{T}).$$

iv)  $L_n$  has a factorization of the type

$$(1.5) \quad L_n u \equiv (D + \bar{p}_1)(D + \bar{p}_2) \cdots (D + \bar{p}_n), \quad \forall u \in AC^{n-1}(\mathcal{T}),$$

where  $Du \equiv u'$  and the  $p_i$ 's are suitable functions.

Factorizations of type (1.3) or (1.5) are useful when studying general properties of disconjugate equations as shown, for example, by Polya [25], Mammana [20], [21], Zedek [39], Hartman [7], [8], Levin [17]; however in studying global or asymptotic problems related to perturbed disconjugate equations of the form  $L_n u = f(t, u, u', \dots, u^{(n-1)})$ , it sometimes appears that factorization (1.3) is not very useful in itself. In fact, when the method of "variation of constants" is applied to the equation, one finds that the possible solutions for the problem at hand (multipoint boundary value problems, asymptotic behavior of solutions, etc.) satisfy certain integral equations which can be easily studied only if the coefficients  $p_i$  of (1.3) are subject to suitable integrability conditions at one or both endpoints of  $\mathcal{T}$ . Trench [32] has shown that every operator  $L$ , disconjugate on  $(a, b)$ , has a factorization (1.3) such that

$$(1.6) \quad \int_a^b (1/p_i) = +\infty, \quad i = 1, \dots, n-1 \quad \text{or} \quad \int_a^b (1/p_i) = +\infty, \quad i = 1, \dots, n-1.$$

The usefulness of such factorizations can be clearly seen in Kartsatos [11], Kusano–Naito [13], Lovelady [18], [19], Trench [33] and also in papers dealing with functional differential equations viewed as perturbations of a disconjugate equation, among which we mention only Grammatikopoulos [3], [4] and Philos, Sficas, Staikos, Stavroulakis [22], [23], [24], [27], [28], [29]. Furthermore the entire asymptotic theory of Willett [34]–[38] could be simplified by the use of Trench's results. The following condition has also been considered

$$(1.7) \quad \int_a^b (1/p_i) < +\infty, \quad i = 1, \dots, n-1,$$

cf. Granata [5], Kusano–Onose [14]–[16], Kartsatos [12]. Such factorizations can sometimes yield results complementary to those obtained by working with factorizations of the Trench type, see [5, Thms. 3.1 and 3.3].

Continuing on Trench's line we shall study the existence of factorizations satisfying (1.7) at one or both endpoints or satisfying (1.6) at both endpoints. Throughout we shall only use factorizations of type (1.3). Section 2 contains definitions of and some general facts about the two types of factorizations, while § 3 sets out the main results of the paper: proofs are to be found in § 5. In § 4 there are some examples.

**2. Two types of canonical factorizations.**

DEFINITION 2.1. Let  $\mathcal{T} = (a, b)$ ,  $-\infty \leq a < b \leq +\infty$ ; the symbol  $D_n(\mathcal{T}) \equiv D_n(a, b)$  will denote the family of all the operators  $L_n$  of type (\*) disconjugate on  $\mathcal{T}$ .

DEFINITION 2.2. A factorization of type (1.3) of an operator  $L \in D_n(a, b)$  is said to be a canonical factorization (C.F. for short) of type (I) [resp. of type (II)] at the endpoint

$a$  if the functions  $p_i$  satisfy not only (1.4) but also the following conditions

$$(2.1) \quad \int_a (1/p_i) = +\infty, \quad i = 1, \dots, n-1,$$

$$(2.2) \quad \left[ \text{resp. } \int_a (1/p_i) < +\infty, \quad i = 1, \dots, n-1 \right].$$

An analogous definition holds at the endpoint  $b$  replacing  $\int_a$  with  $\int^b$ .

Such factorizations will be also termed “global” on  $(a, b)$  since they represent the operator  $L$  on all of  $(a, b)$  and the coefficients  $p_i$  satisfy (1.4) on  $(a, b)$ . On the other hand a factorization will be termed “local” at  $a$  [resp. at  $b$ ] if it represents the operator  $L$  on an interval of the form  $(a, a+r)$  [resp.  $(b-r, b)$ ],  $r > 0$ , or if the  $p_i$ 's satisfy (1.4) on such an interval only. The term factorization, referred to a given operator  $L \in D_n(a, b)$ , will always stand for a global factorization on  $(a, b)$ .

Canonical factorizations of type (I) are those studied by Trench who proved the following fundamental result.

**THEOREM 2.1** (Trench [32]). *Every operator  $L \in D_n(a, b)$  has a C.F. of type (I) at  $a$  and a similar C.F. at  $b$ . Furthermore every C.F. of type (I) at an endpoint is “essentially” unique in the sense that conditions (2.1) determine  $p_0, \dots, p_n$  up to multiplicative constants with product 1.*

Where C.F.'s of type (II) are concerned the situation is different: indeed an operator  $L \in D_n(a, b)$  may have no (global) C.F. of type (II) or may have an infinity of “essentially” different ones. For instance the operator  $d^2/dt^2$ , regarded as an element of  $D_2(-\infty, +\infty)$ , has no global C.F. of type (II) both at  $-\infty$  and at  $+\infty$ . This could be proved by writing in full the identity  $u'' \equiv p_2(p_1(p_0u))'$  and then using simple devices but is left to the reader because it is a particular case of a more general result to be proved below. On the other hand the same operator, regarded as an element of  $D_2(T, +\infty)$  or  $D_2(-\infty, T)$ , where  $T$  is any real number, has an infinite number of essentially different C.F.'s of type (II) at  $+\infty$  [resp. at  $-\infty$ ], namely,

$$u''(t) = (t-c)^{-1} \left[ (t-c)^2 \left( \frac{u(t)}{t-c} \right)' \right]',$$

where  $c$  is any constant  $< T$  [resp.  $> T$ ].

The following result is a counterpart of Theorem 2.1 and asserts the existence of local factorizations of type (II).

**THEOREM 2.2.** *Let  $L \in D_n(a, b)$ ,  $-\infty \leq a < b \leq +\infty$ . Then for every  $t_0$ ,  $a < t_0 < b$ , there exists a C.F. of  $L$  in the interval  $(a, t_0)$  which is of type (II) at  $a$  and a C.F. in  $(t_0, b)$  which is of type (II) at  $b$ .*

**3. Statement of the main result.** Generally speaking it is not true that a global C.F. of type (I) or (II) at an endpoint is also a C.F. at the other, and, even when it is so, it is not necessarily of the same type. We shall characterize all the possible situations and the existence of global factorizations of type (II). In doing so we shall point out the interdependency between canonical factorizations, hierarchical systems, generalized disconjugacy and double asymptotic behavior of solutions. Some definitions are needed.

**DEFINITION 3.1.** Let  $L \in D_n(a, b)$ ; a C.F. of  $L$  is said to be *mixed* if it is of type (I) at one endpoint and of type (II) at the other; it is termed *double* if it is of the same type at both endpoints.

DEFINITION 3.2. Let  $L_n$  be of type  $(*)$  on  $(a, b)$  and let  $t_0 \in [a, b]$ , possibly  $t_0 = \pm\infty$ . An ordered  $n$ -tuple  $(u_1, \dots, u_n)$  of solutions of  $L_n u = 0$  is called a *hierarchical system* at  $t_0$  if there exists a deleted neighborhood  $N$  of  $t_0$  such that

- i)  $u_k(t) \neq 0, t \in N, k = 1, \dots, n,$
- ii)  $\lim_{t \rightarrow t_0} u_k(t)/u_{k+1}(t) = 0, k = 1, \dots, n - 1.$

Note that the order of  $(u_1, \dots, u_n)$  is vital in Definition 3.2 and that every hierarchical system at a point  $t_0$  is a fundamental system of solutions on  $(a, b)$ .

DEFINITION 3.3. An  $n$ -tuple  $(u_1, \dots, u_n)$  is called a *double hierarchical system* on  $(a, b)$  if it is a hierarchical system at both  $a$  and  $b$ ; it is called a *mixed hierarchical system* on  $(a, b)$  if  $(u_1, \dots, u_n)$  is hierarchical at  $a$  and  $(u_n, \dots, u_1)$  is hierarchical at  $b$  (or vice versa).

The locution ‘‘hierarchical system’’ is used by Levin whereas Hartman, Willett and Trench use ‘‘principal system’’. The concept of a mixed hierarchical system is that of a fundamental principal system as given by Willett [36, Def. 1.5].

Hartman [7, Thm. 7.2, p. 331 and Thm. A, p. 353] and Levin [17, Lemma 2.1, p. 58] showed that every disconjugate equation on  $(a, b)$  has a hierarchical system at  $a$  and another one at  $b$ ; on the other hand the existence of a mixed or double hierarchical system is only assured for certain subclasses of disconjugate equations. The following theorem establishes the equivalence between the existence of a global C.F. of type (II), of a mixed C.F. and of a mixed hierarchical system: it is a completion to Trench’s results [32, Thm. 2].

THEOREM 3.1. For  $L \in D_n(a, b)$  the following are equivalent properties:

- 1) The C.F. of type (I) at  $a$  is a C.F. of type (II) at  $b$ ;
- 2) The C.F. of type (I) at  $b$  is a C.F. of type (II) at  $a$ ;
- 3)  $L$  admits of a global C.F. of type (II) at  $a$ ;
- 4)  $L$  admits of a global C.F. of type (II) at  $b$ ;
- 5)  $Lu = 0$  has a mixed hierarchical system on  $(a, b)$ .

Extending the concept of ‘‘zero’’ of a function and the consequent notion of disconjugacy (see Levin [17] and Willett [36], [37]) many other characterizations may be added to those in Theorem 3.1. By so doing we shall show the equivalence between Trench’s results reported in Theorem 3.1 and Willett’s disconjugacy criteria for singular equations [36]. The following Definitions 3.4 and 3.5 may also be found in Willett [36] to whom we refer the reader for further explanations about locutions and properties to be used. In such definitions  $L$  will be a fixed operator of order  $n$  and type  $(*)$  on an open interval  $(a, b), -\infty \leq a < b \leq +\infty$ . The concepts to be defined depend vitally on the choice of  $L$ .

DEFINITION 3.4. A solution  $u$  of  $L_n u = 0$  has a *zero of order  $k, 0 \leq k \leq n - 1$* , at a point  $t_0 \in [a, b]$  (possibly  $t_0 = \pm\infty$ ) provided that there exists a hierarchical system at  $t_0, (u_1, \dots, u_n)$ , and that any one of the following properties (equivalent to each other) holds:

- i) there exists a constant  $c \neq 0$  such that:

$$u(t) = cu_{n-k}(t)[1 + o(1)], \quad t \rightarrow t_0,$$

- ii)  $\lim_{t \rightarrow t_0} u(t)/u_j(t) \begin{cases} = 0, & j = n, n - 1, \dots, n - k + 1, \\ \neq 0, & j = n - k, \end{cases}$

- iii) there exist constants  $c_1, \dots, c_{n-k}$  such that  $c_{n-k} \neq 0$  and

$$u(t) = c_1 u_1(t) + \dots + c_{n-k} u_{n-k}(t), \quad a < t < b.$$

DEFINITION 3.5. Denote the number of zeros, counted as prescribed in Definition 3.5, of a solution  $u$  in an interval  $J \subset [a, b]$  by  $Z_u J$  and let  $Z_u t_0 \equiv Z_u [t_0, t_0]$ . Equation  $L_n u = 0$  is *disconjugate* on  $J$  if for any solution  $u \neq 0$  we have  $Z_u J \leq n - 1$ ; it is *disconjugate at a point  $t_0$*  if it is *disconjugate* on some neighborhood of  $t_0$ .

Zeros will be considered throughout only in the above sense. We may now state:

THEOREM 3.2. *If  $L$  is an operator of type (\*) on  $(a, b)$  the following are equivalent properties:*

- 1)  $L$  is *disconjugate* on  $[a, b]$ ;
- 2)  $L$  is *disconjugate at any one of the two endpoints, say  $a$ , and for every interval  $[a, a + r] \subset [a, b]$  equation  $Lu = 0$  has a mixed hierarchical system on  $(a, a + r)$ ;*
- 3)  $L$  is *disconjugate on  $(a, b)$  and  $Lu = 0$  has a mixed hierarchical system on  $(a, b)$  (hence all properties of Theorem 3.1 hold true);*
- 4) *for every interval  $[\alpha, \beta] \subset [a, b]$   $Lu = 0$  has a mixed hierarchical system on  $(\alpha, \beta)$ .*

Note that Theorems 3.1 and 3.2 are illustrated by the operator  $d^n/dt^n$  on any interval of the form  $(-\infty, T)$  or  $(T, +\infty)$ ,  $T \in \mathbb{R}$ .

Returning to the C.F.'s two cases must still be examined: the existence of double C.F.'s of type (I) or (II). Operators with a double C.F. of type (II) on  $(a, b)$  are obviously a subclass of operators *disconjugate* on  $[a, b]$ : we would like to point out that if at least one of the two endpoints is nonsingular for  $L$  in the Willett sense [36] then the existence of a double C.F. of type (II) is equivalent to the existence of a global C.F. of type (II) at one (suitable) endpoint. However we prefer to postpone both this minor point and some complementary results to a further paper. Instead we shall focus our attention on the existence of double C.F.'s of type (I) which characterize an interesting subclass of *disconjugate operators* on  $(a, b)$  entirely different from that of the foregoing theorems. A last definition is needed.

DEFINITION 3.6. Let  $L_n$  be of type (\*) on  $(a, b)$ . Two ordered  $n$ -tuples of solutions of  $Lu = 0$ ,  $(u_1, \dots, u_n)$  and  $(v_1, \dots, v_n)$ , are *asymptotically equivalent at a point  $t_0 \in [a, b]$ , possibly  $t_0 = \pm\infty$ , provided there exist  $n$  nonzero constants  $c_1, \dots, c_n$  such that:  $u_k(t) = c_k v_k(t)[1 + o(1)]$ ,  $t \rightarrow t_0$ .*

Suppose  $L \in D_n(a, b)$ ; then an immediate consequence of the property of hierarchy is that any two hierarchical systems at the same endpoint are asymptotically equivalent at that endpoint. On the other hand two hierarchical systems at an endpoint may not be asymptotically equivalent at the other endpoint as shown by the following example: let  $L \equiv d^n/dt^n$  on  $(0, +\infty)$ , then both systems  $(1; t; \dots; t^{n-1})$  and  $(1; t+1; \dots; t^{n-1}+1)$  are hierarchical at  $+\infty$  but are not asymptotically equivalent at  $t_0 = 0$ . The following theorem establishes a connection between double C.F.'s of type (I), double hierarchical systems and hierarchical systems asymptotically equivalent at both endpoints.

THEOREM 3.3. *If  $L \in D_n(a, b)$  the following are equivalent properties:*

- 1)  $L$  has a double C.F. of type (I) on  $(a, b)$ ;
- 2)  $L$  has only one (positive constant factors apart) C.F. on  $(a, b)$ ;
- 3)  $Lu = 0$  has a double hierarchical system on  $(a, b)$ ;
- 4) any hierarchical system of  $Lu = 0$  at an endpoint is a double hierarchical system on  $(a, b)$ ;
- 5) any two hierarchical systems at the same endpoint are asymptotically equivalent at both endpoints.

Remarks. 1. This theorem is illustrated by the operator  $d^n/dt^n$  on  $(-\infty, +\infty)$ . 2. A mixed hierarchical system, if any, is unique positive constant factors apart (see e.g., Trench [32, Cor. 2]), whereas all hierarchical systems are double as soon as there is one

such system. 3. It is the author's feeling that property 2) might be replaced by the stronger one: 2a)  $L$  has only "one" factorization of type (1.3)–(1.4) on  $(a, b)$ . Obviously 2a)  $\Rightarrow$  2) by Theorem 2.1, but for the present we cannot say anything about the converse with the exception of the second order case, which is elementary: see § 4, Example A.

**4. Examples.**

**A. Second order operators.** For such operators Theorems 3.1, 3.2, 3.3 cover all possible cases. As a matter of fact if  $Lu \equiv p_2(p_1(p_0u))'$  on  $(a, b)$ , then the possible cases are: i)  $\int_a^b (1/p_1) = \int_a^b (1/p_1) = +\infty$ ; ii)  $\int_a^b (1/p_1) < +\infty$ ; iii)  $\int_a^b (1/p_1) < +\infty$ . In the first case Theorem 3.3 holds; in cases ii) and iii) Theorem 3.1 applies. The same reasoning can be applied to those operators which admit of a factorization (1.3) such that there exist positive constants  $a_i, b_i$  and a function  $p(t) > 0$  on  $\mathcal{T}$  with  $a_i p(t) \leq p_i(t) \leq b_i p(t), t \in \mathcal{T}, i = 1, \dots, n - 1$ . The author [5], [5a] has already pointed out the special asymptotic properties (as  $t \rightarrow +\infty$ ) of the solutions of equations of the form  $Lu + f(t, u) = 0$ , where  $L$  is an operator of such a type on  $(T, +\infty)$ . In the particular case  $p_i \equiv p \forall i$ , this class of operators is contained in that studied by Šeda [30], [31].

**B. Constant coefficient operators on  $(-\infty, +\infty)$ .** Let  $L$  be the operator with constant real coefficients defined by

$$(4.1) \quad Lu \equiv u^{(n)} + a_{n-1}u^{(n-1)} + \dots + a_0u \quad \forall u \in C^n(\mathbb{R}),$$

and let

$$(4.2) \quad r^n + a_{n-1}r^{n-1} + \dots + a_1r + a_0 = 0$$

be its characteristic equation. All the properties spoken about previously can be characterized via the roots of (4.2). We re-collect them in the following theorem whose elementary proof is left to the reader.

**THEOREM 4.1.** *Let  $L$  be the operator (4.1),  $n \geq 1$ , and let  $r_1, \dots, r_n$  be the  $n$  complex roots of (4.2), each counted according to its multiplicity. Then:*

I)  $L$  admits, on  $\mathbb{R}$ , of the two factorizations:  $Lu \equiv \prod_{i=1}^n ((d/dt) - r_i)u$ , which is of the type (1.5), and

$$Lu \equiv e^{r_1 t} [e^{(r_{n-1} - r_1)t} (e^{(r_{n-2} - r_{n-1})t} (\dots (e^{(r_1 - r_2)t} (e^{-r_1 t} u)') \dots)')]'$$

which is of the type (1.3).. Hence  $L$  is disconjugate on  $(-\infty, +\infty)$  iff (4.2) has only real roots.

II)  $L$  is disconjugate on  $[-\infty, +\infty]$  (see Theorems 3.1–3.2) iff all roots of (4.2) are real and distinct.

III)  $L$  has the properties of Theorem 3.3 iff all roots of (4.2) are real and equal.

**C. Factorizations of the operator  $d^n/dt^n$ .** (See also Carlitz [1].) Let  $L \equiv d^n/d^n$  on an interval  $\mathcal{T}$  and consider the two factorizations

$$(4.3) \quad Lu \equiv u^{(n)} \quad (\text{i.e., } p_i \equiv 1, \quad i = 0, \dots, n),$$

$$(4.4) \quad Lu \equiv (t-a)^{1-n} \left[ (t-a)^2 \left( \dots \left[ (t-a)^2 \left( \frac{u}{(t-a)^{n-1}} \right)' \right] \dots \right)' \right]',$$

where  $a \in \mathbb{R}$  is fixed. The identity (4.4) is easily verified by checking that the kernel of the operator  $\tilde{L}$  defined by the right-hand side of (4.4) is the same as that of  $L$ . *First case:*  $\mathcal{T} \equiv (a, b)$  bounded. Then (4.3) is a double C.F. of type (II) on  $\mathcal{T}$ , while (4.4) is a mixed C.F. on  $\mathcal{T}$ , of type (I) at  $a$  and of type (II) at  $b$ . *Second case:*  $\mathcal{T} \equiv (a, +\infty)$ . Both (4.3) and (4.4) are mixed C.F.'s on  $\mathcal{T}$ : the first one is of type (II) at  $a$  and of type (I) at  $+\infty$ ; vice versa for the second one. A double C.F. of type (II) on  $(a, +\infty)$  may be obtained from

(4.4) replacing  $a$  with any real number  $\bar{a} < a$ . *Third case:*  $\mathcal{T} \equiv \mathbb{R}$ . Then (4.3) is the “only one” C.F. of  $L$  on  $\mathcal{T}$ : it is of type (I) at both  $\pm\infty$ .

**5. Proofs.** Proofs of Theorems 2.2 and 3.1 are no more than slight remarks on and alterations of the proof of Trench’s Theorem 1 [32]. Lest we become obscure we must reproduce a part of Trench’s reasonings. In Lemmas 5.1–5.4 we will use the same notations as Trench.

LEMMA 5.1 (Trench [32, Lemma 1]). *If the operator*

$$(5.1) \quad L = \frac{1}{\xi_2} \frac{d}{dt} \frac{1}{\xi_1} \frac{d}{dt} \frac{\cdot}{\xi_0}, \quad \xi_i > 0$$

*is in  $D_2(a, b)$  with  $\int^b \xi_1 < +\infty$  then  $L$  can be represented in the form*

$$(5.2) \quad L = \frac{1}{\eta_2} \frac{d}{dt} \frac{1}{\eta_1} \frac{d}{dt} \frac{\cdot}{\eta_0}, \quad \eta_i > 0,$$

*where*

$$(5.3) \quad \eta_0(t) = \xi_0(t) \cdot \int_t^b \xi_1, \quad \eta_1(t) = \xi_1(t) \cdot \left( \int_t^b \xi_1 \right)^{-2}, \quad \eta_2(t) = \xi_2(t) \cdot \int_t^b \xi_1.$$

*Hence it is  $\int^b \eta_1 = +\infty$  and  $\int_a \eta_1 < +\infty$ , no matter what the behavior of  $\xi_1$  at the point  $a$ .*

According to our terminology this lemma says that if an operator  $L \in D_2(a, b)$  has a global C.F. of type (II) at  $b$  then it has a mixed C.F. on  $(a, b)$ .

LEMMA 5.2. *If in the factorization (5.1) we have  $\int^b \xi_1 = +\infty$ , then in each interval  $(t_0, b)$ ,  $a < t_0 < b$ ,  $L$  can be represented in the form (5.2), where*

$$(5.4) \quad \eta_0(t) = \xi_0(t) \cdot \int_{t_0}^t \xi_1, \quad \eta_1(t) = \xi_1(t) \cdot \left( \int_{t_0}^t \xi_1 \right)^{-2}, \quad \eta_2(t) = \xi_2(t) \cdot \int_{t_0}^t \xi_1.$$

*Hence we have  $\int^b \eta_1 < +\infty$  and  $\int_{t_0} \eta_1 = +\infty$ .*

*Remarks.* 1. Proofs of the foregoing lemmas consist in a straightforward check that (5.1)–(5.2) define the same operator. 2. In Lemma 5.2 it is permissible to choose  $t_0 = a$  only if  $\int_a \xi_1 < +\infty$ . 3. Analogous statements hold when the symbols  $\int^b, \int_t^b, \int_{t_0}^t$  are replaced throughout respectively by  $\int_a, \int_a^t, \int_{t_0}^t$ . The latter also applies to the following two lemmas.

LEMMA 5.3 (Trench [32, Lemma 2]). *If the operator*

$$(5.5) \quad L = \frac{1}{\mu_3} \frac{d}{dt} \frac{1}{\mu_2} \frac{d}{dt} \frac{1}{\mu_1} \frac{d}{dt} \frac{\cdot}{\mu_0}, \quad \mu_i > 0$$

*is in  $D_3(a, b)$  with  $\int^b \mu_1 = +\infty$  and  $\int^b \mu_2 < +\infty$ , then  $L$  can be represented in the form*

$$(5.6) \quad L = \frac{1}{\nu_3} \frac{d}{dt} \frac{1}{\nu_2} \frac{d}{dt} \frac{1}{\nu_1} \frac{d}{dt} \frac{\cdot}{\nu_0}, \quad \nu_i > 0,$$

*where  $\int^b \nu_i = +\infty, i = 1, 2$ . Furthermore*

$$(5.7) \quad \int_a \nu_i < +\infty \quad (i = 1, 2) \quad \text{provided} \quad \int_a \mu_i < +\infty \quad (i = 1, 2).$$

Note that by Lemma 5.1 it is always possible to suppose  $\int^b \mu_1 = +\infty$ . Property (5.7) is not explicitly pointed out by Trench; however it is easy to convince oneself that it does hold by re-reading the original proof and paying due attention to the two possibilities examined therein.

LEMMA 5.4. *If in the factorization (5.5) we have  $\int^b \mu_1 < +\infty$  and  $\int^b \mu_2 = +\infty$  then in each interval  $(t_0, b)$ ,  $a < t_0 < b$ ,  $L$  can be represented in the form (5.6) with  $\int^b \nu_i < +\infty$  ( $i = 1, 2$ ).*

Note that by Lemma 5.2 it is always possible to suppose  $\int^b \mu_1 < +\infty$  provided we consider (5.5) only in an interval  $(a + r, b)$ ,  $r > 0$ .

*Proof of Lemma 5.4.* It is essentially the proof of Lemma 2 [32] with some obvious modifications patterned on Lemma 5.2 which we shall sketch briefly: firstly the symbol  $\int_i^b$  in [32, p. 323] is replaced throughout by  $\int_{t_0}^t$ , where  $t_0$  has been chosen in advance; then a word-for-word repetition of the argument leads one to prove that  $L$  can be represented in the form (5.6) with  $\int^b \nu_1 < +\infty$ . We still have to show that  $\int^b \nu_2 < +\infty$ . We choose  $t_1, t_2$  such that  $t_0 < t_1 < t_2 < b$  and evaluate  $\int_{t_1}^{t_2} \nu_2$ ; by similar arguments as in [32, p. 323] the following relation is easily found

$$\begin{aligned} \int_{t_1}^{t_2} \nu_2 &= -\left(\int_{t_0}^{t_2} \mu_2\right)^{-1} \cdot \left(\int_{t_0}^{t_2} \tilde{\nu}_1\right) + \int_{t_1}^{t_2} \mu_1 + \left(\int_{t_0}^{t_1} \mu_1\right)^{-1} \cdot \left(\int_{t_0}^{t_1} \tilde{\nu}_1\right) \\ &\equiv -I_1 + I_2 + I_3. \end{aligned}$$

Now hold  $t_1$  fixed and let  $t_2 \rightarrow b^-$ :  $I_2, I_3$  are bounded by assumptions; furthermore the decreasing character of the function  $M_2(t) \equiv \left(\int_{t_0}^t \mu_2\right)^{-1}$  and formulas (2.10) of [32], modified as specified above, give

$$I_1 \equiv M_2(t_2) \cdot \int_{t_0}^{t_2} \tilde{\nu}_1(\tau) d\tau \leq \int_{t_0}^{t_2} \tilde{\nu}_1(\tau) M_2(\tau) d\tau \equiv \int_{t_0}^{t_2} \mu_1 \leq \int_{t_0}^b \mu_1 < +\infty.$$

Hence  $\int^b \nu_2 < +\infty$ .  $\square$

*Proof of Theorem 2.2.* Let us consider factorizations on an interval  $(t_0, b)$ . Lemmas 5.2 and 5.4 establish the theorem for  $n = 2, 3$ . If  $n \geq 4$  one may proceed by induction as in Trench [32, Thm. 1]: the whole proof remains unchanged when  $\int_i^b$  is replaced throughout by  $\int_{t_0}^t$ .  $\square$

Before attempting to prove the theorems in § 3 we must reconsider our Definition 3.2. In current literature the inequality in condition i) is found to be systematically replaced by the stronger one: i')  $u_k(t) > 0$ .

Due to the linearity of the operator  $L$  this replacement is obviously immaterial when dealing with hierarchical systems at a single point; but when considering, as in Definition 3.3, the behavior of solutions at both endpoints the difference is substantial. Our definition is the most suitable one for our purposes: for instance the equation  $u'' = 0$  has a double hierarchical system on  $(-\infty, +\infty)$  according to Definition 3.2, but it does not have any such system on the same interval if condition i') is used. On the other hand Theorems 3.1, 3.2 are intimately related to results, concerning mixed hierarchical systems, due to other authors: our first step will thus be to show that conditions i) and i') are equivalent when dealing with mixed hierarchical systems of solutions to disconjugate equations. From now on we put for brevity

$$\begin{aligned} u \ll v, t \rightarrow t_0 &\Leftrightarrow u = o(v), t \rightarrow t_0, \\ u \sim v, t \rightarrow t_0 &\Leftrightarrow u = v[1 + o(1)], t \rightarrow t_0. \end{aligned}$$

*The following lemma is a slight weakening of the hypotheses concerning the signs of  $u_n$  in Proposition 15, p. 117, of Coppel's [2].*

LEMMA 5.5. Let  $L \in D_n(a, b)$ ,  $-\infty \leq a < b \leq +\infty$ , and let  $(u_1, \dots, u_n)$  be a fundamental system of solutions to  $Lu = 0$  on  $(a, b)$  such that:

- i)  $u_k > 0$  on a neighborhood of  $b$ ,  $k = 1, \dots, n$ ;
- ii)  $u_1 \ll \dots \ll u_n$ ,  $t \rightarrow b$ ;
- iii)  $u_k \neq 0$  on a neighborhood of  $a$ ,  $k = 1, \dots, n$ ;
- iv)  $u_n \ll \dots \ll u_1$ ,  $t \rightarrow a$ .

It then follows that  $u_k > 0$  on  $(a, b)$  for  $k = 1, \dots, n$ .

*Proof.* As well known, see [17, Lemma 2.3, p. 61], the hypothesis  $L \in D_n(a, b)$  implies that  $L$  is disconjugate on  $[a, b]$  and  $(a, b]$ . This and definitions 3.4–3.5 imply that a nontrivial solution to  $Lu = 0$  on  $(a, b)$ , say  $v$ , such that  $Z_v b = n - 1$  (as the function  $u_1$  in our hypotheses) or  $Z_v a = n - 1$  (as the function  $u_n$ ) must have the same strict sign on all of  $(a, b)$ . Hence our proposition holds trivially for  $n = 2$ . Let us proceed by induction on  $n$ . Let  $n > 2$  and suppose Lemma 5.5 has been proved for any operator of the class  $D_{n-1}(a, b)$ . Hypotheses i)–ii) imply, by [17, Thm. 2.1, p. 66], that  $W(u_1, \dots, u_k) > 0$  on  $(a, b)$ ,  $k = 1, \dots, n$ . Consider now the equation of order  $n - 1$ ,

$$Lu \equiv W(u_1, \dots, u_{n-1}, u) / W(u_1, \dots, u_n) = 0, \quad \text{which is disconjugate on } (a, b).$$

It has  $(u_1, \dots, u_{n-1})$  as a fundamental system. From i),  $\dots$ , iv) and the inductive hypothesis it follows that  $u_k > 0$  on  $(a, b)$ ,  $k = 1, \dots, n - 1$ . But, by iv),  $u_n$  has  $n - 1$  zeros at  $a$  and hence it is  $> 0$  on  $(a, b)$ .  $\square$

*Proof of Theorem 3.1.* For the equivalences  $1) \Leftrightarrow 2) \Leftrightarrow 5)$  see both Trench [32, Thm. 2] and Lemma 5.5. We now prove  $3) \Rightarrow 2)$  the converse being obvious.

Let  $Lu = p_n(p_{n-1}(\dots(p_0 u)'\dots))'$  with  $p_i > 0$  on  $(a, b)$  and

$$(5.8) \quad \int_a (1/p_i) < +\infty, \quad i = 1, \dots, n - 1.$$

Then if  $n = 2, 3$  Lemmas 5.1 and 5.3 prove the existence of a factorization of  $L$  on  $(a, b)$  which is of type (I) at  $b$  and of type (II) at  $a$ . Let  $n \geq 4$  and let us construct a global C.F. of  $L$  on  $(a, b)$  of type (I) at  $b$ . When we repeat word for word the inductive procedure used by Trench [32, Thm. 1, pp. 323–324] the desired factorization is achieved: what is more, by reviewing the proof, it is easily seen (as pointed out after Lemma 5.3) that, due to (5.8), the new factorization is, like the old one, of type (II) at  $a$ .

The equivalence  $4) \Leftrightarrow 1)$  is proved similarly to  $3) \Leftrightarrow 2)$ , interchanging the roles of  $a$  and  $b$ .  $\square$

*Proof of Theorem 3.2.* The equivalences  $1) \Leftrightarrow 2) \Leftrightarrow 4)$  are a restatement of Willett's results [36]: they were originally enunciated for  $L$  of type (1.1) with continuous coefficients, but they also hold for  $L$  of type (\*) as is explicitly pointed out by Willett.  $1) \Rightarrow 3)$ : see either Willett [36, Thm. 1.2] or Levin [17, Lemma 4.1, p. 80]. We now show  $3) \Rightarrow 1)$ . Let  $(u_1, \dots, u_n)$  be a fundamental system for  $Lu = 0$  on  $(a, b)$  such that

$$(5.9) \quad \begin{cases} u_1 \gg u_2 \gg \dots \gg u_n, & t \rightarrow a, \\ u_1 \ll u_2 \ll \dots \ll u_n, & t \rightarrow b. \end{cases}$$

Let  $u$  be a solution of  $Lu = 0$  with at least  $n$  zeros on  $[a, b]$ ; we must show that  $u \equiv 0$ . Let  $t_i$  be  $n$  such zeros,  $a \leq t_1 \leq t_2 \leq \dots \leq t_n \leq b$ , each of them counted as many times as its multiplicity; three cases will be distinguished. *First case:*  $t_i \neq a \forall i$  or  $t_i \neq b \forall i$ . It follows that  $u \equiv 0$  since  $L$  is disconjugate on  $[a, b)$  and on  $(a, b]$ . *Second case:* there exists  $r \in \{1, \dots, n - 1\}$  such that  $t_i = a$  ( $i = 1, \dots, r$ ) and  $t_i = b$  ( $i = r + 1, \dots, n$ ); i.e.,  $Z_u a = r$  and  $Z_u b = n - r$ . From (5.9) and Definition 3.4 it follows that there exist suitable constants  $c_i$  such that the two relations  $u = c_{r+1}u_{r+1} + \dots + c_n u_n$  and  $u = c_1 u_1 + \dots + c_r u_r$



simultaneously hold. They imply that  $c_i = 0 \forall i$ : i.e.,  $u \equiv 0$ . If  $n = 2$  the proof is complete; if  $n \geq 3$  there is a third case to be examined, namely when  $Z_u a = r, Z_u b = s, Z_u(a, b) \geq n - r - s$ , where  $1 \leq r, s \leq n - 2$  and  $2 \leq r + s \leq n - 1$ .

If  $u$  has the representation  $u = c_1 u_1 + \dots + c_n u_n$ , then we have:

$$Z_u a = r \Rightarrow c_1 = c_2 = \dots = c_r = 0, \quad Z_u b = s \Rightarrow c_n = c_{n-1} = \dots = c_{n-s+1} = 0.$$

Hence  $u$  has a representation of the type

$$(5.10) \quad u = c_{r+1} u_{r+1} + \dots + c_{n-s} u_{n-s},$$

in which the right-hand side has  $n - r - s$  terms. Observe now that a result quoted in Coppel [2, Prop. 15, p. 117] implies that  $(u_1, \dots, u_n)$  is a Cartesian (or Descartes) system on  $(a, b)$ : see definition on page 87 of [2] or on page 67 of [17]. If then  $u \neq 0$ , it would follow from (5.10) that  $u$  has at most  $n - r - s - 1$  zeros on  $(a, b)$  while, by hypothesis,  $Z_u(a, b) \geq n - r - s$ : a contradiction.  $\square$

*Remarks.* 1. Proposition 15 on p. 117 of Coppel's [2] is enunciated for  $L$  of type (1.1) with continuous coefficients but it holds for  $L$  of type (\*) too.

2. By retracing all steps of an argument by Polya [25; proof of Thm. II, p. 317] and using Polya's mean value theorem in the generalized version given by Willett [37, Thm. 1.1], it is possible to give a different proof of the third case examined above. But such an alternative proof, though using the full force of Willett's advanced result, would however be confined to operators with continuous coefficients owing to the fact that Willett's theorem loses its meaning for operators of type (\*).

*Proof of Theorem 3.3.* 1)  $\Leftrightarrow$  2): an obvious consequence of Theorem 2.1.

1)  $\Rightarrow$  3). Suppose  $L$  has the representation (1.3) with  $\int_a (1/p_i) = \int^b (1/p_i) = +\infty$  ( $i = 1, \dots, n - 1$ ). Let  $T$  be any number,  $a < T < b$ , and consider the functions  $u_k$  defined by

$$u_1 = 1/p_0, \quad u_{k+1}(t) = \frac{1}{p_0(t)} \int_T^t \frac{dt_1}{p_1(t_1)} \int_T^{t_1} \frac{dt_2}{p_2(t_2)} \dots \int_T^{t_{k-1}} \frac{dt_k}{p_k(t_k)} \quad (k = 1, \dots, n - 1).$$

They form a fundamental system on  $(a, b)$  and it can be trivially checked that

$$(5.11) \quad u_1 \ll u_2 \ll \dots \ll u_n, \quad \text{both as } t \rightarrow a \text{ and as } t \rightarrow b.$$

3)  $\Rightarrow$  1). By hypothesis  $Lu = 0$  has a fundamental system  $(u_1, \dots, u_n)$  satisfying (5.11). Now let (1.3) be the C.F. of type (I) at  $b$ . We must show that  $\int_a (1/p_i) = +\infty$ ,  $i = 1, \dots, n - 1$ . (A similar proof holds interchanging  $a$  and  $b$ ).

Let  $L_0 u \equiv p_0 u$ ;  $L_i u \equiv p_i (L_{i-1} u)'$  ( $1 \leq i \leq n$ ). Hence  $Lu = L_n u$ . By Corollary 3 of Trench [32] we have  $L_i u_k = 0$  ( $1 \leq k \leq i$ ) and  $L_i u_{i+1} = r_i$  (= constant  $\neq 0$ ) ( $0 \leq i \leq n - 1$ ).

It is immaterial whether we assume  $r_i = 1 \forall i$ . Therefore,  $T$  being a fixed point in  $(a, b)$ , the functions  $u_k$  satisfy the following relations

$$(5.12) \quad u_1 = 1/p_0, \quad u_{k+1}(t) = \frac{1}{p_0(t)} \int_T^t \frac{1}{p_1} \int_T^{t_1} \frac{1}{p_2} \dots \int_T^{t_{k-1}} \frac{1}{p_k} + \sum_{i=u}^k c_{ki} u_i$$

$$(1 \leq k \leq n - 1),$$

where the  $c_{ki}$ 's are suitable constants. To prove  $\int_a (1/p_i) = +\infty$  we use induction on  $i$ . If  $i = 1$  we have from (5.11)

$$u_1(t)/u_2(t) = \left( \int_T^t (1/p_1) + c_{1,1} \right)^{-1} = o(1), \quad t \rightarrow a.$$

Hence  $\int_a^T (1/p_1) = +\infty$ . Suppose now  $i > 1$  and that  $\int_a (1/p_i) = +\infty$  for  $i = 1, \dots, k$ . By (5.11), (5.12) we obtain

$$(5.13) \quad \begin{aligned} 0 &= \lim_{t \rightarrow a} u_{k+1}(t)/u_{k+2}(t) \\ &= \lim_{t \rightarrow a} \int_T^t \frac{1}{p_1} \cdots \int_T^{t_{k-1}} \frac{1}{p_k} / \int_T^t \frac{1}{p_1} \cdots \int_T^{t_k} \frac{1}{p_{k+1}}. \end{aligned}$$

Because of the inductive hypothesis we may apply l'Hôpital's rule  $k$  times to the right-hand limit in (5.13): this gives rise to  $\lim_{t \rightarrow a} (\int_T^t (1/p_{k+1}))^{-1}$ . Since  $p_{k+1} > 0$ , such a limit exists (finite or infinite): hence it must be  $= 0$ . This is the same as  $\int_a^T (1/p_{k+1}) = +\infty$ .

4)  $\Rightarrow$  3): obvious. 3)  $\Rightarrow$  4). Let  $(u_1, \dots, u_n)$  be a double hierarchical system on  $(a, b)$ , i.e., satisfying (5.11), and let  $(v_1, \dots, v_n)$  be any hierarchical system at an endpoint, say  $b$ . From the elementary properties of hierarchical systems it follows that there exist constants  $c_{ij}$  such that:  $v_k = c_{k1}u_1 + \dots + c_{kk}u_k$ ,  $c_{kk} \neq 0 \forall k$ . This and (5.11) imply

$$(5.14) \quad v_k \sim c_{kk}u_k, \quad \text{both as } t \rightarrow a \text{ and as } t \rightarrow b,$$

i.e.,  $(v_1, \dots, v_n)$  is a double hierarchical system.

Relation (5.14) also proves the implication 3)  $\Rightarrow$  5). The proof of 5)  $\Rightarrow$  3) requires an intermediate step supplied by the following

LEMMA 5.6. *If  $L \in D_n(a, b)$  then  $Lu = 0$  has a fundamental system,  $(u_1, \dots, u_n)$ , hierarchical at an endpoint, say  $a$ , and with the following property at  $b$ : "For each  $k \in \{1, \dots, n-1\}$  either  $u_k \ll u_{k+1}$ ,  $t \rightarrow b$  or  $u_{k+1} \ll u_k$ ,  $t \rightarrow b$  is true".*

*Proof of Lemma 5.6.* Let  $(u_1, \dots, u_n)$  be any hierarchical system at  $a$ . Let us show how it is possible to construct another system,  $(v_1, \dots, v_n)$  hierarchical at  $a$  and with the desired property at  $b$ . We recall that, [17, Lemma 2.1], for any two nontrivial solutions  $u$  and  $v$ , and for any point  $t_0 \in [a, b]$ , one of the following three circumstances always occurs:

$$u \ll v, \quad v \ll u, \quad u \sim cv, \quad c = \text{const.} \neq 0 \quad (t \rightarrow t_0).$$

Now we set  $v_1 = u_1$ ; to find  $v_2$  we compare  $u_2$  with  $v_1$  as  $t \rightarrow b$ . If  $v_1 \ll u_2$ ,  $t \rightarrow b$ , or  $u_2 \ll v_1$ ,  $t \rightarrow b$ , we set  $v_2 = u_2$ ; while if there exists a constant  $c_1 \neq 0$  such that

$$(5.15) \quad v_1 = c_1 u_2 + o(u_2), \quad t \rightarrow b,$$

we set  $v_2 = v_1 - c_1 u_2$ . We thus have:

$$\begin{aligned} v_1 &= u_1, & v_2 &\sim \gamma_2 u_2, & t &\rightarrow a \quad (\text{in both cases}), & \gamma_2 &\neq 0, \\ v_2 &= o(u_2) = o(v_1), & t &\rightarrow b \quad (\text{in the second case, because of (5.15)}). \end{aligned}$$

To find  $v_3$  we compare  $u_3$  with  $v_2$ : if  $v_2 \ll u_3$ ,  $t \rightarrow b$ , or  $u_3 \ll v_2$ ,  $t \rightarrow b$ , we set  $v_3 = u_3$ . If  $v_2 = c_2 u_3 + o(u_3)$ ,  $t \rightarrow b$ , we set  $v_3 = v_2 - c_2 u_3$ . As for  $v_2$  it follows that

$$\begin{aligned} v_3 &\sim \gamma_3 u_3, & t &\rightarrow a \quad (\text{in both cases}), & \gamma_3 &\neq 0, \\ v_3 &= o(u_3) = o(v_2), & t &\rightarrow b \quad (\text{in the second case}). \end{aligned}$$

An iteration of the procedure yields the desired basis.  $\square$

We return to the proof of 5)  $\Rightarrow$  3) in Theorem 3.3. Let  $(u_1, \dots, u_n)$  be a system with the same two properties as in Lemma 5.6: we shall show that property 5) implies that such a system is a double hierarchical system on  $(a, b)$ . Consider the new system  $(\bar{u}_1, \dots, \bar{u}_n)$  defined by:  $\bar{u}_k \equiv u_1 + \dots + u_k$ ,  $k = 1, \dots, n$ . It is obviously hierarchical at  $a$  like the former: hence, by 5), it is asymptotically equivalent to  $(u_1, \dots, u_n)$  both at  $a$

and  $b$ . In particular there exists a constant  $c_2 \neq 0$  such that  $\bar{u}_2 \equiv u_1 + u_2 = c_2 u_2 + o(u_2)$ ,  $t \rightarrow b$ ; hence  $u_1 = (c_2 - 1)u_2 + o(u_2)$ ,  $t \rightarrow b$ . Because of the property stated in Lemma 5.6 it must follow that  $c_2 - 1 = 0$ : i.e.,  $u_1 \ll u_2$ ,  $t \rightarrow b$ . By induction on  $i$  we will prove that:  $u_i \ll u_{i+1}$ ,  $t \rightarrow b$ ,  $i = 1, \dots, n-1$ . Suppose this is true for  $i = 1, \dots, k-1$ ; as above there exists a constant  $c_{k+1} \neq 0$  such that

$$\bar{u}_{k+1} \equiv u_1 + \dots + u_{k+1} = c_{k+1} u_{k+1} + o(u_{k+1}), \quad t \rightarrow b,$$

i.e.,  $u_1 + \dots + u_k = (c_{k+1} - 1)u_{k+1} + o(u_{k+1})$ ,  $t \rightarrow b$ .

Using the inductive hypothesis we infer that

$$u_k[1 + o(1)] = (c_{k+1} - 1)u_{k+1} + o(u_{k+1}), \quad t \rightarrow b,$$

and, by Lemma 5.6,  $u_k \ll u_{k+1}$ ,  $t \rightarrow b$ . This completes the proof of  $5) \Rightarrow 3)$  and of Theorem 3.3.  $\square$

**Acknowledgment.** This paper was written while the author was a member of the Italian Committee for Scientific Research C.N.R.-G.N.A.F.A.

#### REFERENCES

- [1] L. CARLITZ, *A theorem on differential operators*, Amer. Math. Monthly, 83 (1976), pp. 351-354.
- [2] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics, vol. 220, Springer-Verlag, Berlin, 1971.
- [3] M. K. GRAMMATIKOPOULOS, *Asymptotic and oscillation criteria for nonlinear retarded differential equations of arbitrary order*, Tech. Rep. 102 (1977), Univ. of Ioannina, Greece.
- [4] ———, *A criterion for the existence of bounded nonoscillatory solutions for nonlinear retarded differential equations*, Ibid., 110 (1977).
- [5] A. GRANATA, *Singular Cauchy problems and asymptotic behaviour for a class of  $n$ -th order differential equations*, Funkcial Ekvac., 20 (1977), pp. 193-212.
- [5a] ———, *Corrigendum and addendum: Singular Cauchy problems and  $\dots$* , Funkcial. Ekvac., submitted.
- [6] P. HARTMAN, *Disconjugate  $n$ -th order differential equations and principal solutions*, Bull. Amer. Math. Soc., 74 (1968), pp. 125-129.
- [7] ———, *Principal solutions of disconjugate  $n$ -th order linear differential equations*, Amer. J. Math., 91 (1969), pp. 306-362.
- [8] ———, *Corrigendum and addendum: Principal solutions of disconjugate  $n$ -th order linear differential equations*, Ibid., 93 (1971), pp. 439-451.
- [9] S. KARLIN, *Total positivity*, vols. I-II, Stanford University Press, Stanford, California, 1968.
- [10] S. KARLIN AND W. STUDDEN, *Tchebycheff systems: with applications in analysis and statistics*, Interscience, New York, 1966.
- [11] A. G. KARTSATOS, *Oscillation properties of even order differential equations*, Bull. Fac. Sci. Ibaraki Univ. Ser. A, no. 2-1 (1969), pp. 9-14.
- [12] ———, *Oscillation and existence of unique positive solutions for nonlinear  $n$ -th order equations with forcing term*, Hiroshima Math. J., 6 (1976), pp. 1-6.
- [13] T. KUSANO AND M. NAITO, *Nonlinear oscillation of fourth order differential equations*, Canad. J. Math., 28 (1976), pp. 840-852.
- [14] T. KUSANO AND H. ONOSE, *A nonoscillation theorem for a second order sublinear retarded differential equation*, Bull. Austral. Math. Soc., 15 (1976), pp. 401-406.
- [15] ———, *Nonoscillation theorems for differential equations with deviating argument*, Pacific J. Math., 63 (1976), pp. 185-192.
- [16] ———, *Nonoscillation theorems for second order differential equations with forcing term*, Quart. J. Math. Oxford, (2) 28 (1977), pp. 191-200.
- [17] A. YU. LEVIN, *Non-oscillation of solutions of the equation  $x^{(n)} + p_1(t)x^{(n-1)} + \dots + p_n(t)x = 0$* , Uspehi Mat. Nauk, 24 (1969), no. 2 (146), pp. 43-96; Russian Math. Surveys 24 (1969), no. 2, pp. 43-99.
- [18] D. L. LOVELADY, *On the oscillatory behavior of bounded solutions of higher order differential equations*, J. Differential Equations, 19 (1975), pp. 167-175.
- [19] ———, *Some oscillation criteria for fourth order differential equations*, Rocky Mountain J. Math., 5 (1975), pp. 593-600.

- [20] G. MAMMANA, *Sopra un nuovo metodo di studio delle equazioni differenziali lineari*, Math. Z., 25 (1926), pp. 734–748.
- [21] ———, *Decomposizione delle espressioni differenziali lineari omogenee in prodotti di fattori simbolici e applicazione relativa allo studio delle equazioni differenziali lineari*, Ibid., 33 (1931), pp. 186–231.
- [22] CH. G. PHILOS, *Oscillatory and asymptotic behavior of all solutions of differential equations with deviating arguments*, Proc. Roy. Soc. Edinburgh Sect. A, 81A (1978), pp. 195–210.
- [23] ———, *Some results on the oscillatory and asymptotic behavior of the solutions of differential equations with deviating arguments*, Technical Report 123 (1977), Univ. of Ioannina, Greece.
- [24] CH. G. PHILOS, Y. G. SFICAS AND V. A. STAIKOS, *Some results on the asymptotic behavior of nonoscillatory solutions of differential equations with deviating arguments*, Ibid., 121 (1977).
- [25] G. POLYA, *On the mean-value theorem corresponding to a given linear homogeneous differential equation*, Trans. Amer. Math. Soc., 24 (1922), pp. 312–324.
- [26] F. ROSATI, *Su una classe di operatori differenziali*, Rend. Circ. Mat. Palermo (2), 19 (1970), pp. 69–88.
- [27] Y. G. SFICAS, *On the oscillatory and asymptotic behavior of damped differential equations with retarded arguments*, Hiroshima Math. J., 6 (1976), pp. 429–450.
- [28] Y. G. SFICAS AND I. P. STAVROULAKIS, *On the oscillatory and asymptotic behavior of a class of differential equations with deviating arguments*, this Journal, to appear.
- [29] I. P. STAVROULAKIS, *Oscillatory and asymptotic properties of differential equations with deviating arguments*, Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Natur., 60 (1976), pp. 611–622.
- [30] V. ŠEDA, *Über die Transformation der linearen Differentialgleichungen  $n$ -ter Ordnung. I*, Časopis Pěst. Mat., 90 (1965), pp. 385–412.
- [31] ———, *On a class of linear differential equations of order  $n$ ,  $n \geq 3$* , Ibid., 92 (1967), pp. 247–261.
- [32] W. F. TRENCH, *Canonical forms and principal systems for general disconjugate equations*, Trans. Amer. Math. Soc., 189 (1974), pp. 319–327.
- [33] ———, *Oscillation properties of perturbed disconjugate equations*, Proc. Amer. Math. Soc., 52 (1975), pp. 147–155.
- [34] D. WILLETT, *Asymptotic behavior of disconjugate  $n$ -th order differential equations*, Canad. J. Math., 23 (1971), pp. 293–314.
- [35] ———, *Generalized de la Vallée Poussin disconjugacy tests for linear differential equations*, Canad. Math. Bull., 14 (1971), pp. 419–428.
- [36] ———, *Disconjugacy tests for singular linear differential equations*, this Journal, 2 (1971), pp. 536–545. *Errata*, Ibid., 3 (1972), p. 559.
- [37] ———, *A generalization of Čaplygin's inequality with applications to singular boundary value problems*, Canad. J. Math., 25 (1973), pp. 1024–1039.
- [38] ———, *Oscillatory theory of  $n$ -th order linear differential equations*, Lecture notes (1973), Univ. of Utah, Salt Lake City.
- [39] M. ZEDEK, *Cayley's decomposition and Polya's  $W$ -property of ordinary linear equations*, Israel J. Math., 3 (1965), pp. 81–86.

## ON THE ANGULAR VARIATION OF SOLUTIONS OF SECOND ORDER LINEAR SYSTEMS\*

STEVEN D. TALIAFERRO†

**Abstract.** Upper bounds for the angular variation of extremal solutions of the second order linear system

$$x'' + P(t)x = 0,$$

where  $P(t)$  is a symmetric  $n \times n$  matrix, are obtained. The angular variation of a solution of the above equation is defined to be the length of its radial projection on the  $n - 1$  dimensional sphere. Also, if  $n = 2$  and  $x(t)$  is an extremal solution, then an upper bound, depending on the angular variation of  $x(t)$ , is obtained for the number of zeros of each component of  $x(t)$ . The proofs are based on variational arguments.

**1. Introduction.** In this paper we will consider the second order linear system

$$(1a) \quad x'' + P(t)x = 0,$$

$$(1b) \quad x(a) = x(b) = 0,$$

where  $P(t)$  is a real, positive definite, symmetric,  $n \times n$  matrix whose elements are continuous in  $t$  for  $a \leq t \leq b$ . We assume throughout this paper that if  $c \in (a, b)$  then (1a) has no nontrivial solution which vanishes at both  $a$  and  $c$ , and that (1a, b) has a nontrivial solution, (i.e. we assume the interval  $(a, b)$  contains no points conjugate to  $a$ , and that  $b$  is conjugate to  $a$ .) Second order linear systems play an important role in the calculus of variations where they appear as the Euler-Lagrange equations of the second variation of the functional  $J(y) = \int_a^b F(x, y, y') dx$ . For references to this material see [2], [4] and Chap. VII of [5].

In particular we will study the size of the angular variation of solutions,  $x(t)$ , of (1a, b). If we write  $x(t) = r(t)\theta(t)$  where  $r(t) = \|x(t)\|$  and  $\theta(t) = x(t)/\|x(t)\|$  then by the angular variation of  $x(t)$  we mean  $\int_a^b \|\theta'(t)\| dt$ . Geometrically, and in two dimensions, if the angular variation of a curve is  $\theta_0$ , then there is some sector of angle  $\theta_0$  and vertex at the origin such that the curve remains in that sector. In  $n$  dimensions, the angular variation of a curve is the length of its radial projection on the  $(n - 1)$  dimensional sphere.

We can write  $P = Q^T D Q$  where  $Q(t)$  is orthogonal and  $D(t)$  is diagonal for  $a \leq t \leq b$ . To motivate the results of this paper, consider the case  $Q(t)$  is constant, i.e. the eigenvectors of  $P$  are constant. Then by making the change of variable  $y = Qx$ , (1a, b) becomes  $y'' + D(t)y = 0$ ,  $y(a) = y(b) = 0$ . Clearly this problem has a nontrivial solution which remains on one of the coordinate axes, and corresponding to this solution, (1a, b) has a nontrivial solution with angular variation zero. Conversely, it will follow from the results of this paper that if  $D$  is constant, i.e. the eigenvalues of  $P$  are constant, then the angular variation of every solution of (1a, b) is less than  $\sqrt{n\pi^2 - 4}$  where  $n$  is the dimension of the space.

Some work of a similar nature has been done by Ahmad and Lazer [1] who proved that if each of the elements of the matrix  $P$  are positive on  $[a, b]$  then (1a, b) has a solution which remains in  $S = \{(x_1, x_2, \dots, x_n) : x_i \geq 0, i = 1, 2, \dots, n\}$ .

Also the concept of angular variation has been used by Schwarz [6] to study disconjugacy of first order systems.

\* Received by the editors November 14, 1977, and in final revised form January 15, 1979.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

**2. Results.** Let the eigenvalues of  $P(t)$  be  $\lambda_i(t)$ ,  $i = 1, 2, \dots, n$ . We can assume each  $\lambda_i(t)$  is continuous on  $[a, b]$  and  $\lambda_1(t) \geq \lambda_2(t) \geq \dots \geq \lambda_n(t)$ .

**THEOREM 1.** *Let*

$$\gamma(\lambda) = \begin{cases} \frac{(a+b-2t)(b-t)}{b-a}, & \text{if } a \leq t \leq \frac{a+b}{2}, \\ \frac{(2t-(a+b))(t-a)}{b-a}, & \text{if } \frac{a+b}{2} \leq t \leq b. \end{cases}$$

*Then the angular variation of every solution of (1a, b) is less than or equal to*

$$\frac{1}{2} \int_a^b \gamma(t)(\lambda_1(t) - \lambda_n(t)) dt.$$

**THEOREM 2.** *The angular variation of every solution of (1a, b) is less than or equal to*

$$\left[ (b-a) \int_a^b \lambda_1(t) dt - 4 \right]^{1/2}.$$

*Remark.* Since  $x^T P(t)x \leq x^T \lambda_1(t) I_n x^T$ ,  $x \in \mathbb{R}^n$ , we have by the results of Morse [4] that if  $c$  is the first zero, larger than  $a$ , of a solution of the scalar equation

$$\rho'' + \lambda_1(t)\rho = 0, \quad \rho(a) = 0,$$

then  $c \in (a, b]$ . Hence by the well-known Lyapunov inequality for scalar equations we have

$$(c-a) \int_a^c \lambda_1(t) dt - 4 > 0$$

and this inequality is also valid for  $c$  replaced with  $b$ . Thus the bound given for the angular variation in Theorem 2 is a real number, and Theorem 2 can be viewed as a generalization of Lyapunov's inequality.

The proof of Theorems 1 and 2 will be postponed until § 3. Next we will show how Theorem 2 can be used to obtain the result mentioned in the Introduction. First we need the following lemma; the short proof of it given below was pointed out to me by W. T. Reid.

**LEMMA 1.** *If  $c \in (a, b)$  then the problem*

$$(2a) \quad \rho'' + \frac{1}{n}(\lambda_1(t) + \dots + \lambda_n(t))\rho = 0,$$

$$(2b) \quad \rho(a) = \rho(c) = 0$$

*has only the trivial solution,  $\rho \equiv 0$ .*

*Proof.* Since, for the equation (1a),  $(a, c]$  contains no points conjugate to  $a$  we have by Theorem 2 of [2, p. 120] that

$$\int_a^c (h'^T h' - h^T P h) dt > 0$$

for continuously differentiable  $n$  dimensional vector valued functions  $h$  with  $h(a) = h(c) = 0$  and  $h \not\equiv 0$ . Taking  $h = e_j u$  where  $u$  is any continuously differentiable scalar function with  $u(a) = u(c) = 0$ ,  $u \not\equiv 0$ , and  $e_j$  is the  $j$ th canonical unit vector we have  $\int_a^c (u'^2 - p_{jj} u^2) dt > 0$ . Summing over  $j$  from 1 to  $n$  and dividing by  $n$  yields

$$(3) \quad \int_a^c \left[ u'^2 - \left( \frac{1}{n} \sum_{j=1}^n p_{jj} \right) u^2 \right] dt > 0.$$

But if  $\rho$  is a nontrivial solution of (2a, b) then multiplying (2a) by  $\rho$ , integrating by parts, and using (2b) yields

$$(4) \quad \int_a^c \left[ \rho'^2 - \left( \frac{1}{n} \sum_{j=1}^n \lambda_j \right) \rho^2 \right] dt = 0.$$

Using the fact that the trace of a matrix is the sum of its eigenvalues we see that (4) contradicts (3) and hence the lemma is proved.

**COROLLARY 1.** *Let  $S = \inf \{ (1/n)(\lambda_1(t) + \dots + \lambda_n(t)) : a \leq t \leq b \}$  and  $D = \sup \{ \frac{1}{2}(\lambda_1(t) - \lambda_n(t)) : a \leq t \leq b \}$ . Then  $b - a \leq \pi/\sqrt{S}$  and the angular variation of all solutions of (1a, b) is less than or equal to  $(D/S)_{12}^5 \pi^2$ .*

*Proof.* By the Sturm comparison theorems and Lemma 1, if  $c \in (a, b)$  then the problem

$$\rho'' + S\rho = 0, \quad \rho(a) = \rho(c) = 0$$

has only the trivial solution. Hence  $b - a \leq \pi/\sqrt{S}$ . So, by Theorem 1, if  $\theta_0$  is the angular variation of a solution of (1a, b) we have

$$\theta_0 \leq D \int_a^b \gamma(t) dt = D(b-a)^2 \frac{5}{12} \leq \frac{D}{S} \frac{5}{12} \pi^2.$$

This completes the proof of Corollary 1.

**COROLLARY 2.** *If  $S$  is as in Corollary 1 and  $M = \max \{ \lambda_1(t) : a \leq t \leq b \}$  then the angular variation of all solutions of (1a, b) is less than or equal to  $((M/S)\pi^2 - 4)^{1/2}$ .*

*Proof.* Since  $b - a \leq \pi/\sqrt{S}$  Corollary 2 follows from Theorem 2.

Note that if the eigenvalues of  $P$  are constant then  $M/S \leq n$ ; hence, by Corollary 2, the angular variation of every solution of (1a, b) is less than  $(n\pi^2 - 4)^{1/2}$ .

We conclude this section with an application of the angular variation of a solution  $x(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$  of (1a, b) when  $P(t)$  is the  $2 \times 2$  matrix  $\begin{pmatrix} p(t) & r(t) \\ r(t) & q(t) \end{pmatrix}$  with  $r(t) \neq 0$ . Since  $P(t)$  is positive definite we have  $p(t) > 0$ ,  $q(t) > 0$ , and  $r^2(t) < p(t)q(t)$  for  $a \leq t \leq b$ . Let  $r_0 = \min_{a \leq t \leq b} |r(t)|$ ,  $p_0 = \max_{a \leq t \leq b} p(t)$ , and  $q_0 = \max_{a \leq t \leq b} q(t)$ .

**THEOREM 3.** *Let  $m_u$  and  $m_v$  be the number of zeros in  $(a, b)$  of  $u$  and  $v$  respectively. Then*

$$m_u \leq \frac{\theta_0}{\tan^{-1}(r_0/p_0)} \quad \text{and} \quad m_v \leq \frac{\theta_0}{\tan^{-1}(r_0/q_0)}$$

where  $\theta_0$  is the angular variation of  $x(t)$ .

*Proof.* We establish the bound only for  $m_u$ ; the bound for  $m_v$  is obtained in a similar way. Let  $m = m_u$ . From (1a) we obtain

$$(5) \quad u'' + p(t)u + r(t)v = 0.$$

If  $c \in (a, b)$  and  $u''(c) = 0$  then  $v(c) \neq 0$ , for otherwise, by (5), we would have  $u(c) = 0$ , which contradicts (1a, b) is disconjugate on  $[a, b]$ . Also from (5) we obtain

$$p(c)u(c) + r(c)v(c) = 0$$

and hence

$$\frac{u(c)}{v(c)} = -\frac{r(c)}{p(c)}.$$

If  $d \in (a, b)$  and  $u(d) = 0$  then  $v(d) \neq 0$  and  $u(d)/v(d) = 0$ .

Let the zeros of  $u$  be  $a = d_0 < d_1 < d_2 < \dots < d_{m+1} = b$ . Choose  $t_1, t_2, \dots, t_{m+1}$  such that  $t_i \in (d_{i-1}, d_i)$  and  $u'(t_i) = 0$ . Choose  $c_2, c_3, \dots, c_{m+1}$  such that  $c_i \in (t_{i-1}, t_i)$  and  $u''(c_i) = 0$ .

For  $i = 2, 3, \dots, m + 1$  we have  $c_i, d_{i-1} \in (t_{i-1}, t_i)$ . Hence, if  $J_i$  is the interval with endpoints  $c_i$  and  $d_{i-1}$  then the intervals  $\{J_i\}_{i=2}^{m+1}$  are pairwise disjoint. Thus

$$\begin{aligned} \theta_0 &\cong \sum_{i=2}^{m+1} \int_{J_i} \|\theta'(t)\| dt \\ &\cong \sum_{i=2}^{m+1} \left| \tan^{-1} \frac{u(c_i)}{v(c_i)} - \tan^{-1} \frac{u(d_{i-1})}{v(d_{i-1})} \right| \\ &= \sum_{i=2}^{m+1} \tan^{-1} \left| \frac{r(c_i)}{p(c_i)} \right| \cong m \tan^{-1} \frac{r_0}{p_0}, \end{aligned}$$

and the proof of Theorem 3 is complete.

For example, if  $r(t) \equiv p(t) \equiv 1$  and  $q(t) > 1$  is continuous, and if  $b$  is the first point past  $a$  for which the boundary value problem

$$\begin{aligned} u'''' + (1 + q(t))u'' + (q(t) - 1)u &= 0, \\ u(a) = u''(a) = u(b) = u''(b) &= 0 \end{aligned}$$

has a nontrivial solution then  $u$  has at most  $(4/\pi)\theta_0$  zeros in  $(a, b)$ . Furthermore if  $m$  and  $M$  are the minimum and maximum of  $q(t)$  on  $[a, b]$  then, by Corollary 1,

$$\frac{4}{\pi} \theta_0 \cong \frac{\sqrt{(M-1)^2 + 4}}{1+m} \frac{5}{3} \pi.$$

**3. Proofs.**

DEFINITION. A pair of continuous, piecewise continuously differentiable functions,  $r: [c, d] \rightarrow R$  and  $\theta: [c, d] \rightarrow R^n$  will be called *admissible on  $[c, d]$*  if  $r(c) = r(d) = 0$ ,  $r$  is not identically zero on  $[c, d]$ , and  $\|\theta(t)\| = 1$  for all  $t \in [c, d]$ .

LEMMA 2. *If  $c < d$ ,  $[c, d]$  is a proper subset of  $[a, b]$ , and  $r$  and  $\theta$  are admissible functions on  $[c, d]$ , then*

$$(6) \quad \int_c^d [r'^2 + r^2(\theta'^T \theta' - \theta^T P \theta)] dt$$

is positive.

*Proof.* Let  $x(t) = r(t)\theta(t)$ . Then  $x$  is continuous and piecewise continuously differentiable on  $[c, d]$ ,  $x(c) = x(d) = 0$ , and  $x$  is not identically zero on  $[c, d]$ . Also (6) equals

$$(7) \quad \int_c^d (x'^T x' - x^T P x) dt.$$

Using the results of Morse [3] and the fact that  $(a, b)$  contains no points conjugate to  $a$ , it follows that  $(c, d)$  contains no points conjugate to  $c$ . Hence by [2, p. 120], (7) is positive.

If  $x(t)$  is a nontrivial solution of (1a, b), then clearly there exist functions  $r(t)$  and  $\theta(t)$  which are admissible on  $[a, b]$  such that  $x(t) = r(t)\theta(t)$  and  $r(t) > 0$  for  $a < t < b$ . The following lemma shows that in some sense  $r(t)$  is convex on  $[a, b]$ .

LEMMA 3. *If  $x(t) = r(t)\theta(t)$  is a nontrivial solution of (1a, b) where  $r(t)$  and  $\theta(t)$  are admissible on  $[a, b]$  and  $r(t) > 0$  on  $(a, b)$  then for each  $t \in (a, b)$  the line segment joining  $(a, 0)$  to  $(t, r(t))$  and the line segment joining  $(t, r(t))$  to  $(b, 0)$  both lie below the graph of  $r$ .*



*Proof.* Clearly  $r(t)$  and  $\theta(t)$  are twice continuously differentiable on  $(a, b)$ . Also, to prove the lemma it suffices to prove that no tangent line to the graph of  $r$  can intersect the  $t$  axis in  $(a, b)$ .

So let  $t_2 \in (a, b)$  and suppose the tangent line to  $r$  at  $t_2$  intersects the  $t$  axis at  $t_1$ , where  $t_1 \in (a, b)$ . We will show this leads to a contradiction if  $r'(t_2) > 0$ ; in a similar way, a contradiction can be reached if  $r'(t_2) < 0$ .

Define

$$\bar{r}(t) = \begin{cases} r(t), & t_2 \leq t \leq b, \\ r'(t_2)(t - t_2) + r(t_2), & t_1 \leq t \leq t_2. \end{cases}$$

Then  $\bar{r}$  is continuously differentiable on  $[t_1, b]$ ,  $\bar{r}''$  is piecewise continuous on  $[t_1, b]$ , and  $\bar{r}(t_1) = \bar{r}(b) = 0$ .

Define  $\bar{\theta}(t)$  to be  $\theta(t)$  for  $t_2 \leq t \leq b$  and define  $\bar{\theta}(t)$  to be a solution of

$$(8) \quad \phi'^T \phi' = \phi^T P(t) \phi, \quad \|\phi(t)\| \equiv 1, \quad \phi(t_2) = \theta(t_2)$$

for  $t_1 \leq t \leq t_2$ . The problem (8) has many solutions. For example we could take  $\phi(t) = \Gamma(s)$  where  $\Gamma(s)$  is a great circle on the unit sphere parameterized with respect to arc length,  $\Gamma(0) = \theta(t_2)$ , and  $s = \gamma(t)$ . Then if  $\gamma(t)$  satisfies the equation

$$(9) \quad \gamma'(t) = \sqrt{\Gamma^T(\gamma(t))P(t)\Gamma(\gamma(t))}, \quad \gamma(t_2) = 0,$$

then  $\phi(t) = \Gamma(\gamma(t))$  will satisfy (8). Since the right-hand side of (9) is bounded for  $t_1 \leq t \leq t_2$  and  $-\infty < \gamma < \infty$ , a solution of (9) exists on  $[t_1, t_2]$ . Clearly  $\bar{\theta}$  is continuous on  $[t_1, b]$  and  $\bar{\theta}'$  is piecewise continuous on  $[t_1, b]$ .

Since  $x(t) = r(t)\theta(t)$  satisfies (1a) we have

$$(10) \quad r''\theta + 2r'\theta' + r\theta'' = -rP\theta.$$

Multiplying (10) on the left with  $\theta^T$  and noting that  $\theta^T\theta = 1$ ,  $\theta^T\theta' = 0$ , and  $\theta^T\theta'' = -\theta'^T\theta'$  we obtain

$$(11) \quad r'' + [\theta^T P \theta - \theta'^T \theta'] r = 0.$$

By virtue of the way  $\bar{r}$  and  $\bar{\theta}$  are defined, they also satisfy (11) for  $t \in [t_1, b]$ . Therefore

$$\begin{aligned} & \int_{t_1}^b [\bar{r}''^2 + \bar{r}^2(\bar{\theta}'^T \bar{\theta}' - \bar{\theta}^T P \bar{\theta})] dt \\ &= \int_{t_1}^b (\bar{r}''^2 + \bar{r}'' \bar{r}) dt = \bar{r}' \bar{r} \Big|_{t_1}^b = 0. \end{aligned}$$

This contradicts Lemma 2 and proves Lemma 3.

LEMMA 4. If  $A$  is a symmetric, real,  $n \times n$  matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$  and if  $x$  and  $y$  are orthogonal vectors in  $R^n$ , then

$$(12) \quad |x^T A y| \leq \frac{1}{2}(\lambda_1 - \lambda_n) \|x\| \|y\|.$$

*Proof.* By virtue of the fact that  $A$  can be diagonalized using an orthogonal matrix, it suffices to prove the lemma in the case  $A$  is a diagonal matrix with  $\lambda_1, \lambda_2, \dots, \lambda_n$  on the diagonal. Clearly it suffices to prove (12) with the absolute value signs removed from the left side of (12). We can also assume  $\|x\| = \|y\| = 1$ . Thus proving the lemma is reduced to showing

$$(13) \quad \lambda_1 x_1 y_1 + \dots + \lambda_n x_n y_n \leq \frac{1}{2}(\lambda_1 - \lambda_n)$$

for  $x$  and  $y$  orthogonal unit vectors.

Note that we can assume  $x_1y_1 \geq x_2y_2 \geq \dots \geq x_ny_n$ , since permuting the coordinates doesn't change the fact that  $x$  and  $y$  are orthogonal unit vectors, but can increase the left side of (13). Suppose  $x_r y_r \geq 0$  and  $x_{r+1} y_{r+1} \leq 0$ . By the orthogonality of  $x, y$

$$\sum_1^n x_i y_i = \sum_1^r x_i y_i + \sum_{r+1}^n x_i y_i = 0$$

and since  $\|x\| = \|y\| = 1$

$$\left(\sum_1^r x_i y_i\right) + \left(-\sum_{r+1}^n x_i y_i\right) = \sum_1^n |x_i y_i| \leq \|x\| \cdot \|y\| = 1;$$

hence

$$0 \leq \sum_1^r x_i y_i = -\sum_{r+1}^n x_i y_i \leq \frac{1}{2}.$$

Thus

$$\begin{aligned} \lambda_1 x_1 y_1 + \dots + \lambda_n x_n y_n &\leq \lambda_1 (x_1 y_1 + \dots + x_r y_r) + \lambda_n (x_{r+1} y_{r+1} + \dots + x_n y_n) \\ &= (\lambda_1 - \lambda_n)(x_1 y_1 + \dots + x_r y_r) \\ &\leq \frac{1}{2}(\lambda_1 - \lambda_n). \end{aligned}$$

*Proof of Theorem 1.* Let  $x, r$ , and  $\theta$  be as in Lemma 3. Since  $x(t)$  satisfies (1a) we have

$$(14) \quad r''\theta + 2r'\theta' + r\theta'' = -rP\theta.$$

Multiplying (14) on the left with  $2r^3\theta'^T$  and noting that  $\theta'^T\theta = 0$  and  $2\theta'^T\theta'' = (\theta'^T\theta')'$ , and then integrating both sides of the resulting equation from  $a$  to  $t$  we get

$$(15) \quad r^4(t)\|\theta'(t)\|^2 = -2 \int_a^t r^4(\tau)\theta'^T(\tau)P(\tau)\theta(\tau) d\tau.$$

Taking the absolute value of both sides of (15) and using Lemma 4 we get

$$(16) \quad r^4(t)\|\theta'(t)\|^2 \leq \int_a^t r^4(\tau)(\lambda_1(\tau) - \lambda_n(\tau))\|\theta'(\tau)\| d\tau.$$

Let  $I(t)$  be the right-hand side of (16). Taking the square root of both sides of (16) and then multiplying both sides by  $r^2(t)(\lambda_1(t) - \lambda_n(t))$  yields

$$(17) \quad I'(t) \leq r^2(t)(\lambda_1(t) - \lambda_n(t))(I(t))^{1/2}.$$

Dividing both sides of (17) by  $2(I(t))^{1/2}$  and integrating from  $a$  to  $t$  gives

$$(18) \quad I(t)^{1/2} \leq \int_a^t r^2(\tau)\lambda(\tau) d\tau$$

where we are letting  $\lambda(t) = (1/2)(\lambda_1(t) - \lambda_n(t))$ . So by (16) and (18) we have

$$(19) \quad r^2(t)\|\theta'(t)\| \leq \int_a^t r^2(\tau)\lambda(\tau) d\tau.$$

By a similar argument we have

$$(20) \quad r^2(t)\|\theta'(t)\| \leq \int_t^b r^2(\tau)\lambda(\tau) d\tau.$$

By virtue of Lemma 3 we have

$$(21) \quad \frac{r(\tau)}{r(t)} \leq \begin{cases} \frac{b-\tau}{b-t}, & a \leq \tau \leq t \leq \frac{a+b}{2}, \\ \frac{\tau-a}{t-a}, & \frac{a+b}{2} \leq t \leq \tau \leq b. \end{cases}$$

Dividing both sides of (19) by  $r^2(t)$ , integrating the resulting inequality from  $a$  to  $(a+b)/2$  and using the first part of (21) gives

$$(22) \quad \int_a^{(a+b)/2} \|\theta'(t)\| dt \leq \int_a^{(a+b)/2} \gamma(t)\lambda(t) dt.$$

Doing the same thing with (20) and the second part of (21) yields

$$(23) \quad \int_{(a+b)/2}^b \|\theta'(t)\| dt \leq \int_{(a+b)/2}^b \gamma(t)\lambda(t) dt.$$

Inequalities (22) and (23) taken together prove Theorem 1.

*Proof of Theorem 2.* Let  $x, r,$  and  $\theta$  be as in Lemma 3. By Lemma 3, no tangent line to the graph of  $r$  can intersect the  $t$  axis in  $(a, b)$ . Hence

$$-\frac{1}{b-t} \leq \frac{r'(t)}{r(t)} \leq \frac{1}{t-a} \quad \text{for } a < t < b.$$

Therefore, for  $\varepsilon > 0,$

$$(24) \quad \begin{aligned} \int_{a+\varepsilon}^{b-\varepsilon} \frac{r''}{r} dt &= \frac{r'}{r} \Big|_{a+\varepsilon}^{b-\varepsilon} + \int_{a+\varepsilon}^{b-\varepsilon} \left(\frac{r'}{r}\right)^2 dt \\ &\leq \frac{r'}{r} \Big|_{a+\varepsilon}^{b-\varepsilon} + \int_{a+\varepsilon}^{(a+b)/2} \frac{dt}{(t-a)^2} + \int_{(a+b)/2}^{b-\varepsilon} \frac{dt}{(b-t)^2} \\ &= \frac{r'}{r} \Big|_{a+\varepsilon}^{b-\varepsilon} + \frac{2}{\varepsilon} - \frac{4}{b-a}. \end{aligned}$$

Since  $r$  satisfies (11),  $r''(a) = r''(b) = 0$ . So

$$\frac{r'(a+\varepsilon)}{r(a+\varepsilon)} = \frac{1}{\varepsilon} + o(1), \quad \frac{r'(b-\varepsilon)}{r(b-\varepsilon)} = -\frac{1}{\varepsilon} + o(1).$$

Hence, letting  $\varepsilon$  tend to zero in (24) we get

$$(25) \quad \int_a^b \frac{r''}{r} dt \leq -\frac{4}{b-a}.$$

So by (11), (25) and Schwarz's inequality we have

$$(26) \quad \frac{1}{b-a} \left( \int_a^b \|\theta'\| dt \right)^2 \leq \int_a^b \|\theta''\|^2 dt \leq \int_a^b (\theta^T P \theta) dt - \frac{4}{b-a}.$$

Multiplying the inequality (26) by  $(b-a)$ , taking the square root of both sides, and noting  $\theta^T P \theta \leq \lambda_1(t)$ , we obtain Theorem 2.

**4. Concluding remarks.** As we have seen, if the eigenvalues of  $P(t)$  are constant on  $[a, b]$  then the angular variation of all solutions of (1a, b) is bounded by a quantity which is independent of  $P(t)$  and depends only on the dimension of the space. This is no longer

true if we don't require the eigenvalues of  $P(t)$  to be constant as can be shown by an example which is too long to be included.

**Acknowledgment.** I would like to thank the reviewer for simplifying the proof of Lemma 4 and pointing out that Theorem 2 generalized Lyapunov's inequality.

#### REFERENCES

- [1] S. AHMAD AND A. LAZER, *On the components of extremal solutions of second order systems*, this Journal, 8 (1977), pp. 16–23.
- [2] I. GEL'FAND AND S. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [3] M. MORSE, *The foundations of the calculus of variations in the large in  $m$ -space*, Trans. Amer. Math. Soc., 31 (1929), pp. 379–404.
- [4] ———, *A generalization of the Sturm separation and comparison theorems in  $n$ -space*, Math. Ann., 103 (1930), pp. 52–69.
- [5] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
- [6] B. SCHWARZ, *Curves on the unit sphere and disconjugacy of differential systems*, J. Math. Anal. Appl., 39 (1972), pp. 75–86.

## CONSEQUENCES OF ANALYTICITY IN LINEAR ELASTOSTATICS AND RELATED SYSTEMS\*

BRIAN STRAUGHAN†

**Abstract.** Recent results of Oleinik and Radkevich are investigated in the context of linear elastostatics, a mixture of two such solids and a possible model for a fibre reinforced elastic material. The results described establish Liouville theorems, and theorems for uniqueness, analytic continuation and continuous dependence.

**1. Introduction.** In this paper we are concerned with applying some recent results of Oleinik and Radkevich (see Oleinik [21]) to anisotropic linear elastostatics and related systems of equations. These results provide uniqueness theorems, unique continuation theorems and interesting bounds for solutions in unbounded domains. The systems dealt with here are anisotropic, inhomogeneous linear elastostatics, a mixture of two anisotropic linear elastic solids and a theory of fibre reinforced elastic materials. In each case the equations are considered only on all of  $\mathbb{R}^3$ ; while it is possible to consider boundary value problems (see Oleinik [21]) this is not done here as the details would possibly obscure the results which are being described. Indeed, the potential of Oleinik and Radkevich's work on elliptic systems has probably not yet been fully explored in continuum mechanics and new applications are likely to be of interest.

The exposition of Oleinik's [21] results given here is only for elliptic systems of order two and for those defined on  $\mathbb{R}^3$ ; this in no way covers all the situations embraced by the general theory given in [21]. However, we should like to point out that at the heart of all of her results is the following theorem on Banach space valued functions. (This theorem is proved with the aid of Baire's category theorem; see Oleinik and Radkevich [22]–[24].)

Let  $\Omega$  be a domain in  $\mathbb{R}^{n+1}$  and let  $(x_1, x) \in \mathbb{R}^{n+1}$ , i.e.  $(x_1, x) = (x_1, x_2, \dots, x_{n+1})$ . For  $q = 1, 2, \dots, n + 1$  define the differentiation operator  $D_{x_q} = -i\partial/\partial x_q$  and  $D^\alpha = D_{x_2}^{\alpha_2} \dots D_{x_{n+1}}^{\alpha_{n+1}}$  where  $|\alpha| = \alpha_2 + \dots + \alpha_{n+1}$ .

**THEOREM 1.1** (Oleinik and Radkevich [22]–[24].) *Let  $B(\Omega)$  be a Banach space consisting of distributions  $u \in D'(\Omega)$  ( $D'(\Omega)$  is the dual space of  $C_0^\infty(\Omega)$ ) with the norm  $\|u\|_B$  in which the convergence of a sequence in the norm on  $B(\Omega)$  implies its convergence in  $D'(\Omega)$ . Suppose that for a domain  $G$  with  $G \subset \Omega$  and for every  $u \in B(\Omega)$  there exists a constant  $\varepsilon$  depending on  $u$  and a domain*

$$Q_\varepsilon(G) = \{x_1, y_1, x \mid (x_1, x) \in G, |y_1| < \varepsilon\},$$

*such that  $u(x_1, x)$  can be extended into  $Q_\varepsilon(G)$  as an analytic function  $u(x_1 + iy_1, x)$  of  $x_1 + iy_1$ , with  $|D^\alpha u|$  bounded in  $Q_\varepsilon(G)$  for  $|\alpha| \leq k, k \geq 1$ . Then, there exist positive constants  $\delta$  and  $C$  such that for any  $u \in B(\Omega)$  the following estimate is valid:*

$$(1.1) \quad \sup_{Q_\delta(G)} \sum_{|\alpha| \leq k} |D^\alpha u| \leq C \|u\|_{B(\Omega)}.$$

**2. Second order elliptic systems and unique continuation theorems.** We consider only  $\mathbb{R}^3$  and so let  $(x_1, x)$  denote the point in  $\mathbb{R}^3$  given by  $(x_1, x_2, x_3)$ , i.e.  $x \equiv (x_2, x_3)$ . Let

\* Received by the editors December 19, 1978, and in revised form March 26, 1979.

† Department of Mathematics, University of Glasgow, Glasgow G12 8QW, Scotland.

then  $\Omega = \mathbb{R}^2 \times \{|x_1| < A\}$  for some constant  $A$ . In  $\Omega$  we consider systems of the form

$$(2.1) \quad \sum_{j=1}^N \sum_{\alpha_1+|\alpha|\leq m_{kj}} a_{kj}^{\alpha_1\alpha}(x) D_{x_1}^{\alpha_1} D_x^\alpha v_j = F_k, \quad k = 1, \dots, N,$$

where  $D^\alpha = D_{x_2}^{\alpha_2} D_{x_3}^{\alpha_3}$  and  $|\alpha| = \alpha_2 + \alpha_3$ , with  $a_{kj}^{\alpha_1\alpha}$  analytic functions of  $x^1$ .

System (2.1) is supposed uniformly elliptic (see Agmon, Douglis and Nirenberg [1], Oleinik [21]) and so attached to the system are integers  $s_1, \dots, s_N, t_1, \dots, t_N$ , corresponding to the equations and unknowns, respectively. The coefficients  $m_{kj}$  are such that  $m_{kj} \leq s_k + t_j$  and in this work we only consider  $s_k + t_j \leq 2$ .<sup>2</sup> If  $s_k + t_j < 0$  then  $m_{kj} = 0$ , and the  $s_k$  are chosen so that  $s_k \leq 0$ . Moreover, an integer  $m$  is defined by  $m = \frac{1}{2} \sum_{j=1}^N (s_j + t_j)$ . The equations (2.1) form a uniformly elliptic system if there is a constant  $\lambda$  such that

$$(2.2) \quad \begin{aligned} \lambda (|\xi_1|^2 + |\xi|^2)^m &\leq |\det p_{kj}(x, \xi_1, \xi)| \\ &\leq \lambda^{-1} (|\xi_1|^2 + |\xi|^2)^m, \end{aligned}$$

where

$$(2.3) \quad p_{kj}(x, \xi_1, \xi) = \sum_{\alpha_1+|\alpha|=s_k+t_j} a_{kj}^{\alpha_1\alpha}(x) \xi_1^{\alpha_1} \xi^\alpha,$$

and where  $\xi^\alpha = \xi_2^{\alpha_2} \xi_3^{\alpha_3}$ .

For a set  $Q$  in  $\mathbb{R}^q$  the usual norm on  $C^k(Q)$  is defined, i.e. if  $f \in C^k(Q)$ ,

$$(2.4) \quad \|f\|_Q^k = \sup_Q \sum_{|\beta|\leq k} |D^\beta f| < \infty.$$

The following theorems form the basis of the remainder of the work. We emphasize, however, that they are only special cases of the general theory developed by Oleinik and Radkevich.

**THEOREM 2.1** (see Oleinik [21]). *Let  $v = (v_1, \dots, v_N)$  be a solution to system (2.1) in  $\Omega$  with  $F_k \equiv 0, k = 1, \dots, N$ , and suppose  $v_j \in C^{t_j+1}(\Omega)$ . Suppose that  $\|a_{kj}^{\alpha_1\alpha}\|_{\mathbb{R}^2}^{1-s_j} \leq M$  for some constant  $M$  and  $\alpha_1+|\alpha|\leq m_{kj}, k, j = 1, \dots, N$ .*

*Let  $\Omega_1 = \omega_1 \times \{|x_1| < A - 2\}$ ,  $\omega_1 \subset \omega$ . Then, the function  $v_j$  with all derivatives up to order  $t_{j-1}$  can be extended into the domain  $Q_\delta(\Omega_1) = \{x_1, y_1, x | (x_1, x) \in \Omega_1, |y_1| < \delta\}$ , as an analytic function of  $x_1 + iy_1$  and for  $|\alpha| \leq t_j - 1$  the following inequality is valid:*

$$(2.5) \quad \sup_{Q_\delta(\Omega_1)} |D_x^\alpha v_j| \leq C \left( \sum_{k=1}^N \sup_\Omega |v_k| \right), \quad j = 1, \dots, N,$$

where  $\delta$  and  $C$  are constants depending on  $N, m$  and the constant  $\lambda$  in (2.2).

The next result is a similar analytic continuation theorem for solutions to the inhomogeneous version of (2.1).

**THEOREM 2.2** (see Oleinik [21]). *Let  $\Omega, \Omega_1$  and the bounds on  $a_{kj}^{\alpha_1\alpha}$  be as in Theorem 2.1. Suppose the functions  $F_k(x_1, x), k = 1, \dots, N$ , and their derivatives up to order  $1 - s_k$  may be extended into a domain  $Q_{\delta_0}(\Omega) = \{x_1, y_1, x | (x_1, x) \in \Omega, |y_1| < \delta_0\}$  as analytic functions of  $x_1 + iy_1$ . Suppose also that  $v_j \in C^{t_j+1}(\Omega)$  is a solution to (2.1) in  $\Omega$ . Then,  $v_j$  can be extended into the domain  $Q_\delta(\Omega_1) = \{x_1, y_1, x | (x_1, x) \in \Omega_1, |y_1| < \delta\}$  with*

<sup>1</sup> A function  $\psi(x)$  defined on  $\mathbb{R}^m$  is said to be analytic at a (real) point  $x_0 \in \mathbb{R}^m$  if it is representable by an absolutely convergent power series in the variables  $x^j - x_0^j$  in a real neighborhood of  $x_0$ . This implies  $\psi$  can be defined as an analytic function in a complex neighborhood of  $x_0$ , see John [15, p. 52].

<sup>2</sup> This is sufficient to include the examples in §§ 3-5. For some systems in continuum mechanics, e.g. multipolar elasticity (see Green and Rivlin [11]) it may be desirable to remove this restriction.

derivatives of  $v_j$  up to order  $t_j - 1$  as analytic functions of  $x_1 + iy_1$  and for  $|\alpha| \leq t_{j-1}$  and constants  $\delta, C$  depending on  $m, N$  and  $\lambda$ , we have

$$(2.6) \quad \|D^\alpha v_j\|_{O_\delta(\Omega_1)}^0 \leq C \left\{ \sum_{k=1}^N \|v_k\|_{\Omega}^0 + \sum_{m=1}^N \|F_m\|_{O_{\delta_0}^s(\Omega)}^{1-s} \right\}.$$

Both of the above theorems are proved using the a priori estimates for elliptic systems (see Agmon, Douglis and Nirenberg [1]) and the Morrey–Nirenberg [20] method for establishing analyticity in elliptic systems (see Oleinik [21]). As our main interest is in the application of these theorems we do not include the rather technical proofs.

**3. Anisotropic linear elasticity.** The equations governing the displacement field of an anisotropic linear elastic body in equilibrium are

$$(3.1) \quad (a_{pikh}(x)u_{k,h})_{,j} + \rho f_p = 0,$$

where  $u_p$  are the components of displacement,  $\mathbf{f}$  is the body force,  $\rho$  the density,  $a_{pikh}$  are the elasticities and standard indicial notation is assumed. It is supposed  $a_{pikh}$  are real analytic functions of  $x = (x_2, x_3)$  and the equations are here taken to be defined on  $\mathbb{R}^3$ .

A thorough study of uniqueness for various boundary value problems for (3.1) is given by Knops and Payne [17] and a similar account of existence theory may be found in the works of Fichera [4], [5]. The above system may be regarded as an elliptic system in the sense of Agmon, Douglis and Nirenberg and so the results of [1], [25] are applicable. In this work, however, we are concerned directly only with Oleinik’s results; these do incidentally have implications for the uniqueness question. Application of the generalized ellipticity concept of [1], [25] in the context of elasticity theory was discussed by Hayes and Horgan [26].

To consider (3.1) as an elliptic system in the sense of Agmon, Douglis and Nirenberg [1], [25] (see § 2) we may take  $s_k = 0, t_k = 2, k = 1, 2, 3, m = 3$  and then the ellipticity condition is

$$(3.2) \quad |\det a_{pikh}(x)\xi_j\xi_h| \neq 0,$$

$\forall \xi \neq \mathbf{0}$  (see e.g. Knops and Payne [17, p. 20]). Moreover, the uniform ellipticity condition (2.2) is

$$(3.3) \quad \lambda (|\xi_1|^2 + |\xi|^2)^3 \leq |\det a_{pikh}\xi_j\xi_h| \leq \lambda^{-1} (|\xi_1|^2 + |\xi|^2)^3,$$

for some constant  $\lambda$  independent of  $x (\in \mathbb{R}^2)$ .

Oleinik’s work [21] essentially selects one direction, taken here as the  $x_1$  direction, and employs analyticity by extending  $u_p$  to be an analytic function of the variable  $x_1 + iy_1, x_1, y_1 \in \mathbb{R}$ . In particular she seeks solutions which for (3.1) may be represented as

$$(3.4) \quad u_p(x_1, x) = e^{i\mu x_1} U_p(x; \mu),$$

where  $x = (x_2, x_3)$  and  $\mu = \mu_1 + i\mu_2, \mu_1, \mu_2 \in \mathbb{R}$ .

We shall suppose the elasticities satisfy (3.3) and deal with solutions of the form (3.4). The following theorems employ estimates (2.5) and (2.6) to obtain results for  $U_p$  and consequently for  $u_p$ . The proofs of these results are given by Oleinik [21] in the general linear elliptic setting; however, brief details are included for completeness, for the elastic case.

The first result concerns the dependence of  $u_p$  on the real part of the complex amplitude  $\mu$ .

**THEOREM 3.0.** *Let  $\Omega, \Omega_1, \omega, \omega_1$  and  $Q_\delta(\Omega_1)$  be as defined in § 2 and consider a solution to the homogeneous system for (3.1). From (2.5) we find that if  $a_{pjkh}$  and their first and second derivatives are uniformly bounded on  $\mathbb{R}^2$  then*

$$(3.5) \quad \sum_{p=1}^3 \sup_{\omega_1} |U_p(x; \mu)| \leq K \exp \{-\delta|\mu_1| + 2|\mu_2|\} \left( \sum_{q=1}^3 \sup_{\omega} |U_q| \right),$$

for a constant  $K$ .

The above result is analogous to Oleinik [21, Thm. 12]. From this inequality we see that solutions to (3.1) of the form (3.4) which are bounded in  $\omega$  decrease exponentially in  $\omega_1$  with increasing  $\mu_1$ .

The next application of Theorem 2.1 gives a type of Liouville theorem for solutions to (3.1) in  $\mathbb{R}^3$ . Another Liouville theorem for elliptic systems is given in Agmon, Douglis and Nirenberg [1, p. 70]. Although it is well known that Liouville theorems may be obtained for elliptic equations via Hopf’s maximum principles, I do not know if this approach is applicable to systems.

**THEOREM 3.1.** *Let  $u_p \in C^3(\Omega)$  be a solution to (3.1) of the form (3.4) with  $\|a_{pjkh}\|_{\mathbb{R}^2} \leq M < \infty$  and with  $f_m \equiv 0$  in the domain  $\Omega = \mathbb{R}^3$ . Suppose there is a constant  $\delta_1 (> 0)$  such that in  $\omega = \mathbb{R}^2$*

$$(3.6) \quad \sum_1^3 |U_m(x; \mu)| \leq \exp \{\delta_1|x|\}.$$

Then, if  $-\delta|\mu_1| + 2|\mu_2| + 2\delta_1 + \log C < 0$  where  $\delta$  and  $C$  are given constants,  $u_p \equiv 0$  in  $\Omega$ .

*Proof* (cf. Oleinik [21, Thm. 4]). Let  $\omega_R = \{x \in \mathbb{R}^2 \mid |x| < R\}$ , for some  $R > 0$ . Applying Theorem 2.1, (2.5) becomes

$$\sup_{Q_\delta(\Omega_1)} |u_p| \leq C \left( \sum_1^3 \sup_{\Omega} |u_m| \right), \quad p = 1, 2, 3,$$

for constants  $\delta, C$ . Set  $\omega = \omega_{R+2}, \omega_1 = \omega_R, \Omega = \omega_{R+2} \times \{|x_1| < R+2\}$  and  $\Omega_1 = \omega_R \times \{|x_1| < R\}$ . Hence, we deduce that

$$\sum_1^3 \sup_{\omega_R} |U_j| \leq \exp(\log C' - \delta|\mu_1| + 2|\mu_2|) \left( \sum_1^3 \sup_{\omega_{R+2}} |U_k| \right),$$

where  $C' = 4C$ . Now,  $\delta$  and  $C'$  do not depend on  $R$ , and so we may use a “bootstrap” argument to extend the above inequality to  $\omega_{R+2M}$  for  $M \in \mathbb{Z}, M > 1$ , to see that

$$(3.7) \quad \sum_1^3 \sup_{\omega_R} |U_j| \leq \exp \{M(\log C' - \delta|\mu_1| + 2|\mu_2|)\} \left( \sum_1^3 \sup_{\omega_{R+2M}} |U_k| \right).$$

Finally, let us invoke hypothesis (3.6) on the right hand side to obtain

$$(3.8) \quad \sum_1^3 \sup_{\omega_R} |U_j| \leq \exp \{M(\log C' - \delta|\mu_1| + 2|\mu_2| + 2\delta_1) + \delta_1 R\}.$$

The coefficient  $\log C' - \delta|\mu_1| + 2|\mu_2| + 2\delta_1 < 0$  by hypothesis (taking  $C' = C$  in the statement of the theorem) and so we may let  $M \rightarrow \infty$  in (3.8) to see that  $U_k \equiv 0$  in  $\omega_R$ . However,  $R$  is arbitrary and so  $U_k \equiv 0$  in  $\omega$ . The theorem follows.

The next theorem gives a decay result in a neighborhood of infinity.

**THEOREM 3.2.** *Let  $u_p \in C^3(\Omega)$  be a solution to (3.1) of the form (3.4) with  $\|a_{pjkh}\|_{\mathbb{R}^2} \leq M < \infty$  and with  $f_k \equiv 0$ . Let  $\delta, \delta_1$  and  $C'$  be as in Theorem 3.1, let  $R$  be a*



positive constant and suppose (3.6) holds in  $\{|x| > R\}$ . If there is a positive constant  $\delta_2$  such that

$$(3.9) \quad 4\delta_1 + \delta_2 + \log C' + 2|\mu_2| - \delta|\mu_1| < 0,$$

then

$$(3.10) \quad \sum_1^3 |U_k(x; \mu)| \leq C'' \exp(-\delta_2|x|),$$

for  $(x_2, x_3) \in \{|x| > R\}$  and some constant  $C''$ .

*Proof* (cf. Oleinik [21, Theorem 7]). The proof follows similar lines to that of the last theorem. However,  $\omega_R$  is replaced by  $S_k^x \cap \omega$  where  $\omega = \{|x| > R\}$ ,  $S_k^x =$  open ball in  $\mathbb{R}^2$ , center  $x$ , radius  $k$ . With  $p(x) = |x| - R$  and  $p_e(x) =$  greatest integer  $\leq p(x)$ , hypothesis (3.6) leads to

$$(3.11) \quad \begin{aligned} & \sum_1^3 \sup_{S_k^x \cap \omega} |U_j| \\ & \leq \exp[(4+R)\delta_1] \cdot \exp\left\{\frac{1}{2}(p(x)-1)(4\delta_1+2|\mu_2| \right. \\ & \quad \left. + \log C' - \delta|\mu_1|) - \frac{1}{2}(p(x)-1-p_e(x)) \right. \\ & \quad \left. \cdot (\log C' + 2|\mu_2| - \delta|\mu_1|)\right\}. \end{aligned}$$

(3.10) follows directly from (3.11) by use of the arbitrariness of  $x$ .

We consider next the inhomogeneous problem for (3.1). The result described may be viewed as a continuous dependence theorem on the values of the body force  $f_j$ .

**THEOREM 3.3.** *Let  $u_k \in C^3(\Omega)$  be a solution to (3.1) with  $\|a_{pjkh}\|_{\mathbb{R}^2}^1 \leq M < \infty$  in  $\Omega \times \mathbb{R}$ . Suppose that for any  $R > 1$ ,*

$$\sum_{j=1}^3 \|H_j\|_{\omega_R}^1 \leq C_1 \exp(\delta_3 R),$$

for some constants  $C_1, \delta_3(>0)$ , where  $\omega_R = \{|x| < R\}$ , and  $\rho f_p = e^{i\mu x_1} H_p$ . If there are positive constants  $\delta, \delta_1, C'$  such that

$$\delta|\mu_1| > \log C' + \mu_2 + \delta_\beta; \quad \beta = 1, 3,$$

and if in  $\omega$

$$\sum_1^3 |U_k(x)| \leq C_2 \exp(\delta_1|x|),$$

then

$$\sum_1^3 \sup_{\omega_R} |U_j| \leq C_3 \exp(\delta_3 R).$$

The proof of this theorem is very similar to the proof of Theorems 3.1 and 3.2, except Theorem 2.2 is used in place of Theorem 2.1 (cf. Oleinik [21, Thm. 8]).

It is worth observing that Theorem 3.1 gives a uniqueness theorem for solutions to (3.1) which have form (3.4) and which are bounded exponentially as in (3.6). Moreover, by a modification of the proof of Theorem 3.1 we may rederive a proof of uniqueness in the dynamic linear elastic problem for negative definite elasticities, a result first proved by Hayes and Knops [14] (see also Knops and Payne [17, § 8.2]). Nevertheless, for the dynamic problem uniqueness holds without any definiteness as shown by Knops and Payne [16].

**4. A mixture of two anisotropic linear elastic solids.** This section considers theorems equivalent to those of § 3 for a mixture of two elastic solids. The theory we employ was developed by Green and Steel [12] and explicit forms for the linear theory were presented by Knops and Steel [18].

Let us consider a mixture of two anisotropic linear elastic solids occupying  $\mathbb{R}^3$ . Then, employing the notation of [18], the components of displacement of each solid with respect to a reference configuration are denoted by  $\omega_p$  and  $\eta_p$  and in terms of these components the equations of equilibrium in the mixture are

$$(4.1) \quad \begin{aligned} \sigma_{kp,k} - \pi_p + \rho_1 F_p &= 0, \\ \pi_{kp,k} + \pi_p + \rho_2 G_p &= 0, \end{aligned}$$

where  $\sigma_{kp}$  and  $\pi_{kp}$  are the partial stresses in each solid,  $\pi_p$  is a diffusive force,  $\rho_1$  and  $\rho_2$  are the densities of each solid, and  $F_p$  and  $G_p$  denote the body force per unit mass of each solid.

The constitutive equations are

$$(4.2) \quad \sigma_{kp} = a_{kp} + b_{kp} + A_{kprs}\omega_{r,s} + C_{kprs}\eta_{r,s}$$

$$(4.3) \quad \pi_{kp} = b_{pk} + D_{kprs}\omega_{r,s} + B_{kprs}\eta_{r,s}$$

$$(4.4) \quad \pi_p = \frac{\rho_1}{\rho} b_{rs}\eta_{r,sp} - \frac{\rho_2}{\rho} (a_{sr} + b_{sr})\omega_{r,sp}$$

In these equations  $a_{kp} + b_{kp}$  and  $b_{pk}$  are the respective initial stresses in each solid and the coefficients  $A_{kprs}$ ,  $B_{kprs}$ ,  $C_{kprs}$  and  $D_{kprs}$  are given explicitly in terms of the coefficients of the free energy for the mixture in equations (3.3) of [18]. We shall assume the coefficients in (4.2)–(4.4) depend only on  $x = (x_2, x_3)$ , are real analytic, and together with their first derivatives are uniformly bounded on  $\mathbb{R}^2$ .

It appears necessary with the approach adopted here to additionally assume that

$$(4.5) \quad (a_{kp} + b_{kp})_{,k} = 0, \quad b_{pk,k} = 0.$$

These conditions are obviously satisfied for an initially unstressed mixture.

To write (4.1) as an elliptic system in the sense of Agmon, Douglis and Nirenberg [1], set  $s_k = 0$ ,  $t_k = 2$ ,  $k = 1, \dots, 6$ , and then  $m = 6$ . The uniform ellipticity condition (2.2) is

$$(4.6) \quad \lambda^{-1} (|\xi_1|^2 + |\xi|^2)^6 \geq |\det l_{JK}| \geq \lambda (|\xi_1|^2 + |\xi|^2)^6,$$

for each real  $(\xi_1, \xi_2, \xi_3) \neq 0$  and the  $6 \times 6$  matrix  $l_{JK}$  is given by

$$(4.7) \quad l_{JK} = \begin{pmatrix} \Lambda_{11}(pr) & \Lambda_{12}(pr) \\ \Lambda_{21}(pr) & \Lambda_{22}(pr) \end{pmatrix}$$

where the  $\Lambda$ 's are the following  $3 \times 3$  matrices

$$(4.8) \quad \begin{aligned} \Lambda_{11} &= A_{kprs}\xi_s\xi_k + \frac{\rho_2}{\rho} (a_{sr} + b_{sr})\xi_s\xi_p, \\ \Lambda_{12} &= C_{kprs}\xi_s\xi_k - \frac{\rho_1}{\rho} b_{rs}\xi_s\xi_p, \\ \Lambda_{21} &= D_{kprs}\xi_s\xi_k - \frac{\rho_2}{\rho} (a_{sr} + b_{sr})\xi_s\xi_p, \\ \Lambda_{22} &= B_{kprs}\xi_s\xi_k + \frac{\rho_1}{\rho} b_{rs}\xi_s\xi_p. \end{aligned}$$

For the mixture problem the solution corresponding to (3.4) is sought as

$$(4.9) \quad \begin{aligned} \omega_p(x_1, x) &= e^{i\mu x_1} \Omega_p(x; \mu), \\ \eta_p(x_1, x) &= e^{i\mu x_1} Y_p(x; \mu). \end{aligned}$$

Adapting the methods outlined in § 3, we find the analogue of (3.5) to be

$$(4.10) \quad \sum_{p=1}^3 \left\{ \sup_{\omega_1} |\Omega_p| + \sup_{\omega_1} |Y_p| \right\} \leq K \exp(-\delta|\mu_1| + 2|\mu_2|) \sum_{q=1}^3 \left\{ \sup_{\omega} |\Omega_q| + \sup_{\omega} |Y_q| \right\},$$

if  $\bar{\omega}_1 \subset \omega$  and  $F_k \equiv G_k \equiv 0$ .

The corresponding Liouville theorem (Theorem 3.1) shows that if a solution to (4.1) of the form (4.9) in  $\mathbb{R}^3$ , with  $F_k \equiv G_k \equiv 0$ , is such that in  $\mathbb{R}^2$

$$(4.11) \quad \sum_{p=1}^3 (|\Omega_p| + |Y_p|) \leq \exp\{\delta_1|x|\},$$

and  $-\delta|\mu_1| + 2|\mu_2| + 2\delta_1 + \log C < 0$ , then  $\omega_p \equiv \eta_p \equiv 0$  in  $\mathbb{R}^3$ .

If the solution (4.9) with  $F_k \equiv G_k \equiv 0$  is such that (3.9) holds and condition (4.11) is satisfied for  $|x| > R$ , then in  $\{x \in \mathbb{R}^2 \mid |x| > R\}$ ,

$$(4.12) \quad \sum_1^3 (|\Omega_p| + |Y_p|) \leq \exp(-\delta_2|x|),$$

for a positive constant  $\delta_2$ .

Finally, there is the result analogous to Theorem 3.3 for the inhomogeneous problem for (4.1) which shows that if the body forces may be written as  $\rho_1 F_p = e^{i\mu x_1} \mathcal{F}_p$  and  $\rho_2 G_p = e^{i\mu x_1} \mathcal{G}_p$  with  $\sum_1^3 (\|\mathcal{F}_p\|_{\omega_R}^1 + \|\mathcal{G}_p\|_{\omega_R}^1) \leq C_1 \exp(\delta_3 R)$ , and there is a solution of the form (4.9) with

$$\delta|\mu_1| > \log C' + \mu_2 + \delta_\beta, \quad \beta = 1, 3,$$

such that in  $\omega$

$$\sum_1^3 (|\Omega_p| + |Y_p|) \leq C_2 \exp(\delta_1|x|),$$

then

$$(4.13) \quad \sum_1^3 \sup_{\omega_R} (|\Omega_p| + |Y_p|) \leq C_3 \exp(\delta_3 R),$$

where the constants  $\delta, \delta_1, \delta_3$  etc. are as in Theorem 3.3.

**5. An inextensible linear elastic material.** We shall consider a linear elastic solid in  $\mathbb{R}^3$ , but one which is inextensible in the  $x_3$ -direction. In this case (see Hayes and Horgan [13]) there is introduced a  $C^1$  function  $\phi(x)$  which is associated with an arbitrary tension in the  $x_3$ -direction. For the class of problems to be studied here it is sufficient to consider the third component of displacement,  $u_3$  to be identically zero, see [13]. The relevant equations are then

$$(5.1) \quad \begin{aligned} (c_{\alpha j \beta l} u_{\beta, l})_{, j} &= 0 \\ (c_{3 j \beta l} u_{\beta, l})_{, j} - (c_{33 \beta s} u_{\beta, s})_{, 3} + \phi_{, 3} &= 0 \end{aligned}$$

where Greek indices take values 1, 2, while Latin indices as before assume the values 1 to 3. The elastic coefficients  $c_{pqrs}$  are assumed constant as in [13], although this is not necessary for the following results to remain valid.

Equations (5.1) form an elliptic system in the sense of Agmon, Douglis and Nirenberg [1] in the variables  $(u_1, u_2, \phi_{,3})$ , if we choose  $t_1 = t_2 = 2, s_1 = s_2 = 0, t_3 = 0, s_3 = 0$ , and so  $m = 2$ . The uniform ellipticity condition may be verified to be

$$(5.2) \quad \lambda^{-1}(|\xi_1|^2 + |\xi|^2)^2 \cong \frac{1}{2} \varepsilon_{\alpha\delta} \varepsilon_{\beta\gamma} c_{\alpha j \beta i} c_{\delta p \gamma m} \xi_j \xi_i \xi_p \xi_m | \cong \lambda (|\xi_1|^2 + |\xi|^2)^2,$$

for every real  $\xi \neq 0$ . (Let us observe that (5.1) differs from the elliptic systems of §§ 3 and 4 in the selection of the integers  $t_p, s_p$ .)

In a manner analogous to (3.4) solutions to (5.1) are sought of the form

$$(5.3) \quad u_\alpha(x_1, x) = e^{i\mu x_1} U_\alpha(x), \quad \phi_{,3}(x_1, x) = e^{i\mu x_1} \Phi(x).$$

Results analogous to inequality (3.5) and Theorems 3.1 and 3.2 may now be established. The left hand side of (3.5) is replaced by  $\sup_{\omega_1} |U_1(x)| + \sup_{\omega_1} |U_2(x)| + \sup_{\omega_1} |\Phi(x)|$ , whereas the left hand sides of (3.6) and (3.10) become  $|U_1(x)| + |U_2(x)| + |\Phi(x)|$ . Furthermore, we may establish a result analogous to Theorem 3.3 for the inhomogeneous version of (5.1) for an appropriate body force.

**6. Concluding remarks.** Implications for uniqueness for the equations of §§ 4 and 5 follow in the same manner as the corresponding ones in § 3.

While we have only dealt with the equations for three types of elastic materials, it should be possible to adapt Oleinik's work to several other systems. In particular, we mention the multipolar theories of Green and Rivlin [11] and the associated dislocation theories of Fox [6], the rod, shell and plate theories of Green, Naghdi, Laws and Wenner, see e.g. [7–10], the micropolar theory of Eringen [3] (see also Knops and Straughan [19]), higher order mixture theories, see e.g. Atkin and Craine [2], and thermoelasticity. However, as noted earlier, for some of these theories the case of second order elliptic systems discussed in § 2 is insufficient and the more general theory given by Oleinik [21] is necessary.

**Acknowledgment.** I should like to thank a referee for helpful comments regarding presentation and for bringing references [25] and [26] to my attention.

REFERENCES

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II*, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.  
 [2] R. J. ATKIN AND R. E. CRAINE, *Continuum theories of mixtures: applications*, J. Inst. Math. Appl., 17 (1976), pp. 153–207.  
 [3] A. C. ERINGEN, *Linear theory of micropolar elasticity*, J. Math. Mech., 15 (1966), pp. 909–923.  
 [4] G. FICHERA, *Linear elliptic differential systems and eigenvalue problems*, Lecture notes in Mathematics, vol. 8, Springer-Verlag, Berlin, 1965.  
 [5] ———, *Existence theorems in elasticity*, Handbuch der Physik, vol. VIa/2, C. Truesdell, ed., Springer-Verlag, Berlin, 1972.  
 [6] N. FOX, *A continuum theory of dislocations for polar elastic materials*, Quart. J. Mech. Appl. Math., 19 (1966), pp. 343–355.  
 [7] A. E. GREEN, N. LAWS AND P. M. NAGHDI, *Rods, plates and shells*, Proc. Cambridge Philos. Soc., 64 (1968), pp. 895–913.  
 [8] A. E. GREEN AND P. M. NAGHDI, *Shells in the light of generalised continua*, IUTAM Symposium Copenhagen 1967 on Theory of thin shells, Springer-Verlag, Berlin, 1969.  
 [9] ———, *On uniqueness in the linear theory of elastic shells and plates*, J. Mécanique, 10 (1971), pp. 251–261.

- [10] A. E. GREEN, P. M. NAGHDI AND M. L. WENNER, *Linear theory of Cosserat surface and elastic plates of variable thickness*, Proc. Cambridge Philos. Soc., 69 (1971), pp. 227–254.
- [11] A. E. GREEN AND R. S. RIVLIN, *Simple force and stress multipoles*, Arch. Rational Mech. Anal., 16 (1964), pp. 325–353.
- [12] A. E. GREEN AND T. R. STEEL, *Constitutive equations for interacting continua*, International J. Engrg. Sci., 4 (1966), pp. 483–500.
- [13] M. HAYES AND C. O. HORGAN, *On the displacement boundary-value problem for inextensible elastic materials*, Quart. J. Mech. Appl. Math., 27 (1974), pp. 287–297.
- [14] M. HAYES AND R. J. KNOPS, *On the displacement boundary value problem of linear elastodynamics*, Quart. Appl. Math., 26 (1968), pp. 291–293.
- [15] F. JOHN, *Plane Waves and Spherical Means Applied to Partial Differential Equations*, Interscience, New York, 1955.
- [16] R. J. KNOPS AND L. E. PAYNE, *Uniqueness in classical elastodynamics*, Arch. Rational Mech. Anal., 27 (1968), pp. 349–355.
- [17] ———, *Uniqueness Theorems in Linear Elasticity*, Springer Tracts in Natural Philosophy, vol. 19, Springer-Verlag, Berlin, 1971.
- [18] R. J. KNOPS AND T. R. STEEL, *On the stability of a mixture of two elastic solids*, J. Comp. Mats., 3 (1969), pp. 652–663.
- [19] R. J. KNOPS AND B. STRAUGHAN, *Continuous dependence theorems in the theory of linear elastic materials with microstructure*, Internat. J. Engrg. Sci., 14 (1976), pp. 555–565.
- [20] C. MORREY AND L. NIRENBERG, *On the analyticity of solutions of linear elliptic systems of partial differential equations*, Comm. Pure Appl. Math., 10 (1957), pp. 271–290.
- [21] O. A. OLEINIK, *The analyticity of solutions of partial differential equations and its applications*, Trends in Applications of Pure Mathematics to Mechanics, Pitman, London, 1976.
- [22] O. A. OLEINIK AND E. B. RADKEYVICH, *On eigenfunctions and the behaviour of solutions of systems of partial differential equations depending on a parameter*, Rev. Roum. Math. Pure Appl., 19 (1974), pp. 47–54.
- [23] ———, *On the analyticity of solutions of linear differential equations and systems*, Dokl. Akad. Nauk. SSSR., 207 (1972), pp. 785–788.
- [24] ———, *On analyticity of solutions of linear partial differential equations*, Mat. Sb., 90 (1973), pp. 592–607.
- [25] A. DOUGLIS AND L. NIRENBERG, *Interior estimates for elliptic systems of partial differential equations*, Comm. Pure Appl. Math., 8 (1955), pp. 503–538.
- [26] M. HAYES AND C. O. HORGAN, *On the Dirichlet problem for incompressible elastic materials*, J. Elasticity, 4 (1974), pp. 17–25.

## INTEGRAL AVERAGES AND THE OSCILLATION OF SECOND ORDER ORDINARY DIFFERENTIAL EQUATIONS\*

G. J. BUTLER†

**Abstract.** Some of the more important and useful tests for the oscillation of the second order scalar linear differential equation  $y'' + qy = 0$  are given by the classical Fite–Wintner theorem and its generalizations by Wintner and by Hartman. These tests involve the behavior of the integral of  $q$  or, more generally, the average behavior of the integral.

Several years ago, Waltman extended the Fite–Wintner theorem to nonlinear equations. We show that the Wintner and Hartman theorems also extend to a large class of nonlinear equations which includes the Emden–Fowler equation. Further generalizations of the averaging technique for the linear equation due to Coles and to Willett are also shown to extend to some degree to nonlinear equations.

**1. Introduction.** Consider the second order scalar linear ordinary differential equation

$$(L) \quad y''(t) + q(t)y(t) = 0, \quad t \in I = [T, \infty).$$

In the study of this equation from the point of view of disconjugacy on  $I$ , many criteria for oscillation have been found which involve the behavior of the integral of  $q$ . Three of the more important such conditions which guarantee that all solutions of (L) oscillate on  $I$  are the following:

$$(A1) \quad \int_T^\infty q(s) ds = \infty \quad (\text{Fite [9], Wintner [18]}).$$

$$(A2) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_T^t \int_T^s q(\tau) d\tau ds = \infty \quad (\text{Wintner [18]}).$$

$$(A3) \quad \begin{aligned} & -\infty < \overline{\lim}_{t \rightarrow \infty} \frac{1}{t} \int_T^t \int_T^s q(\tau) d\tau ds \\ & < \overline{\lim}_{t \rightarrow \infty} \frac{1}{t} \int_T^t \int_T^s q(\tau) d\tau ds \leq \infty \quad (\text{Hartman [10]}). \end{aligned}$$

Several years ago, Coles [7] and Willett [17] extended these criteria by considering weighted averages of the integral of  $q$  of the form

$$A_\phi(t, T) = \frac{\int_T^t \phi(s) (\int_T^s q(\tau) d\tau) ds}{\int_T^t \phi(s) ds}.$$

Thus Willett [17] showed that there is a class  $\Phi_0$  of nonnegative locally integrable, but not integrable, functions, which contains all such bounded functions, such that if for some  $\phi \in \Phi_0$ , we have

$$(A4) \quad -\infty < \overline{\lim}_{t \rightarrow \infty} A_\phi(t) < \overline{\lim}_{t \rightarrow \infty} A_\phi(t) \leq \infty \quad \text{or} \quad \lim_{t \rightarrow \infty} A_\phi(t) = \infty,$$

then all solutions of (L) oscillate on  $I$ .

Willett's result is actually stronger than that stated, but in this form it is clearly seen to be an extension of the criteria (A2) and (A3), which together correspond to (A4) with  $\phi \equiv 1$ .

\* Received by the editors June 9, 1978, and in revised form April 16, 1979.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

In this paper we are concerned with the possibility of averaging techniques for studying the oscillatory behavior of nonlinear equations of the form

$$(N) \quad y''(t) + q(t)f(y(t)) = 0$$

with a particular interest when  $f(y) = |y|^\alpha \operatorname{sgn} y$ ,  $\alpha > 0$ , in which case we shall denote the above equation by  $(N_\alpha)$ .

In [16], Waltman showed that the Fite–Wintner condition (A1) is an oscillation criterion for  $(N_{2n+1})$ ,  $n$  a natural number. Indeed his method of proof extends to the case of any monotone, nondecreasing, continuously differentiable function  $f$  satisfying  $yf(y) > 0$  for  $y \neq 0$ .

Our main result is that for a certain class  $\mathcal{F}$  of functions  $f$ , both the Wintner condition (A2) and the Hartman condition (A3) serve as oscillation criteria for (N). The class of such equations handled in this way includes all equations of the form  $(N_\alpha)$ . In the case  $\alpha > 1$ , we answer a question raised by Wong in [20], [21]. When  $0 < \alpha < 1$ , we can obtain a stronger result than the oscillation criteria (A2) and (A3) and hence we are able to improve a result of Kamenev [11].

As far as applying general averaging conditions of the type (A4) to (N) is concerned, we have had only limited success; however we are able to obtain refinements to conditions (A2) and (A3) that enable us to resolve the oscillatory nature of  $(N_\alpha)$  when  $q(t) = t^\lambda p(t)$ , where  $p$  is a nonconstant periodic function of mean zero; for all values of  $\lambda$ , when  $\alpha > 1$ , and for all  $\lambda \geq 1$ , when  $0 < \alpha < 1$ . This answers a question raised in [21].

When  $f'$  is bounded away from zero, many of the arguments used for handling (L) go through for (N) with only minor changes and one can obtain averaging criteria (A4) for (N) (see also [12]).

Before proceeding to a description of the class  $\mathcal{F}$  and a precise statement of our main results, we make a few preliminary remarks.

Throughout we make the underlying assumption that  $q$  and  $f$  are continuous on the real line and that  $f$  is continuously differentiable, except possibly at 0, and satisfies  $f'(y) \geq 0$ ,  $yf(y) > 0$  for  $y \neq 0$  (which implies that  $f(0) = 0$ ). We shall denote the set of all such functions  $f$  by  $\mathcal{C}$ .

By a solution of (N) we shall always mean a nontrivial solution defined on some half-line  $I = [T, \infty)$ . Without a sign restriction on  $q$ , there may exist noncontinuable solutions of (N) [3]; however there will always exist infinitely many continuable solutions under rather mild additional conditions on  $q$  (see [6]). A solution of (N) oscillates if it has infinitely many zeros on  $I$ ; the conditions on  $q$  and  $f$  guarantee that these zeros can only cluster at  $\infty$ . If all solutions of (N) oscillate, (N) is said to be oscillatory.

We make the following imprecise, but we hope helpful, remark concerning the study of oscillation of (N); roughly speaking there are, as in the case of (L), two main conditions under which (N) is oscillatory. The first is that  $q$  is, in some sense, “sufficiently positive”; we refer to [1], [2], [4], [8], [11]–[16], [19]–[21] for results of this nature as well as condition (A2) of this paper. The second condition is that  $q$  is “sufficiently oscillatory” in its behavior, and this is the idea behind condition (A3) and its generalization in condition (A4).

**2. The class  $\mathcal{F}$  and statement of results.** For  $f \in \mathcal{C}$ , we define  $\Omega(x) = \Omega_f(x)$  on  $(-\infty, 0) \cup (0, \infty)$  by

$$(D1) \quad \Omega(x) = \begin{cases} \int_x^1 du/f(u), & x > 0, \\ \int_x^{-1} du/f(u), & x < 0. \end{cases}$$

$\Omega$  is monotone decreasing from  $(0, \infty)$  to  $(a_+, b_+)$  and monotone increasing from  $(-\infty, 0)$  to  $(a_-, b_-)$ , where  $-\infty \leq a_{\pm} < 0 < b_{\pm} \leq \infty$ , and we have

(1) 
$$\Omega'(x) = -1/f(x), \quad x \neq 0,$$

(2) 
$$\Omega''(x) = f'(x)/f^2(x) \geq 0, \quad x \neq 0.$$

Define  $\gamma(x), \delta(x) (= \gamma_f(x), \delta_f(x))$  on  $(-\infty, 0) \cup (0, \infty)$  by

(D2) 
$$\gamma(x) = \begin{cases} \int_x^1 \left| \frac{\Omega''(u)}{\Omega(u)} \right|^{1/2} du, & x > 0, \\ \int_x^{-1} \left| \frac{\Omega''(u)}{\Omega(u)} \right|^{1/2} du, & x < 0. \end{cases}$$

(D3) 
$$\delta(x) = \begin{cases} \int_x^1 (\Omega''(u))^{1/2} du, & x > 0, \\ \int_x^{-1} (\Omega''(u))^{1/2} du, & x < 0. \end{cases}$$

(Here we interpret  $\gamma(\pm 1)$  to be  $\lim_{x \rightarrow \pm 1} \gamma(x)$ , which is easily shown to be zero.)

Define  $\Gamma(x) (= \Gamma_f(x))$  for  $0 < |x| < 1$  and  $\Delta(x) (= \Delta_f(x))$  for  $x \neq 0, \pm 1$  by

(D4) 
$$\Gamma(x) = \frac{\gamma(x)}{\log \Omega(x)}, \quad \Delta(x) = \frac{|\delta(x)|}{|\Omega(x)|^{1/2}}.$$

Finally, define  $\mathcal{F}$  to be the subset of functions  $f$  of  $\mathcal{C}$  for which

- (i)  $\lim_{x \rightarrow +0} \Omega(x) = b_+ < \infty$  or  $\underline{\lim}_{x \rightarrow +0} \Gamma(x) > 1$ ,
- (ii)  $\lim_{x \rightarrow -0} \Omega(x) = b_- < \infty$  or  $\overline{\lim}_{x \rightarrow -0} \Gamma(x) < -1$ ,
- (H) (iii)  $\underline{\lim}_{x \rightarrow 0} \Delta(x) > 0$ ,
- (iv)  $\underline{\lim}_{|x| \rightarrow \infty} \Delta(x) > 0$ .

The defining conditions for  $\mathcal{F}$  are not very pleasant, but may be shown to hold for  $f(y) = |y|^\alpha \operatorname{sgn} y, \alpha > 0$ ; any finite linear combination of such functions that is in the set  $\mathcal{C}$ ; any analytic function in  $\mathcal{C}$ . Indeed membership of  $\mathcal{F}$  is determined by behavior near  $y = 0$  and near  $|y| = \infty$  (subject to already being in  $\mathcal{C}$ ); hence functions with the asymptotic behavior indicated above will also be in  $\mathcal{F}$ , and we can also allow asymptotic behavior of the type  $|y|^\alpha |\log 1/y|^\beta \operatorname{sgn} y, \alpha, \beta > 0$ .

*Example 1.*  $f(y) = |y|^\alpha \operatorname{sgn} y, \alpha > 1$ .

$$\Omega(x) \sim \frac{x^{1-\alpha}}{\alpha-1} \text{ as } x \rightarrow +0, \quad \Gamma(x) \sim \sqrt{\frac{\alpha}{\alpha-1}} \text{ as } x \rightarrow +0,$$

$$\Delta(x) \sim 2\sqrt{\frac{\alpha}{\alpha-1}} \text{ as } x \rightarrow +0 \text{ and as } x \rightarrow +\infty.$$

*Example 2.*  $f(y) = y$ .

$$\Omega(x) = \log 1/x, \quad \Gamma(x) \rightarrow \infty \text{ as } x \rightarrow +0,$$

$$\Delta(x) \rightarrow \infty \text{ as } x \rightarrow +0 \text{ and as } x \rightarrow +\infty.$$



*Example 3.*  $f(y) = |y|^\alpha \operatorname{sgn} y, 0 < \alpha < 1.$

$$\Omega(x) \sim \frac{1}{1-\alpha} \text{ as } x \rightarrow +0, \quad b_+ < \infty,$$

$$\Delta(x) \sim 2 \frac{\alpha}{1-\alpha} \text{ as } x \rightarrow +0 \text{ and as } x \rightarrow +\infty.$$

(In each of these examples, there is a corresponding result as  $x \rightarrow -0$  etc.)

We shall also wish to consider the subset  $\mathcal{F}_0$  of functions  $f$  of  $\mathcal{F}$  for which  $f$  is twice differentiable on  $(-\infty, 0) \cup (0, \infty)$  with  $xf''(x) \geq 0.$   $\mathcal{F}_0$  includes the functions  $|y|^\alpha \operatorname{sgn} y, \alpha \geq 1.$

**THEOREM 1.** *Let  $f \in \mathcal{F}$  and let  $q$  be continuous. Then either of the two conditions (A2) or (A3) implies that (N) is oscillatory.*

**COROLLARY 1.** *Conditions (A2) and (A3) are both oscillatory criteria for  $(N_\alpha), \alpha > 0.$*

**THEOREM 2.** *Let  $f \in \mathcal{F}$  and suppose that  $b_\pm < \infty.$  Then a sufficient condition for (N) to be oscillatory is that  $\lim_{t \rightarrow \infty} (1/t) \int_T^t \int_T^s q(\tau) d\tau ds$  does not exist as a finite limit or  $-\infty.$*

**COROLLARY 2.** *The above condition is sufficient for the oscillation of  $(N_\alpha)$  for  $0 < \alpha < 1.$*

**THEOREM 3.** *Let  $f \in \mathcal{F}_0$  and let  $q(t) = t^\lambda p(t),$  where  $\lambda > 0$  and  $p$  is a nonconstant, continuous periodic function of period  $\omega$  and mean zero, that is  $\int_0^\omega p(t) dt = 0.$  Then (N) is oscillatory.*

**3. A technical lemma.** Our results will be obtained from an analysis of the Riccati equation associated with (N) and for this, the following lemma will be crucial.

**LEMMA 1.** *Let  $f \in \mathcal{F}$  such that  $b_+ = \infty.$  Let  $\Omega^{-1} (= \Omega_f^{-1})$  be the inverse function from  $(a_+, \infty)$  to  $(0, \infty)$  associated with the restriction of  $\Omega$  to  $(0, \infty),$  and for  $x \in (a_+, \infty)$  define  $F(x)$  to be  $[f'(\Omega^{-1}(x))]^{1/2}.$  Let  $x(t)$  be any eventually positive, continuously differentiable function on  $[T, \infty)$  such that  $x(t)$  is in the domain of  $\Omega^{-1}$  for all  $t \in I$  and such that  $x'(t)F(x(t))$  is not in  $L^2(I).$  Then*

$$\overline{\lim}_{t \rightarrow \infty} \frac{1}{x(t)} \int_T^t \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds > 1.$$

*An analogous result holds if  $b_- = \infty.$*

*Proof.* By (H)(i), there exists  $\lambda > 1$  such that

$$(3) \quad \underline{\lim}_{x \rightarrow +0} \Gamma(x) > \lambda.$$

Suppose the lemma is false. Then for  $t \geq t_1 \geq T,$  say, we have

$$(4) \quad \int_T^t \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds \leq \lambda^{1/2} x(t).$$

Set the left-hand side of (4) equal to  $u(t).$  Then

$$(5) \quad 0 \leq u(t) \leq \lambda^{1/2} x(t), \quad t \geq t_1,$$

$$(6) \quad u''(t) = (x'(t))^2 F^2(x(t)) \geq 0, \quad t \geq t_1.$$

Since  $x'(t)F(x(t))$  is not square integrable on  $[T, \infty),$  it follows from (6) that  $u'(t) \rightarrow \infty, u(t) \rightarrow \infty,$  as  $t \rightarrow \infty,$  and choosing  $t_2 \geq t_1$  so that  $u'(t) > 0, u(t) > 0$  for  $t \leq t_2,$  (5)

and (6) imply for  $t \geq t_2$ , that

$$(7) \quad \left[ \frac{u''(t)}{u(t)} \right]^{1/2} \geq \frac{x'(t)F(x(t))}{\lambda^{1/4} x^{1/2}(t)}.$$

Noting that the left-hand side of (7) is  $(u''(t)/u'(t))^{1/2}(u'(t)/u(t))^{1/2}$ , we may integrate the above inequality between  $t_2$  and  $t \geq t_2$ , and use the Schwarz inequality to obtain

$$(8) \quad \begin{aligned} \left( \log \frac{u'(t)}{u'(t_2)} \right)^{1/2} \left( \log \frac{u(t)}{u(t_2)} \right)^{1/2} &\geq \lambda^{-1/4} \int_{x(t_2)}^{x(t)} F(z)z^{-1/2} dz \\ &\geq \lambda^{-1/4} \int_{x(t_2)}^{u(t)\lambda^{-1/2}} F(z)z^{-1/2} dz, \end{aligned}$$

using (5), since the integrand is positive. Making the change of variable  $\xi = \Omega^{-1}(z)$ , we have

$$(9) \quad \begin{aligned} \int_{x(t_2)}^{u(t)\lambda^{-1/2}} F(z)z^{-1/2} dz &= \int_{h(t)}^{h_0} \left( \frac{f'(\xi)}{\Omega(\xi)} \right)^{1/2} \frac{d\xi}{f(\xi)} \\ &= \int_{h(t)}^{h_0} \left( \frac{\Omega''(\xi)}{\Omega(\xi)} \right)^{1/2} d\xi, \end{aligned}$$

where we have used (2) and put

$$h(t) = \Omega^{-1}(u(t)\lambda^{-1/2}), \quad h_0 = \Omega^{-1}(x(t_2)).$$

By (D2), the right-hand side of (9) is equal to  $\gamma(h(t)) + c$ , where  $c = -\gamma(h_0)$ . Since  $u(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , we have  $h(t) \rightarrow 0$  as  $t \rightarrow \infty$ , and using (3), (D4), we deduce that for  $t$  sufficiently large,

$$(10) \quad \begin{aligned} \gamma(h(t)) + c &\geq \lambda \log \Omega(h(t)) + c = \lambda \log (u(t)\lambda^{-1/2}) + c \\ &\geq \lambda^{3/4} \log \left( \frac{u(t)}{u(t_2)} \right), \end{aligned}$$

for  $t \geq t_3 \geq t_2$ , say.

From (8), (9) and (10), we find that

$$(11) \quad \frac{u'(t)}{u'(t_2)} \geq \left( \frac{u(t)}{u(t_2)} \right)^\lambda, \quad t \geq t_3,$$

which we may write as

$$(12) \quad u'(t)/u^\lambda(t) \geq k = u'(t_2)/u^\lambda(t_2) > 0.$$

Since  $\lambda > 1$ , (12) leads directly to a contradiction on integrating from  $t_3$  to  $\infty$ , and the lemma is proved.

For the proof of Theorem 3 we shall require a slight modification of the above lemma. For  $k \geq 1$ , define  $\Phi_k$  to be the set of bounded, positive continuous function  $\phi$  for which  $\sup_{t \in I} \phi(t) \leq k \inf_{t \in I} \phi(t)$ .

LEMMA 2. Assume the hypotheses and notation of Lemma 1, except that now  $f \in \mathcal{F}_0$ . Then there exists  $k > 1$ , depending on  $f$ , such that if  $\phi \in \Phi_k$  and if  $x(t)$  is continuously differentiable on  $I$  such that  $\int_T^t \phi(s)x'(s) ds$  is eventually positive,  $x(t) \in \text{dom. } \Omega^{-1}$  for all  $t \in I$  and  $x'(t)F(x(t)) \notin L^2(I)$ , then

$$\liminf_{t \rightarrow \infty} \frac{1}{\int_T^t \phi(s)x'(s) ds} \cdot \int_T^t \phi(s) \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds > 1.$$

*Proof.* We proceed along the lines of the proof of Lemma 1. Choose  $\lambda > 1$  such that

$$(3') \quad \lim_{x \rightarrow +0} \Gamma(x) > \lambda.$$

Let  $k$  be any number such that  $1 < k < \lambda$ . If the lemma is false, there exists  $\phi \in \Phi_k$  and  $x(t)$  satisfying the hypotheses of the lemma, such that for  $t \geq t_1$ , say,

$$(4') \quad \int_T^t \phi(s) \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds \leq \lambda^{1/2} \int_T^t \phi(s) x'(s) ds.$$

Setting the left-hand side of (4') equal to  $u(t)$ , we have

$$(5') \quad 0 \leq u(t) \leq \lambda^{1/2} \int_T^t \phi(s) x'(s) ds, \quad t \geq t_1,$$

$$(6') \quad (u'(t)/\phi(t))' = (x'(t))^2 F^2(x(t)) \geq 0, \quad t \geq t_1.$$

As in the proof of Lemma 1, we may find  $t_2 \geq t_1$  so that  $u(t) > 0$ ,  $u'(t) > 0$  for  $t \geq t_2$  and obtain

$$(7') \quad \begin{aligned} \sqrt{\frac{1}{\phi(t)} \frac{(u'(t)/\phi(t))'}{(u'(t)/\phi(t))} \frac{u'(t)}{u(t)}} &= \sqrt{\frac{(u'(t)/\phi(t))'}{u(t)}} \\ &\geq \frac{\lambda^{-1/4} |x'(t)| F(x(t))}{\left(\int_T^t \phi(s) x'(s) ds\right)^{1/2}}, \quad t \geq t_2. \end{aligned}$$

Let  $M = \sup_{t \in I} \phi(t)$ ,  $m = \inf_{t \in I} \phi(t)$ . For  $x \in \text{dom. } \Omega^{-1}$ , we have

$$\begin{aligned} (F^2(x))' &= f''(\Omega^{-1}(x))(\Omega^{-1}(x))' = -f''(\Omega^{-1}(x))f(\Omega^{-1}(x)), \quad \text{by (1),} \\ &\leq 0, \end{aligned}$$

since  $f \in \mathcal{F}_0$  and  $\Omega^{-1}(x) > 0$ . Since  $F(x) > 0$ , it follows that  $F$  is a decreasing function of  $x$  in its domain of definition.

Now  $x(t) \leq \int_T^t |x'(s)| ds + |x(T)|$  and so

$$F(x(t)) \geq F\left(\int_T^t |x'(s)| ds + |x(T)|\right), \quad t \geq t_2.$$

Since  $|\int_T^t \phi(s) x'(s) ds|^{1/2} \leq M^{1/2}(\int_T^t |x'(s)| ds + |x(T)|)$ , it follows from integrating (7') between  $t_2$  and  $t > t_2$ , and using the Schwarz inequality, that

$$\begin{aligned} m^{-1/2} \left[ \log \left( \frac{u'(t)}{u'(t_2)} \cdot \frac{\phi(t_2)}{\phi(t)} \right) \right]^{1/2} \left[ \log \left( \frac{u(t)}{u(t_2)} \right) \right]^{1/2} &\geq \lambda^{-1/4} M^{-1/2} \int_{t_2}^t \frac{|x'(s)| F(w(s))}{w^{1/2}(s)} ds \\ &= \lambda^{-1/4} M^{-1/2} \int_{w(t_2)}^{w(t)} F(z) z^{-1/2} dz, \end{aligned}$$

where

$$w(t) = \int_T^t |x'(s)| ds + |x(T)|.$$

By (5'),  $u(t) \leq \lambda^{1/2} M(\int_T^t |x'(s)| ds + |x(T)|)$ ,  $t \geq t_2$  and so

$$(8') \quad \begin{aligned} \left[ \log \left( \frac{u'(t)}{u'(t_2)} \cdot \frac{\phi(t_2)}{\phi(t)} \right) \right]^{1/2} \left[ \log \left( \frac{u(t)}{u(t_2)} \right) \right]^{1/2} \\ \geq \lambda^{-1/4} m^{1/2} M^{-1/2} \int_{w(t_2)}^{u(t)\lambda^{-1/2}M^{-1}} F(z) z^{-1/2} dz. \end{aligned}$$

As in Lemma 1, we may show that the right hand side of (8') exceeds  $\lambda^{1/2} m^{1/2} M^{-1/2} \log(u(t)/u(t_2))$ , for  $t$  sufficiently large and deduce that

$$\frac{u'(t)}{u'(t_2)} \frac{\phi(t_2)}{\phi(t)} \geq \left(\frac{u(t)}{u(t_2)}\right)^\sigma, \quad \sigma = \lambda m M^{-1} \geq \frac{\lambda}{k} > 1.$$

Thus for sufficiently large  $t$ , we have  $u'(t) \geq (m/M)(u(t)/u(t_2))^\sigma$ , which yields a contradiction as before.

**4. Proofs of the theorems.** Suppose that (N) possesses a nontrivial, nonoscillatory solution  $y$  on  $I = [T, \infty)$ . Without loss of generality, we shall assume that  $y(t) > 0$  on  $I$  and put  $z(t) = y'(t)/f(y(t))$  to obtain the Riccati equation

$$(13) \quad z'(t) = -q(t) - f'(y(t))z^2(t)$$

which we may integrate to obtain

$$(14) \quad z(t) = z(T) - \int_T^t q(s) ds - \int_T^t f'(y(s))z^2(s) ds.$$

Now  $\int_T^t z(s) ds = -\int_{y(t)}^{y(T)} du/f(u) = -\Omega(y(t)) + c_0$  where  $c_0 = \Omega(y(T))$ , and so

$$y(t) = \Omega^{-1} \left[ c_0 - \int_T^t z(s) ds \right]$$

and  $f'(y(t)) = F^2(x(t))$  where  $F(x) = [f'(\Omega^{-1}(x))]^{1/2}$ ,  $x(t) = c_0 - \int_T^t z(s) ds$ . Thus we may rewrite (14) (after rearranging) as

$$(15) \quad \int_T^t q(s) ds = z(T) - \int_T^t F^2(x(s))(x'(s))^2 ds + x'(t).$$

*Proof of Theorem 1.* We integrate (15) between  $T$  and  $t \geq T$  and then divide throughout by  $t$ , obtaining

$$(16) \quad \frac{1}{t} \int_T^t \int_T^s q(\tau) d\tau ds = (1 - T/t)z(T) - \frac{1}{t} \int_T^t \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds + \frac{1}{t}(x(t) - x(T)).$$

There are two cases to consider:

I.  $x'(t)F(x(t)) \in L^2(I)$ . Let  $t_n \geq T$  be chosen so that  $[\int_{t_n}^\infty (x'(t))^2 F^2(x(t)) dt]^{1/2} < 1/n$ ,  $n = 1, 2, \dots$ . For  $t \geq t_n$ , we may use the Schwarz inequality to obtain

$$(17) \quad \begin{aligned} \frac{1}{n} &> \left[ \int_{t_n}^t (x'(s))^2 F^2(x(s)) ds \right]^{1/2} \geq (t - t_n)^{-1/2} \left| \int_{t_n}^t F(x(s))x'(s) ds \right| \\ &= (t - t_n)^{-1/2} \left| \int_{t_n}^t [f'(\Omega^{-1}(x(s)))]^{1/2} x'(s) ds \right| \\ &= (t - t_n)^{-1/2} \left| \int_{\Omega^{-1}(x(t_n))}^{\Omega^{-1}(x(t))} ([f'(u)]^{1/2}/f(u)) du \right| \\ &= (t - t_n)^{-1/2} \left| \int_{\Omega^{-1}(x(t))}^{\Omega^{-1}(x(t_n))} [\Omega''(u)]^{1/2} du \right|, \quad \text{by (2)} \\ &\geq (t - t_n)^{-1/2} [|\delta(\Omega^{-1}(x(t)))| - c_n], \quad \text{by (D3)} \end{aligned}$$

where  $c_n = |\delta(\Omega^{-1}(x(t_n)))|$ .

By (H)(iii) and (iv), there exist  $A, r > 0$  such that  $\Delta(u) \cong A$  whenever  $|u| < r$  or  $|u| > 1/r$ , that is

$$(18) \quad |\delta(u)| \cong A|\Omega(u)|^{1/2}, \quad |u| < r \quad \text{or} \quad |u| > \frac{1}{r}.$$

From (17) and (18), it follows that for any  $t > t_n$ , either  $r \leq \Omega^{-1}(x(t)) \leq 1/r$  or  $1/n \geq (t - t_n)^{-1/2}[A|x(t)|^{1/2} - c_n]$  which we may write as

$$(19) \quad r \leq \Omega^{-1}(x(t)) \leq \frac{1}{r} \quad \text{or} \quad \frac{|x(t)|}{t - t_n} \leq \frac{1}{A^2} \left[ \frac{1}{n} + c_n(t - t_n)^{-1/2} \right]^2.$$

Since  $\Omega$  is continuous on  $(0, \infty)$ , there exists  $B > 0$  such that  $r \leq \Omega^{-1}(x(t)) \leq 1/r$  implies  $|x(t)| \leq B$  and so (19) implies that for all  $t > t_n$ ,

$$(20) \quad \begin{aligned} \frac{x(t)}{t - t_n} &\leq \max \left[ \frac{B}{t - t_n}, \frac{1}{A^2} \left\{ \frac{1}{n} + c_n(t - t_n)^{-1/2} \right\}^2 \right] \\ &\leq \frac{2}{A^2 n^2} \quad \text{if } t > t_n \text{ is sufficiently large,} \end{aligned}$$

and so we must have

$$(21) \quad \lim_{t \rightarrow \infty} \frac{x(t)}{t} = 0.$$

It follows from the integrability of  $(x'(t))^2 F^2(x(t))$  and from (21) that the right-hand side of (16) has a finite limit, and so the left-hand side of (16) has a finite limit, as  $t \rightarrow \infty$ , which will contradict each of conditions (A2), (A3).

The second case to consider is

II.  $x'(t)F(x(t)) \notin L^2(I)$ . If  $b_+ < \infty$ , then  $x(t) = \Omega(y(t))$  is bounded above by  $C$ , say. On the other hand there exist  $\tau_n \rightarrow \infty$  such that  $\int_T^t \int_T^s (x'(\tau))^2 F^2(x(\tau)) \, d\tau \, ds > nt$  whenever  $t \geq \tau_n$  and so for  $t \geq \tau_n$ , we see that the right-hand side of (16) is bounded above by  $z(T) + (C - x(T))/t - n$  and it follows that the left-hand side of (16) has the limit  $-\infty$  as  $t \rightarrow \infty$ , contradicting both (A2) and (A3).

Suppose then that  $b_+ = \infty$ . If  $x(t)$  is not eventually positive, choose  $s_n \rightarrow \infty$  with  $x(s_n) \leq 0$  and arguing as above we deduce that  $\liminf_{t \rightarrow \infty} (1/t) \int_T^t \int_T^s q(\tau) \, d\tau \, ds = -\infty$ , contradicting (A2) and (A3). If  $x(t) > 0$  for  $t \geq t_1 \geq T$ , say, we may apply Lemma 1 to obtain

$$\overline{\lim}_{t \rightarrow \infty} \frac{1}{x(t)} \int_T^t \int_T^s (x'(\tau))^2 F^2(x(\tau)) \, d\tau \, ds > 1$$

and so there exists  $\lambda > 1, T_n \rightarrow \infty$  such that

$$\frac{1}{x(T_n)} \int_T^{T_n} \int_T^s (x'(\tau))^2 F^2(x(\tau)) \, d\tau \, ds > \lambda$$

which implies that

$$-\frac{1}{T_n} \left[ \int_T^{T_n} \int_T^s (x'(\tau))^2 F^2(x(\tau)) \, d\tau \, ds - x(T_n) \right] \leq -\frac{(1 - \lambda^{-1})}{T_n} \int_T^{T_n} \int_T^s (x'(\tau))^2 F^2(x(\tau)) \, d\tau \, ds$$

$\rightarrow -\infty$  as  $n \rightarrow \infty$ , since  $x'(t)F(x(t)) \notin L^2(I)$ , and using (16) we again find that  $\liminf_{t \rightarrow \infty} (1/t) \int_T^t \int_T^s q(\tau) \, d\tau \, ds = -\infty$ , contradicting (A2) and (A3).

This completes the proof of the theorem, and the corollary follows from the remarks in § 2.

*Proof of Theorem 2.* If we assume that  $b_+ < \infty$  and that (N) has a positive nonoscillatory solution on  $[T, \infty)$ , then the only possibilities in the preceding argument that need to be considered are

I.  $x'(t)F(x(t)) \in L^2(I)$ , in which case it was shown that the left-hand side of (16) must have a finite limit as  $t \rightarrow \infty$ , and

II.  $x'(t)F(x(t)) \notin L^2(I)$ ,  $b_+ < \infty$ , in which case it was shown that the left-hand side of (16) must tend to  $-\infty$  as  $t \rightarrow \infty$ .

Theorem 2 now follows.

*Proof of Theorem 3.* Since  $f \in \mathcal{F}_0$ , we may choose  $k > 1$  in accordance with the conclusion of Lemma 2. We have

$$\int_T^t q(s) ds = t^\lambda P(t) + O(t^{\lambda-1}),$$

where  $P$  is a nonconstant periodic function of period  $\omega$  and mean zero. Define  $P_-(t)$  to be

$$\begin{cases} P(t), & P(t) \leq 0, \\ 0, & \text{otherwise,} \end{cases}$$

and for  $\varepsilon > 0$ , define

$$P_+^\varepsilon(t) = \begin{cases} P(t), & P(t) \geq \varepsilon \\ 0, & \text{otherwise.} \end{cases}$$

Since  $P$  has mean zero we may choose  $\varepsilon$  sufficiently small that

$$(22) \quad \frac{k+1}{2} \int_{n\omega}^{(n+1)\omega} P_+^\varepsilon(t) dt + \int_{n\omega}^{(n+1)\omega} P_-(t) dt \geq 0, \quad n = 1, 2, \dots$$

Define  $\phi(t)$  to be a continuous function with

$$\phi(t) = \begin{cases} k, & \text{if } P(t) \geq \varepsilon, \\ 1, & \text{if } P(t) \leq 0 \end{cases}$$

and  $1 \leq \phi(t) \leq k$  for all  $t$ . Then  $\phi \in \Phi_k$ .

For  $N\omega \leq t < (N+1)\omega$ , we have

$$\begin{aligned} & \int_T^t \phi(s) \int_T^s q(\tau) d\tau ds \geq \int_T^t \phi(s) s^\lambda P(s) ds + O(t^\lambda) \\ & = \int_0^t \phi(s) s^\lambda P(s) ds + O(t^\lambda) \\ (23) \quad & \geq k \sum_{n=0}^{N-1} (n\omega)^\lambda \int_{n\omega}^{(n+1)\omega} P_+^\varepsilon(s) ds + \sum_{n=0}^N ((n+1)\omega)^\lambda \int_{n\omega}^{(n+1)\omega} P_-(s) ds + O(t^\lambda) \\ & \geq \frac{k-1}{2} \left( \sum_{n=0}^{N-1} (n\omega)^\lambda \right) \int_0^\omega P_+^\varepsilon(s) ds + O(N^\lambda), \quad \text{by (22)} \\ & \geq At^{\lambda+1} \end{aligned}$$

for some positive constant  $A$ , independent of  $t$ , for all sufficiently large  $t$ .

Now we return to (15), multiply throughout by  $\phi(t)$ , integrate between  $T$  and  $t \geq T$  and divide throughout by  $\int_T^t \phi(s) ds$  to obtain

$$(24) \quad \frac{1}{\int_T^t \phi(s) ds} \int_T^t \phi(s) \int_T^s q(\tau) d\tau ds = z(T) - \frac{1}{\int_T^t \phi(s) ds} \left[ \int_T^t \phi(s) \int_T^s (x'(\tau))^2 F^2(x(\tau)) d\tau ds - \int_T^t \phi(s)x'(s) ds \right].$$

By (23), the left-hand side of (24) tends to  $+\infty$  as  $t \rightarrow \infty$ . Define  $w(t)$  to be  $|x(T)| + \int_T^t |x'(s)| ds$ . Then  $x(t) \leq w(t)$ ,  $(x'(t))^2 = (w'(t))^2$ . Since  $f \in \mathcal{F}_0$ ,  $F(x)$  is decreasing in  $x$  (see proof of Lemma 2) and so we have  $0 \leq F(w(t)) \leq F(x(t))$ . Now suppose that  $x'(t)F(x(t)) \in L^2(I)$ . Then  $w'(t)F(w(t)) \in L^2(I)$ , and following the analysis of case II of Theorem 1, we may deduce that  $w(t)/t \rightarrow 0$  as  $t \rightarrow \infty$ . It follows that  $(1/t) \int_T^t \phi(s)x'(s) ds \rightarrow 0$  as  $t \rightarrow \infty$  and so the right-hand side of (24) has a finite limit as  $t \rightarrow \infty$ , yielding a contradiction. On the other hand, if  $x'(t)F(x(t)) \notin L^2(I)$ , we may use Lemma 2 to deduce that the right-hand side of (24) has the limit infimum  $-\infty$  as  $t \rightarrow \infty$ , and again we have the contradiction. This completes the proof of the theorem. (We note that  $f \in \mathcal{F}_0$  implies that  $b_+ = \infty$ .)

*Remarks.* If  $q(t) = t^\lambda p(t)$  where  $\lambda$  is a real number and  $p$  is a nonconstant periodic function of mean zero, then for  $\alpha > 1$ ,  $(N_\alpha)$  is oscillatory iff  $\lambda \geq -1$ . The case  $\lambda > 0$  follows from Theorem 3, the case  $\lambda = 0$  was proved in [5], and for  $\lambda < 0$  is a consequence of Theorem 2.3 of [4]. For  $0 < \alpha < 1$ ,  $(N_\alpha)$  is oscillatory if  $\lambda \geq 1$  or  $\lambda = 0$ , the case  $\lambda \geq 1$  following from Theorem 2. The case  $\lambda = 0$  was proved in [5]. If  $\lambda < 0$  or  $0 < \lambda < 1$ , the oscillatory character is unknown, but we conjecture oscillation iff  $\lambda \geq -\alpha$ .

For the case  $\alpha = 1$ , we refer to [17].

REFERENCES

[1] F. V. ATKINSON, *On second order nonlinear oscillation*, Pacific J. Math., 5 (1955), pp. 643–647.  
 [2] S. BELOHOREC, *Oscillatory solutions of certain nonlinear differential equations of the second order*, Mat. Casopis Sloven. Akad. Vied., 11 (1961), pp. 250–255.  
 [3] T. A. BURTON AND R. GRIMMER, *On the continuability of second order differential equations*, Proc. Amer. Math. Soc., 29 (1971), pp. 277–283.  
 [4] G. J. BUTLER, *On the oscillatory behaviour of a second order nonlinear differential equation*, Ann. Mat. Pura Appl., 105 (1975), pp. 73–92.  
 [5] ———, *Oscillation theorems for a nonlinear analogue of Hill's equation*, Quart. J. Math., (2) 27 (1976), 159–171.  
 [6] ———, *The existence of continuable solutions of a second order differential equation*, Canad. J. Math., 29 (1977), 472–479.  
 [7] W. J. COLES, *An oscillation criterion for second order differential equations*, Proc. Amer. Math. Soc., 19 (1968), pp. 755–759.  
 [8] ———, *Oscillation criteria for nonlinear second order equations*, Ann. Mat. Pura Appl., 82 (1969), pp. 123–134.  
 [9] W. B. FITE, *Concerning the zeros of the solutions of certain differential equations*, Trans. Amer. Math. Soc., 19 (1918), pp. 341–352.  
 [10] P. HARTMAN, *On nonoscillatory linear differential equations of second order*, Amer. J. Math., 74 (1952), pp. 389–400.  
 [11] I. V. KAMENEV, *Some specifically nonlinear oscillation theorems*, Mat. Zametki, 10 (1971), pp. 129–134 (In Russian).  
 [12] ———, *Oscillation criteria related to averaging of solutions of ordinary differential equations of second order*, Differencial'nye Uravnenija, 10 (1974), pp. 246–252. (In Russian.)  
 [13] I. T. KIGURADZE, *A note on the oscillation of  $u'' + a(t)|u|^n \operatorname{sgn} u = 0$* , Casopis Pest. Mat., 92 (1967), pp. 343–350. (In Russian.)

- [14] J. W. MACKI AND J. S. W. WONG, *Oscillation of solutions to second order nonlinear differential equations*, Pacific J. Math., 24 (1968), pp. 111–117.
- [15] H. ANOSE, *Oscillation criteria for second order nonlinear differential equations*, Proc. Amer. Math. Soc., 51 (1974), pp. 67–73.
- [16] P. WALTMAN, *An oscillation theorem for a nonlinear second order equation*, J. Math. Anal. Appl., 10 (1965), pp. 439–441.
- [17] D. W. WILLETT, *On the oscillatory behavior of the solutions of second order linear differential equations*, Ann. Polon. Math., 21 (1969), pp. 175–194.
- [18] A. WINTNER, *A criterion of oscillatory stability*, Quart. Appl. Math., 7 (1949), pp. 115–117.
- [19] J. S. W. WONG, *On second order nonlinear oscillation*, Funkcial Ekvac., 11 (1969), pp. 207–234.
- [20] ———, *A second order nonlinear oscillation theorem*, Proc. Amer. Math. Soc., 40 (1973), pp. 487–491.
- [21] ———, *Oscillation theorems for second order nonlinear differential equations*, Bull. Inst. Math. Acad. Sinica, 3 (1975), pp. 283–309.



## A FREE BOUNDARY OPTIMIZATION PROBLEM. II\*

ANDREW ACKER†

**Abstract.** Given a compact set  $Q \subset \mathbb{R}^2$ , a function  $a(p) > 0$  continuous on  $\mathbb{R}^2$ , and a sufficiently large constant  $A > 0$ , we determine (under suitable assumptions) the doubly-connected region  $\Omega \subset \mathbb{R}^2$  encircling (but not intersecting)  $Q$  which has the least capacitance subject to the constraint that  $|\Omega| := \iint_{\Omega} a^2(p) \, dx \, dy \leq A$ .

**1. Introduction and main results.** Our purpose is to generalize the isoperimetric inequality [2, Thm. 2 (Case 1)], which involved the capacity of a condenser. We will use the notation introduced in [2, § 1]. This investigation is motivated by the following free boundary optimization problem.

*Problem 1.* Given a simply-connected, compact set  $Q \subset \mathbb{R}^2$  (whose boundary  $\partial Q$  has bounded curvature), a function  $a(p) > 0$  continuous on  $\mathbb{R}^2$ , and any constant  $A \in \mathbb{R}_+ = (0, \infty)$ , we seek a doubly-connected region  $\Omega$  which has minimum capacity subject to the constraints that  $S^* \supset Q$  (i.e.,  $\Omega$  encircles  $Q$  without intersecting it) and  $|\Omega| := \iint_{\Omega} a^2(p) \, dx \, dy \leq A$ .

In [2], Problem 1 was solved for all  $A \in \mathbb{R}_+$  in the special case where  $Q$  is convex and  $a(p) \equiv 1$  on  $\mathbb{R}^2$ . In this paper, we succeed in solving Problem 1 for all sufficiently large  $A \in \mathbb{R}_+$  under considerably more general assumptions concerning  $Q$  and the function  $a(p)$ .

We seek a solution  $\Omega$  with the property that

$$(1) \quad S^* = Q,$$

i.e.,  $\Omega$  does not separate away from the geometric constraint  $Q$ . It follows by applying the Poincaré variation formula for capacity [2, (3)] that a sufficiently regular region  $\Omega$  satisfying (1) can solve Problem 1 only if  $\Omega$  satisfies the following conditions:

$$(2) \quad |\nabla U(p)| = c \cdot a(p) \quad \text{on } \Gamma$$

and

$$(3) \quad |\nabla U(p)| \geq c \cdot a(p) \quad \text{on } \Gamma^* = \partial Q$$

for some value  $c \in \mathbb{R}_+$ , and

$$(4) \quad |\Omega| = A.$$

(For  $p \in \Gamma^* \cup \Gamma$ , we define  $|\nabla U(p)| = \lim_{q \rightarrow p} |\nabla U(q)|$ ,  $q \in \Omega$ , when the limit exists.)

For each  $c \in \mathbb{R}_+$ , (1) and (2) constitute a well posed free boundary problem under the conditions of the following theorem. (See [1, Lemma 11], [3] and [5].)

**THEOREM 1.** *Assume that  $Q$  is starlike relative to some point  $p_0 \in Q$ , and that  $\lambda \cdot a(p_0 + \lambda \cdot (p - p_0))$  is (weakly) monotone increasing in  $\lambda \in [1, \infty)$  for each  $p \in \partial Q$ . Then*

- (a) *For each  $c \in \mathbb{R}_+$ , there exists a unique region  $\Omega_c$  such that  $S_c^* = Q$  and  $|\nabla U_c(p)| = c \cdot a(p)$  on  $\Gamma_c$ .*
- (b)  *$\Omega_{c'} \subset \Omega_c$  whenever  $0 < c \leq c' < \infty$ , and  $\bigcup_{c \in \mathbb{R}_+} \Gamma_c = \mathbb{R}^2 \setminus Q$ .*
- (c)  *$Q \cup \Omega_c$  is starlike relative to  $p_0$  for each  $c \in \mathbb{R}_+$ .*

\* Received by the editors December 9, 1977 and in revised form May 15, 1978.

† Mathematisches Institut I, Universität Karlsruhe (TH), 75 Karlsruhe 1, Federal Republic of Germany.

Under the assumptions of Theorem 1, there is for each  $A \in R_+$  a unique  $c \in R_+$  such that  $|\Omega_c| = A$ . Since each region  $\Omega_c, c \in R_+$ , is the unique solution of (1) and (2) at the area  $A = |\Omega_c|$ , it is clear that Problem 1 has no sufficiently regular solution at the area  $A = |\Omega_c|$  which satisfies (1) unless  $\Omega_c$  satisfies (3), i.e.,  $|\nabla U_c(p)| \cong c \cdot a(p)$  on  $\Gamma_c^* = \partial Q$ . Thus, it remains to answer two questions.

1. For what values of  $c \in R_+$  does  $\Omega_c$  satisfy (3)?
2. Assuming  $\Omega_c$  satisfies (3), is  $\Omega_c$  a solution of Problem 1 at the area  $A = |\Omega_c|$ ?

Both questions are answered (the second affirmatively) by the following theorem, which is our main result.

**THEOREM 2.** *Assume in Problem 1 that  $Q$  is starlike relative to  $p_0 \in Q$  and that  $\lambda \cdot a(p_0 + \lambda \cdot (p - p_0))$  is (weakly) monotone increasing in  $\lambda \in [1, \infty)$  for each  $p \in \partial Q$ . If  $\log(a(p))$  is subharmonic in  $R^2 \setminus Q$ , then:*

- (a) *There is a value  $0 \leq C_0 \leq \infty$  such that  $|\nabla U_c(p)| \cong c \cdot a(p)$  on  $\partial Q$  only for those  $c \in R_+$  satisfying  $c \leq C_0$ . In fact  $|\nabla U_c(p)| \cong c \cdot a(p)$  throughout  $CL(\Omega_c)$  if  $c \in [0, C_0] \cap R_+$  (where  $CL = \text{Closure}$ ).*
- (b) *If  $c \in [0, C_0] \cap R_+$ , then*

$$(5) \quad K \cong K_c$$

*for any region  $\Omega$  satisfying  $S^* \supset Q$  and  $|\Omega| \leq |\Omega_c|$ . Thus,  $\Omega_c$  solves Problem 1 at the area  $A = |\Omega_c|$ .*

- (c) *If  $C_0 < c < \infty$ , then (5) does not hold for all admissible  $\Omega$ , and  $\Omega_c$  does not solve Problem 1 at the area  $A = |\Omega_c|$ .*

**Remark 1.** For the purpose of discussing uniqueness in the context of Theorem 2, we define  $C_1 \in (0, \infty]$  as follows: If  $Q = \{p \in R^2 : |p - p_0| \leq r_0\}$ , then  $C_1 = (1/\alpha \cdot \log(r_1/r_0))$ , where  $r_1 \in [r_0, \infty]$  is the maximum value such that  $|p - p_0| \cdot a(p) = \alpha \in R_+$  throughout  $R(r_0, r_1) := \{p \in R^2 : r_0 < |p - p_0| < r_1\}$ . Otherwise  $C_1 = \infty$ . Notice that  $C_0 = \infty$  whenever  $C_1 < \infty$ . Under the assumptions of Theorem 2, if  $0 < c \leq C_1 < \infty$  or if  $C_1 = \infty$  and  $c \in [0, C_0] \cap R_+$ , then the proof of (5) in § 5 can be extended to show that actually

$$(6) \quad K > K_c$$

for any region  $\Omega \neq \Omega_c$  satisfying  $S^* \supset Q$  and  $|\Omega| \leq |\Omega_c|$ . Thus  $\Omega_c$  is the unique solution of Problem 1 at the area  $A = |\Omega_c|$ . Equation (6) does not hold if  $C_1 < c < \infty$ . In this case, the (only) solutions of Problem 1 at the area  $A = |\Omega_c|$  are the annuli  $R(\sigma, \tau)$ , where  $r_0 \leq \sigma < \tau \leq r_1, \tau < \infty$ , and  $(\tau/\sigma) = \exp(A/(2\pi))$ .

**Remark 2.** Assume  $a(p) \cong \delta \cdot (1 + |p|)^{\delta-1}$  in  $R^2$  for some  $0 < \delta < 1$ . Then under the assumptions of Theorem 2, one can show that  $C_0 > 0$ , and therefore that Problem 1 has a solution satisfying (1) for any sufficiently large  $A \in R_+$ . If  $a(p) \equiv 1$  in  $R^2$ , then  $C_0 = \infty$  (i.e., Problem 1 has a solution satisfying (1) for all  $A \in R_+$ ) if and only if  $Q$  is convex, as was shown in [2, Thm. 2].

**Remark 3.** Under the assumptions of Theorem 2, the regions  $\Omega_c, c \in [0, C_0] \cap R_+$ , have the following area-minimizing property equivalent to Theorem 2(b):  $\Omega_c$  minimizes  $|\Omega|$  in the class of all  $\Omega$  which are conformally equivalent to  $\Omega_c$  and satisfy  $S^* \supset Q$ .

**Remark 4.** Under the assumptions of Theorem 1, it was shown in [1, Thms. 10 and 12] that

$$(7) \quad K - K_c \cong c^2 \cdot (|\Omega_c| - |\Omega|)$$

for any  $c \in R_+$  and  $\Omega$  satisfying  $S^* = Q$ , and that in fact each region  $\Omega_c, c \in R_+$ , uniquely minimizes capacitance in the class of all  $\Omega$  satisfying  $S^* = Q$  and  $|\Omega| \leq A := |\Omega_c|$ . If

$C_0 \in \mathbb{R}_+$ , one can show using (5) and (7) that (7) holds for any  $c \in \mathbb{R}_+$  and any  $\Omega$  satisfying  $S^* \supset Q$  and  $|\Omega| \geq |\Omega_{C_0}|$ .

**2. The proof of Theorem 2(a).** For each  $c \in \mathbb{R}_+$ ,  $|\nabla U_c(p)| \geq c \cdot a(p)$  on  $\partial Q$  if and only if the same inequality holds throughout  $CL(\Omega_c)$ . Indeed, this immediately follows by applying the maximum principle to the function  $\phi_c(p) := \log(|\nabla U_c(p)|/(c \cdot a(p)))$ , which is continuous on  $CL(\Omega_c)$  and superharmonic on  $\Omega_c$ , and which vanishes on  $\Gamma_c$ . Thus, to prove Theorem 2(a) we must show that

$$(8) \quad H = [0, C_0] \cap \mathbb{R}_+$$

for some  $C_0 \in [0, \infty]$ , where  $H = \{c \in \mathbb{R}_+ : |\nabla U_c(p)| \geq c \cdot a(p) \text{ on } \partial Q\}$ . Since  $H$  is closed relative to  $\mathbb{R}_+$ , it suffices to show for each  $c \in H$  that  $c - \delta \in H$  for all sufficiently small  $\delta > 0$ . We will prove this under the further assumption that  $a(p)$  is analytic in each coordinate variable. For each  $c \in \mathbb{R}_+$ , it follows using Lewy's theorem [4] (see Lemma 5) that  $(U_c(p)/c)$  has a harmonic continuation  $V_c(p)$  to  $\Omega_c \cup O_c$ , where  $O_c$  is a neighborhood of  $\Gamma_c$ . If  $\phi_c(p) := \log(|\nabla V_c(p)|/a(p))$  on  $\Omega_c \cup O_c$ , then by the strong maximum principle, either  $\phi_c(p) \equiv 0$  in  $\Omega_c$  (only possible when  $\Omega_c$  is an annulus) or else  $\phi_c(p) > 0$  in  $\Omega_c$ , implying that  $D_n \phi_c(p) > 0$  on  $\Gamma_c$  (where  $D_n$  differentiates in the interior normal direction). In either case, we conclude easily that  $\phi_c(p) \leq 0$  in  $\bar{O}_c \setminus \Omega_c$ , where  $\bar{O}_c \subset O_c$  is a suitable neighborhood of  $\Gamma_c$ . Thus, for  $\delta > 0$  sufficiently small (so that  $\Omega_{c-\delta} \cup \Gamma_{c-\delta} \subset \Omega_c \cup \bar{O}_c$ ), we have that  $|\nabla V_c(p)| \leq a(p) = |\nabla V_{c-\delta}(p)|$  on  $\Gamma_{c-\delta}$ . Therefore, if  $W_\delta(p) = V_{c-\delta}(p) - V_c(p)$ , then  $D_n W_\delta(p) \geq a(p) - |\nabla V_c(p)| \geq 0$  on  $\Gamma_{c-\delta}$ . Since  $W_\delta(p) = (1/(c-\delta)) - (1/c)$  on  $\partial Q$ , it follows from the maximum principle that  $W_\delta(p) \leq (1/(c-\delta)) - (1/c)$  in  $\Omega_{c-\delta}$ , implying that  $|\nabla V_{c-\delta}(p)| \geq |\nabla V_c(p)| \geq a(p)$  on  $\partial Q$ . Therefore  $c - \delta \in H$ , proving (8) in the case where  $a(p)$  is analytic in each coordinate variable.

To prove (8) in the general case, we will show that  $(0, \bar{c}] \subset H$  if  $\bar{c} \in H$ . For  $\bar{c} \in H$  fixed and for each  $n \in \mathbb{N}$ , one can define a function  $a_n(p)$  with the properties assumed of  $a(p)$  such that  $a_n(p)$  is analytic in each coordinate variable,  $a_n(p) \leq a(p)$  on  $\partial Q$ ,  $a_n(p) \geq a(p)$  on  $\Gamma_{\bar{c}}$ , and  $|a_n(p) - a(p)| < (1/n)$  in  $\Omega_{\bar{c}}$ . It suffices to show that  $(0, \bar{c}] \subset H_n := \{c \in \mathbb{R}_+ : |\nabla U_{n,c}(p)| \geq c \cdot a_n(p) \text{ on } \partial Q\}$  for each  $n \in \mathbb{N}$ , where  $\Omega_{n,c}$  is defined such that  $S_{n,c}^* = Q$  and  $|\nabla U_{n,c}(p)| = c \cdot a_n(p)$  on  $\Gamma_{n,c}$ . Now  $\bar{c} \in H_n$  for each  $n \in \mathbb{N}$ , since  $\Omega_{n,\bar{c}} \subset \Omega_{\bar{c}}$ , and therefore  $|\nabla U_{n,\bar{c}}(p)| \geq |\nabla U_{\bar{c}}(p)| \geq \bar{c} \cdot a(p) \geq \bar{c} \cdot a_n(p)$  on  $\partial Q$ . Therefore  $(0, \bar{c}] \subset H_n$  for each  $n$ , since (8) has already been shown to hold when  $a(p)$  is analytic in each coordinate variable.

**3. Preliminary lemmas for the proof of Theorem 2(b).** Let  $F(z) : \mathbb{R} \times (0, 1) \rightarrow \Omega$  be a  $K$ -periodic, analytic mapping onto  $\Omega$ .  $F(z)$  can be defined to be the analytic continuation of  $E^{-1}(-jz)$  to  $\mathbb{R} \times (0, 1)$ , where  $E(z) = U + jV$  ( $j = \sqrt{-1}$ ) and  $V(p)$  is a harmonic conjugate of  $U(p)$ . The mappings  $\hat{F}_i(z)$  and  $\hat{F}_i^*(z)$  are defined analogously relative to the regions  $\hat{\Omega}_i$  and  $\hat{\Omega}_i^*$  defined in this section.

LEMMA 3. Let  $Q$  be starlike relative to  $p_0 \in Q$ , and assume  $\lambda \cdot a(p_0 + \lambda \cdot (p - p_0))$  is (weakly) increasing in  $\lambda \in [1, \infty)$  for each  $p \in \partial Q$ . Let  $\Omega_{\hat{c}}$ ,  $\hat{c} \in \mathbb{R}_+$ , and  $\Omega$  be regions such that  $S_{\hat{c}}^* = Q$ ,  $|\nabla U_{\hat{c}}(p)| = \hat{c} \cdot a(p)$  on  $\Gamma_{\hat{c}}$ ,  $S^* \supset Q$ , and  $A_0 := |S_c \setminus S| = |S^* \setminus Q|$ . Then for any  $\delta > 0$  and  $n \in \mathbb{N}$  satisfying  $n \cdot \delta \leq A_0$ , there exists a sequence of regions  $\hat{\Omega}_i$ ,  $i = 0, \dots, n$ , with the following properties:

- (a)  $\hat{\Omega}_0 = \Omega_{\hat{c}}$ .
- (b)  $\hat{S}_i^* \subset \hat{S}_{i+1}^*$  and  $\hat{S}_i \supset \hat{S}_{i+1}$ ,  $i = 0, \dots, n-1$ .
- (c)  $\hat{S}_i^*$  and  $\hat{S}_i^* \cup \hat{\Omega}_i$  are both starlike relative to  $p_0$  for all  $i$ .
- (d1) For each  $i$ ,  $|\nabla \hat{U}_i(p)| = c_i \cdot a(p)$  on  $\hat{\Gamma}_i$  for some constant  $c_i > 0$ .
- (d2) If  $\log(a(p))$  is subharmonic in  $\mathbb{R}^2 \setminus Q$ , and if  $\hat{c} \leq C_0$ , then  $|\nabla \hat{U}_i(p)| \geq c_i \cdot a(p)$  throughout  $\hat{\Omega}_i$  for each  $i$ .

(e)  $|(\dot{S}_{i+1}^* \setminus \dot{S}_i^*) \cap S^*| = |(\dot{S}_i \setminus \dot{S}_{i+1}) \setminus S| = \delta, i = 0, \dots, n-1$ . Thus  $|(\dot{S}_i^* \setminus Q) \cap S^*| = |(\dot{S}_i \setminus \dot{S}_i) \setminus S| = i \cdot \delta$  for each  $i$ .

*Proof.* For  $\delta > 0$  fixed, we prove the lemma by induction on  $n$ . The assertions are trivial for  $n = 0$ . If  $\dot{\Omega}_0, \dots, \dot{\Omega}_n$  satisfy (a)–(e), and if  $(n + 1) \cdot \delta \leq A_0$ , then by Theorem 1 there exists a unique smallest region  $\dot{\Omega}_n \supset \dot{\Omega}_n$  such that  $\dot{S}_n^* = \dot{S}_n^*, |(\dot{S}_n \setminus \dot{S}_n) \setminus S| = \delta$ , and  $|\nabla \dot{U}_n(p)| = \dot{c}_n \cdot a(p)$  on  $\dot{\Gamma}_n$  for a constant  $0 < \dot{c}_n < c_n$ . We define  $\dot{\Omega}_{n+1} = \dot{\Omega}_n(0, \lambda)$ , where  $0 < \lambda < 1$  is the largest value such that  $|S^* \cap \dot{\Omega}_n(\lambda, 1)| = \delta$ . (Here  $\Omega(\alpha, \beta) := \{p \in \Omega : \alpha < U(p) < \beta\}, 0 \leq \alpha < \beta \leq 1$ .) That  $\dot{\Omega}_0, \dots, \dot{\Omega}_{n+1}$  satisfy (a)–(e) (with  $(n + 1)$  substituted for  $n$ ) is easily seen from the construction of  $\dot{\Omega}_{n+1}$  and the following considerations.

*Concerning (c):* Since  $\dot{S}_n^*$  is starlike relative to  $p_0$ , Theorem 1(c) implies that  $\dot{S}_{n+1}^* \cup \dot{\Omega}_{n+1} := \dot{S}_n^* \cup \dot{\Omega}_n$  is starlike relative to  $p_0$ . Therefore, assuming  $p_0 = 0$ , we have that  $|\psi_n(z)| \leq \pi/2$  on  $R \times \{0, 1\}$ , where  $\psi_n(z) := \arg(D_z \dot{F}_n(z) / \dot{F}_n(z))$  is harmonic in  $R \times (0, 1)$ . Therefore  $|\psi_n(z)| \leq \pi/2$  throughout  $R \times [0, 1]$ , implying that  $\dot{S}_{n+1}^* := \dot{S}_n^* \cup \Omega_n(\lambda, 1)$  is star-like relative to  $p_0 = 0$ . *Concerning (d1) and (d2):*  $|\nabla \dot{U}_{n+1}(p)| = (|\nabla \dot{U}_n(p)| / \lambda) = (\dot{c}_n / \lambda) \cdot a(p) = c_{n+1} \cdot a(p)$  on  $\dot{\Gamma}_{n+1} := \dot{\Gamma}_n$ . Furthermore, if  $\log(a(p))$  is subharmonic in  $R^2 \setminus Q$  and  $|\nabla \dot{U}_n(p)| \geq c_n \cdot a(p)$  in  $\dot{\Omega}_n$ , then  $\lambda \cdot (|\nabla \dot{U}_{n+1}(p)| - c_{n+1} \cdot a(p)) = (|\nabla \dot{U}_n(p)| - \dot{c}_n \cdot a(p)) \geq 0$  in  $\dot{\Omega}_{n+1} \subset \dot{\Omega}_n$  by application of Theorem 2(a).

In the proof of Theorem 2(b) in § 5, we will make use of the following stronger assumptions concerning  $Q$  and the function  $a(p)$ :

(A1)  $Q$  is compact and starlike relative to each point  $p \in B_\rho(p_0) := \{p \in R^2 : |p - p_0| < \rho\}$ , where  $\rho > 0$  is sufficiently small. Also,  $\partial Q$  has bounded curvature.

(A2)  $\lambda \cdot a(p + \lambda \cdot (q - p))$  is (weakly) increasing in  $\lambda \in [1, \infty)$  for each  $p \in B_\rho(p_0)$  and  $q \in \partial Q$ , where  $\rho > 0$  is sufficiently small. Moreover,  $\log(a(p))$  is subharmonic in  $R^2 \setminus Q$  and  $a(p) \geq a > 0$  in  $R^2$ ,  $a$  a constant.

(A3)  $a(p) = a(x, y)$  is a real analytic function of each coordinate variable in  $R^2 \setminus Q$ .

LEMMA 4. Assume in Lemma 3 that  $Q$  and the function  $a(p)$  satisfy (A1) and (A2), and that  $\hat{c} \leq C_0$ . Then all the estimates given in [2, Lemma 7] (except [2, (19)]) apply to the regions  $\dot{\Omega}_i, i = 0, \dots, n = [A_0 / \delta]$  defined in Lemma 3.<sup>1</sup> Further, for any  $\alpha \in [0, A_0)$  we have:

$$(9) \quad \bar{d}(0, \dot{\Gamma}_i) \leq R(\alpha) < \infty$$

uniformly over all sufficiently small  $\delta > 0$  and  $i = 0, \dots, [A_0 / \delta]$ .

*Proof.* We omit most of the details, since the proofs already given in [2] (for  $Q$  convex and  $a(p) \equiv 1$ ) still apply after certain adjustments. The main difference is that  $\dot{S}_i^*$  and  $\dot{S}_i^* \cup \dot{\Omega}_i, i = 0, \dots, n$ , are no longer convex. Therefore Lemma 3(d2) replaces [2, Lemma 5(c)] in the proof of [2, (13)]. In order to prove [2, (14)–(18)] by the procedure in [2], the inequalities:  $\eta \leq |\omega \setminus S| \leq \mu \cdot \bar{d}(\gamma^*, \gamma)$  and [2, (20)] must be replaced by similar estimates based on the fact that  $\dot{S}_i^*$  and  $\dot{S}_i^* \cup \dot{\Omega}_i, i = 0, \dots, n$ , are all starlike relative to  $B_\rho(p_0)$  (as follows from (A1), (A2), and Lemma 3(c)). As replacement for the first inequality, we have  $\eta \leq |\omega \setminus S| \leq (2\pi \bar{a}^2 \bar{r}^2 \cdot \bar{d}(\gamma^*, \gamma) / \rho)$ , where  $\omega = \dot{\Omega}_i$  for some  $i, \bar{r} = \bar{d}(p_0, \Gamma), \bar{a} = \sup\{a(p) : p \in S^* \cup \Omega\}, |\cdot|$  refers to area weighted by  $a^2(p), \Phi(p) := \{p + \lambda(p - q) : q \in B_\rho(p_0), \lambda \geq 0\}$ , and  $\eta = \inf\{|\Omega \cap \Phi(p)| : p \in S^*\}$ . [2, (20)] can be replaced by the inequality:  $\varepsilon \leq (\bar{r} \delta \bar{C}_i / (\rho \bar{a}^2 N(\alpha)))$ , where  $\bar{C}_i$  and  $N(\alpha)$  are defined in [2]. Equation (9) is obtained as in the proof of [2, (18)].

LEMMA 5. In Lemma 3, if  $Q$  and the function  $a(p)$  satisfy (A1), (A2), and (A3), and if  $\hat{c} \leq C_0$ , then for any fixed  $\alpha \in [0, A_0)$  there exist constants  $\eta \in R_+$  and  $0 < \sigma \leq \tau < \infty$  such that  $\dot{F}_i(z) : R \times (0, 1) \rightarrow \dot{\Omega}_i$  can be analytically continued to  $R \times (-\eta, 1)$ , and

<sup>1</sup> Here,  $[x]$  denotes the greatest integer function.

$\sigma \leq |D_z \hat{F}_i(z)| \leq \tau$  throughout  $R \times (-\eta, \frac{3}{4})$ , both uniformly for all  $i = 0, \dots, [\alpha/\delta]$  and all sufficiently small  $\delta > 0$ .

*Proof.* That each function  $\hat{F}_i(z)$  can be analytically continued beyond the boundary  $R \times \{0\}$  follows from Lemma 3(d1) and Lewy's theorem [4]. That the analytic continuations have the asserted uniform properties can be shown essentially by applying the estimates in Lemma 4 to the proof in [4].

LEMMA 6. Assume in Lemma 3 that  $Q$  and the function  $a(p)$  satisfy (A1), (A2), and (A3), and that  $\hat{c} \leq C_0$ . Then for any fixed  $\alpha \in [0, A_0]$ , we have:

$$(10) \quad |\nabla \hat{U}_i(p) - \nabla \hat{U}_i(q)| \leq M \cdot |p - q| \text{ in } \hat{\Omega}_i(0, \frac{1}{2}),$$

$$(11) \quad \text{Curvature}(\hat{\Gamma}_i) \leq M$$

and

$$(12) \quad \text{Curvature}(\hat{\Gamma}_i^*) \leq M$$

for some constant  $M \in R_+$ , uniformly over all  $i = 0, \dots, [\alpha/\delta]$  and all sufficiently small  $\delta > 0$ .

*Proof.* Since  $\log(|D_z \hat{F}_i(z)|)$  is harmonic in  $R \times (-\eta, 1)$ , it follows using the estimate in Lemma 5 that the functions  $\nabla \log(|D_z \hat{F}_i(z)|)$ , are uniformly bounded over  $R \times (-\eta/2, 1/2)$  for all  $i = 0, \dots, [\alpha/\delta]$  and all sufficiently small  $\delta > 0$ . Equation (10) follows from this, since  $|\nabla \hat{U}_i(p)| \cdot |D_z \hat{F}_i(z)| = 1$  for any  $\hat{F}_i(z) = p \in \hat{\Omega}_i$ . Equation (11) follows by a similar argument based on Lemma 5 and the fact that the curvature of  $\hat{\Gamma}_i$  at  $p \in \hat{\Gamma}_i$  is given by  $\text{Curv}(\hat{\Gamma}_i; p) = |D_x \arg(D_z \hat{F}_i(x))|/|D_z \hat{F}_i(x)|$ , where  $p = \hat{F}_i(x)$ ,  $x \in R$ . The more detailed proof of (12) is deferred to the Appendix.

**4. Heuristic demonstration of Theorem 2(b).** For  $c \in R_+$  fixed, let  $\hat{\Omega} = R^2 \setminus (Q \cup S \cup S_c)$ , where  $\hat{c} \in (0, c]$  is chosen such that  $A_0 := |S_c \setminus S| = |S^* \setminus Q|$ . Since  $|\hat{\Omega}| = |\Omega| + |S^* \setminus Q| - |S_c \setminus S| = |\Omega| \leq |\Omega_c|$ , we conclude from [1, Thm. 12] that  $\hat{K} > K_c$  whenever  $\hat{\Omega} \neq \Omega_c$ . Therefore (5) follows if

$$(13) \quad K \geq \hat{K}.$$

For  $n \in N$  large (and for  $\delta = A_0/n$ ), we define the regions  $\hat{\Omega}_0, \dots, \hat{\Omega}_n$  such that  $\hat{S}_i = S \cup \hat{S}_i$  and  $\hat{S}_i^* = S^* \cap \hat{S}_i^*$  for each  $i$ , where  $\hat{\Omega}_0, \dots, \hat{\Omega}_n$  were defined (relative to  $\Omega_c$  and  $\Omega$ ) in Lemma 3. (See [2, Figs. 4 and 5].) Notice that  $\hat{\Omega}_0 = \hat{\Omega} (\Rightarrow \hat{K}_0 = \hat{K})$  and  $\hat{\Omega}_n = \Omega (\Rightarrow \hat{K}_n = K)$ . To demonstrate the plausibility of (13), we will argue that the values  $\hat{K}_0, \hat{K}_1, \dots, \hat{K}_n$  are essentially monotone increasing. To the extent that the Poincaré variational formula for capacity [2, (3)] is applicable, we have approximately that

$$(14) \quad \delta \hat{K} := \hat{K}_{i+1} - \hat{K}_i \approx \int_{\gamma_i^*} |\nabla \hat{U}_i(p)|^2 \cdot \delta n(p) \cdot |dp| - \int_{\gamma_i} |\nabla \hat{U}_i(p)|^2 \cdot \delta n(p) \cdot |dp|,$$

$i = 0, \dots, n-1$ , where  $\gamma_i^* = \hat{\Gamma}_i^* \cap \text{Interior}(S^*)$  and  $\gamma_i = \hat{\Gamma}_i \setminus S$ . Using Lemma 3(d1 and d2) and [2, Lemma 4], we see that  $|\nabla \hat{U}_i(p)| \leq |\nabla \hat{U}_i(p)| = c_i \cdot a(p)$  on  $\gamma_i$ , whereas  $|\nabla \hat{U}_i(p)| \geq |\nabla \hat{U}_i(p)| \geq c_i \cdot a(p)$  on  $\gamma_i^*$ . Moreover,  $|\hat{\Omega}_{i+1}| - |\hat{\Omega}_i| = |\hat{S}_i \setminus \hat{S}_{i+1}| - |\hat{S}_{i+1}^* \setminus \hat{S}_i^*| = 0$  for each  $i = 0, \dots, n-1$ , by Lemma 3(e). Thus, (14) becomes

$$(15) \quad \begin{aligned} \hat{K}_{i+1} - \hat{K}_i &\geq c_i^2 \cdot \left( \int_{\gamma_i^*} a^2(p) \cdot \delta n(p) \cdot |dp| - \int_{\gamma_i} a^2(p) \cdot \delta n(p) \cdot |dp| \right) \\ &\approx c_i^2 \cdot (|\hat{\Omega}_{i+1}| - |\hat{\Omega}_i|) = 0. \end{aligned}$$

**5. Proof of Theorem 2(b).** As was shown in § 4, it suffices to prove (13). However, the argument in § 4 leading to (13) is incomplete. In fact the harmonic measure  $\hat{U}_i(p)$  of  $\hat{\Gamma}_i^*$  in  $\hat{\Omega}_i$  need not exist, since  $\hat{S}_i^* = S^* \cap \hat{S}_i^*$  could contain isolated points. Further, the error in (14) and (15) has not been examined.

In this section, we will first prove (13) (and hence (5)) in the case where  $Q$  and the function  $a(p)$  satisfy (A1), (A2), and (A3), and then extend (5) to the general case. Under assumptions (A1), (A2), and (A3), the proof of (13) is very similar to the proof already given in [2, § 6] that  $K \cong \hat{K}$ . Indeed, if (in [2, § 6]) we interpret  $|\Omega|$ ,  $|S^* \setminus Q|$ , etc., to be areas weighted by  $a^2(p)$ , and if  $\Omega_c$ ,  $c \in R_+$ , and  $\hat{\Omega}_i$ ,  $i = 0, \dots, n$ , are understood to be the regions defined in Theorem 1 and Lemma 3 (instead of [2, Thm. 1 and Lemma 6]), then the proof in [2, § 6] that  $K \cong \hat{K}$  is valid up to [2, (27)] after one makes the following minor correction: Given  $\varepsilon > 0$  (in [2, (22)]), one must choose  $\alpha \in [0, A_0)$  such that  $|\Omega \cap \Phi(p)| \geq A_0 - \alpha$  for each  $p \in \hat{\Omega}_\varepsilon$ , where  $\Phi(p) = \{p + \lambda(p - q) : q \in B_\rho(p_0), \lambda \geq 0\}$ . One can then show that  $\hat{S}_i \subset \hat{S}_\varepsilon^*$  for  $\alpha \leq i \cdot \delta \leq A_0$  just as before (preceding [2, (23)]), even though  $\hat{S}_i^* \cup \hat{\Omega}_i$  is only starlike relative to  $B_\rho(p_0)$  rather than convex.

In [2, (27)], we have, using Lemma 3(d2), that

$$(16) \quad |\nabla \tilde{U}_i(p)| \geq |\nabla \dot{U}_i(p)| \geq c_i \cdot a(p) \quad \text{in } \tilde{\Omega}_i := \hat{\Omega}_i \cap \hat{\Omega}_{i+1} \quad i = 0, \dots, n - 1.$$

By applying Lemmas 3(d1) and 4, we find that

$$(17) \quad |\nabla \tilde{U}_i(p)| \leq (1 + M \cdot \delta) \cdot c_i \cdot a(p) \quad \text{on } \tilde{\Gamma}_i = \hat{\Gamma}_i, \quad i = 0, \dots, [\beta/\delta].$$

In (17), and throughout this section,  $M$  represents a (fixed, but arbitrary) finite constant which is independent of sufficiently small  $\delta > 0$ . By substituting (16) and (17) into [2, (27)], we see that [2, (28)] generalizes to

$$(18) \quad I(\tilde{\Omega}_i; W_i) \geq (c_i/\lambda) \cdot \left( \int_{\gamma_i^*(\lambda)} a(p) \cdot W_i(p) \cdot |dp| - (1 + M \cdot \delta) \cdot \int_{\tilde{\Gamma}_i} a(p) \cdot W_i(p) \cdot |dp| \right),$$

$i = 0, \dots, [\beta/\delta] - 1$ . By substituting the boundary conditions (in [2]) for  $W_i(p)$  and taking the limit as  $\lambda \rightarrow 1 - 0$ , we obtain

$$(19) \quad \Delta_i \geq c_i \cdot \left( \int_{\gamma_{i+1}^*} a(p) \cdot (1 - \dot{U}_i(p)) \cdot |dp| - (1 + M \cdot \delta) \cdot \int_{\gamma_{i+1}} a(p) \cdot \dot{U}_{i+1}(p) \cdot |dp| \right),$$

$i = 0, \dots, [\beta/\delta] - 1$ , where  $\gamma_i^*$  and  $\gamma_i$  are defined preceding [2, (26)]. Therefore, in order to prove [2, (24)], and hence [2, (22)] and (13), it suffices to show that

$$(20) \quad J_i^* := \int_{\gamma_{i+1}^*} a(p) \cdot (1 - \dot{U}_i(p)) \cdot |dp| \geq c_i \cdot \delta - M \cdot \delta^2$$

and

$$(21) \quad J_i := \int_{\gamma_i} a(p) \cdot \dot{U}_{i+1}(p) \cdot |dp| \leq c_{i+1} \cdot \delta + M \cdot \delta^2.$$

Indeed, one easily sees by substituting (20) and (21) into (19) and utilizing the estimates in Lemma 4 that

$$(22) \quad \Delta_i \geq -M \cdot \delta^2, \quad i = 0, \dots, [\beta/\delta] - 1.$$

We now prove (20). For any  $p \in \hat{\Gamma}_{i+1}^*$ ,  $i = 0, \dots, n = (A_0/\delta)$ , let  $l_i(p)$  be the (shortest) line segment which connects  $p$  to  $\hat{\Gamma}_i^*$  and is perpendicular to  $\hat{\Gamma}_{i+1}^*$  at  $p$ , and let

$\delta n_i(p)$  be the length of  $l_i(p)$ . It follows from [2, (17)] and  $\bar{S}^* \supset \dot{S}_\varepsilon^*$  (where  $\gamma_i^* = \dot{\Gamma}_i^* \cap \bar{S}$ ) that  $\sigma_i^* \supset (\dot{S}_{i+1}^* \setminus \dot{S}_i^*) \cap S^*$  and  $|\sigma_i^*| \cong |(\dot{S}_{i+1}^* \setminus \dot{S}_i^*) \cap S^*| = \delta, i = 0, \dots, [\beta/\delta] - 1$ , if  $\delta > 0$  is sufficiently small. Thus, one can show using [2, (17)], (9), (12), and (A3) (which implies the Lipschitz continuity of  $a(p)$  over any bounded set) that

$$(23) \quad \int_{\gamma_{i+1}^*} a^2(p) \cdot \delta n(p) \cdot |dp| + M \cdot \delta^2 \cong |\sigma_i^*| \cong \delta > 0, \quad i = 0, \dots, [\beta/\delta] - 1.$$

On the other hand, it follows using Lemma 3(d2) that

$$(24) \quad 1 - \dot{U}_i(p) = \int_{\tau_i(p)} |\nabla \dot{U}_i(q)| \cdot |dq| \cong c_i \int_{\tau_i(p)} a(q) \cdot |dq| \cong c_i \cdot a(p') \cdot L(\tau_i(p))$$

for any  $i = 0, \dots, n$  and  $p \in \gamma_{i+1}^*$ , where  $\tau_i(p)$  is the curve of steepest ascent of  $\dot{U}_i$  from  $p$  to  $\dot{\Gamma}_i^*$ ,  $p' \in \tau_i(p)$ , and  $L(\cdot)$  denotes arc length. One can show using [2, (17)] and (12) that

$$(25) \quad L(\tau_i(p))^* \cong \delta n_i(p) - M \cdot \delta^2 \quad \text{on } \gamma_{i+1}^*, \quad i = 0, \dots, [\beta/\delta] - 1.$$

Furthermore,  $L(\tau_i(p)) \leq M \cdot \delta$  on  $\gamma_{i+1}^*, i = 0, \dots, [\beta/\delta] - 1$ , as follows from  $a(p) \geq a$ , (24), and various estimates in Lemma 4. Therefore, after substituting (25) into (24), one can show using (9), (A3), and [2, (15)] that

$$(26) \quad 1 - \dot{U}_i(p) \cong c_i \cdot a(p) \cdot \delta n(p) - M \cdot \delta^2 \quad \text{on } \gamma_{i+1}^*, \quad i = 0, \dots, [\beta/\delta] - 1.$$

By multiplying (26) by  $a(p)$  and integrating over  $\gamma_{i+1}^*$ , we obtain

$$(27) \quad J_i^* \cong c_i \cdot \int_{\gamma_{i+1}^*} a^2(p) \cdot \delta n_i(p) \cdot |dp| - M \cdot \delta^2, \quad i = 0, \dots, [\beta/\delta] - 1.$$

Now (20) follows by combining (23) and (27) and using [2, (15)]. We omit the quite similar proof of (21).

This completes the proof of Theorem 2(b) under the additional assumptions (A1), (A2), and (A3). It remains to show that these further assumptions are superfluous. Let  $Q$  and the function  $a(p)$  have the properties assumed in Theorem 2, and let  $c \in [0, C_0] \cap R_+$ . For each  $n \in N$ , the set  $Q_n := CL(Q \cup \Omega_c(1 - 1/2n, 1))$  is starlike relative to  $B_{\rho_n}(p_0)$  for  $\rho_n > 0$  sufficiently small, as follows by applying the strong maximum principle to the function:  $\arg(D_z F_c(z)/F_c(z))$ , as in the proof of Lemma 3(c). In order to show that assumptions (A1), (A2), and (A3) are superfluous, it suffices to construct a sequence of functions  $\{a_n(p)\}$  uniformly convergent to  $a(p)$  over  $\Omega_c \cup \Omega$ , such that for each  $n \in N$ ,  $a_n(p)$  satisfies (A2) and (A3) relative to  $Q_n$ ,  $a_n(p) \leq a(p)$  on  $\partial Q_n$ , and  $a_n(p) \geq a(p)$  on  $\Gamma_c$ . Indeed, if the functions  $a_n(p), n \in N$ , exist, and if  $\omega_n$  is defined such that  $s_n^* = Q_n$  and  $|\nabla u_n(p)| = c_n \cdot a_n(p)$  on  $\gamma_n$  (where  $c_n = c/(1 - 1/2n)$ ), then it follows that  $|\nabla u_n(p)| \geq c_n \cdot a_n(p)$  on  $\gamma_n^* = \partial Q_n$ . Thus, if  $\{\bar{\omega}_n\}$  is any sequence of regions convergent to  $\Omega$  (in an appropriate sense), such that  $\bar{s}_n^* \supset Q_n$  and  $|\bar{\omega}_n|_n \leq |\omega_n|_n$  for each  $n \in N$  (where  $|\cdot|_n$  denotes area weighted by  $a_n^2(p)$ ), then  $\bar{k}_n \geq k_n$  for each  $n$ , since (5) holds under the additional assumptions. Therefore  $K - K_c = \lim_{n \rightarrow \infty} (\bar{k}_n - k_n) \geq 0$  under the assumptions of Theorem 2 only.

For the construction of the functions  $a_n(p), n \in N$ , we can assume without loss of generality that  $p_0 = 0$  and  $\lambda \cdot a(\lambda p)$  is weakly increasing in  $\lambda \in R_+$  for each  $p \in R^2$ . Let  $\psi(p)$  be a sufficiently smooth function on  $R^2$  such that  $\psi(p) < 0$  in  $Q_1, \psi(p) > 0$  outside  $Q \cup \Omega_c, \psi(\lambda p) + \log(\lambda)$  is increasing in  $\lambda \in R_+$  for all  $p \in R^2$ , and  $\nabla^2 \psi(p) \geq \delta > 0$  in  $B_{3\nu}(0)$ , where  $\nu = \sup\{|p| : p \in \Omega \cup \Omega_c\}$ . If  $a(p)$  is Lipschitz continuous in  $B_{3\nu}(0)$ , then all above requirements for the sequence  $\{a_n(p)\}$ , except that each function  $a_n(p)$  satisfy

(A3), are fulfilled by the functions  $\alpha_n(p)$ ,  $n \in N$ , defined on  $R^2$  as follows.

$$\begin{aligned} \alpha_n(p) &= \max \{a(p) \cdot \exp(\psi(p)/n), \underline{a} \cdot \exp(\mu \cdot (|p|^2 - (2\nu)^2)_+)\}, \quad |p| \leq 3\nu \\ &= \underline{a} \cdot \exp(\mu \cdot (|p|^2 - (2\nu)^2)), \quad |p| > 3\nu, \end{aligned}$$

where  $(x)_+ = \max \{x, 0\}$ ,  $\underline{a} = \inf \{a(p) : p \in B_{3\nu}(0)\}$ , and  $\mu \in R_+$  is chosen such that  $\alpha_n(p) = \underline{a} \cdot \exp(5\mu\nu^2)$  for  $|p| = 3\nu$ . In general, the functions  $a_n(p)$ ,  $n \in N$ , defined (using polar coordinates, i.e.,  $p = (r, \theta + 2m\pi)$ ,  $m = 0, \pm 1, \pm 2, \dots$ ) by

$$a_n(r, \theta) = \exp \left[ \int_{-\infty}^{\infty} \int_0^{\infty} G_n(\log(r/r'), \theta - \theta') \cdot \log(\alpha_n(r', \theta')) \cdot (dr'/r') d\theta' \right],$$

$(r, \theta) \in R_+ \times R$ , satisfy all requirements, where  $G_n(x, y) = (1/\pi\varepsilon_n) \cdot \exp(-(x^2 + y^2)/\varepsilon_n)$ ,  $\varepsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ , and each  $\varepsilon_n > 0$  is sufficiently small.

**Appendix: The proof of (12).** For  $\delta > 0$  fixed (in Lemma 3), we define  $\beta_i(y) = \max \{|D_x \arg(D_z \dot{F}_i(x + jy))| : x \in R\}$ ,  $i = 0, 1, \dots, [A_0/\delta]$ . The functions  $\beta_i(y)$ ,  $i = 0, 1, \dots, [A_0/\delta] - 1$ , are defined analogously relative to the  $\dot{F}_i(z)$ . The curvature of  $\dot{\Gamma}_i^*$  at  $p \in \dot{\Gamma}_i^*$  is given by

$$\begin{aligned} \text{(A.1)} \quad \text{Curv}(\dot{\Gamma}_i^*; p) &= |D_x \arg(D_z \dot{F}_i(x + j))| \cdot |\nabla \dot{U}_i(p)| \\ &= |D_x \arg(D_z \dot{F}_i(x + j))| \cdot |\nabla \dot{U}_i(p)|, \end{aligned}$$

where  $\dot{F}_i(\dot{x} + j) = p = \dot{F}_i(\dot{x} + j)$ ,  $\dot{x}, \dot{x} \in R$ . Therefore

$$\text{(A.2)} \quad \beta_i(1) \leq \mu_i \cdot \dot{\beta}_i(1),$$

where  $\mu_i = \max \{|\nabla \dot{U}_i(p)|/|\nabla \dot{U}_i(p)| : p \in \dot{\Gamma}_i^* = \dot{\Gamma}_i^*\}$ . Further, one can define  $\dot{F}_{i+1}(z) = \dot{F}_i((1 - \varepsilon_i)z)$  on  $R \times [0, 1]$  (where  $\dot{U}_i(p) = 1 - \varepsilon_i$  on  $\dot{\Gamma}_{i+1}^*$ ), implying that

$$\text{(A.3)} \quad \dot{\beta}_{i+1}(y) = (1 - \varepsilon_i) \cdot \dot{\beta}_i((1 - \varepsilon_i)y), \quad 0 \leq y \leq 1.$$

Since  $D_x \arg(D_z \dot{F}_i(z))$  is harmonic in  $R \times (0, 1)$ , we have

$$\text{(A.4)} \quad \dot{\beta}_i(1 - \varepsilon_i) \leq \varepsilon_i \cdot \dot{\beta}_i(0) + (1 - \varepsilon_i) \cdot \dot{\beta}_i(1).$$

By combining (A.2), (A.3), and (A.4), we obtain

$$\text{(A.5)} \quad \dot{\beta}_{i+1}(1) - \dot{\beta}_i(1) \leq (\mu_i - 1) \cdot \dot{\beta}_i(1) + \varepsilon_i \cdot \dot{\beta}_i(0).$$

Now, for any fixed  $\alpha \in [0, A_0)$ , there exist finite constants  $M_1, M_2$ , and  $M_3$  such that if  $\delta > 0$  is sufficiently small and  $i \in \{0, 1, \dots, [\alpha/\delta]\}$ , then  $0 \leq \mu_i - 1 \leq M_1 \cdot \delta$  and  $0 \leq \varepsilon_i \leq M_2 \cdot \delta$  (both due to Lemma 4), and  $0 \leq \dot{\beta}_i(0) \leq M_3$  (due to Lemma 5). Equation (A.5) reduces under these conditions to

$$\text{(A.6)} \quad ((\dot{\beta}_{i+1}(1) - \dot{\beta}_i(1))/\delta) \leq M_1 \cdot \dot{\beta}_i(1) + M_2 \cdot M_3, \quad i = 0, \dots, [\alpha/\delta] - 1.$$

Now (12) follows directly from (A.6) and the fact (following from (A.1)) that  $\beta_1(1) \leq (\text{Curvature}(\partial Q)/C)$ ,  $C$  defined in [2, (12)].

REFERENCES

[1] A. ACKER, *Heat flow inequalities with applications to heat flow optimization problems*, this Journal, 8 (1977), pp. 604-618.  
 [2] ———, *A free boundary optimization problem*, this Journal, 9(1978), pp. 1179-1191.



- [3] A. BEURLING, *On free-boundary problems for the Laplace equation*, Seminars on Analytic Functions, Vol. I, Institute for Advanced Study, Princeton, N.J., 1957, pp. 248–263.
- [4] H. LEWY, *A note on harmonic functions and a hydrodynamical application*, Proc. Amer. Math. Soc., 3 (1952), pp. 111–113.
- [5] D. E. TEPPER, *Free boundary problem, the starlike case*, this Journal, 6 (1975), pp. 503–505.

### ERRATA: A FREE BOUNDARY OPTIMIZATION PROBLEM\*

ANDREW ACKER†

The following misprints appear in this paper:

1. p. 1183, line 12:  $|(S_m^* \setminus Q) \cap bs^*|$  should read  $|(S_m^* \setminus Q) \cap S^*|$ .
2. p. 1188, line 19:  $j = 2, 3m, 4$  should read  $j = 2, 3, 4$ .
3. p. 1188, line 26:  $b1 - \hat{U}_i(p)$  should read  $1 - \hat{U}_i(p)$ .

---

\* This Journal, 9 (1978), pp. 1179–1191. Received by the editors July 1, 1979.

† Mathematisches Institut I, Universität Karlsruhe (TH), 75 Karlsruhe 1, Federal Republic of Germany.

- [3] A. BEURLING, *On free-boundary problems for the Laplace equation*, Seminars on Analytic Functions, Vol. I, Institute for Advanced Study, Princeton, N.J., 1957, pp. 248–263.
- [4] H. LEWY, *A note on harmonic functions and a hydrodynamical application*, Proc. Amer. Math. Soc., 3 (1952), pp. 111–113.
- [5] D. E. TEPPER, *Free boundary problem, the starlike case*, this Journal, 6 (1975), pp. 503–505.

### ERRATA: A FREE BOUNDARY OPTIMIZATION PROBLEM\*

ANDREW ACKER†

The following misprints appear in this paper:

1. p. 1183, line 12:  $|(S_m^* \setminus Q) \cap bs^*|$  should read  $|(S_m^* \setminus Q) \cap S^*|$ .
2. p. 1188, line 19:  $j = 2, 3m, 4$  should read  $j = 2, 3, 4$ .
3. p. 1188, line 26:  $b1 - \hat{U}_i(p)$  should read  $1 - \hat{U}_i(p)$ .

---

\* This Journal, 9 (1978), pp. 1179–1191. Received by the editors July 1, 1979.

† Mathematisches Institut I, Universität Karlsruhe (TH), 75 Karlsruhe 1, Federal Republic of Germany.

## STRONG SOLUTIONS FOR INFINITE-DIMENSIONAL RICCATI EQUATIONS ARISING IN TRANSPORT THEORY\*

HENDRIK J. KUIPER† AND STEVEN M. SHEW‡

**Abstract.** The main result gives conditions under which the Riccati equation  $S'(t) = A(t)S(t) + S(t)B(t) + S(t)C(t)S(t) + D(t)$  with initial condition  $S(0) = S_0$  has a strongly differentiable solution. In addition, equations of more general form, but with more restrictive initial conditions, are shown to have solutions which are differentiable with respect to the uniform topology. These results, as well as their proofs, are discussed in the context of an important problem in transport theory.

**1. Introduction.** Let  $\mathcal{X}$  be a separable Hilbert space over the complex field and let  $\mathcal{L}(\mathcal{X})$  be the Banach algebra of bounded linear operators on  $\mathcal{X}$ . When we wish to distinguish between the uniform, strong, and weak operator topologies on  $\mathcal{L}(\mathcal{X})$  we do so by using subscripts:  $\mathcal{L}(\mathcal{X})_u$ ,  $\mathcal{L}(\mathcal{X})_s$ , and  $\mathcal{L}(\mathcal{X})_w$ . The Riccati initial value problem

$$(1) \quad \begin{aligned} S' &= A(t)S + SB(t) + SC(t)S + D(t), & 0 < t < T, \\ S(0) &= S_0 \in \mathcal{L}(\mathcal{X}), \end{aligned}$$

where  $S$  is an  $\mathcal{L}(\mathcal{X})$ -valued function (or distribution), arises in certain problems in transport theory as well as in optimal control theory. In such settings the operators  $A(t)$  and  $B(t)$  are often unbounded closed operators rather than bounded operators on  $\mathcal{X}$ . This complicates questions of existence and uniqueness. Indeed, one must first decide as to what will constitute a solution. We call  $S$  a *distributional solution* if (1) is interpreted as an equation of vector-valued distributions and  $S'$  is the distributional derivative of  $S$ . When the derivative is to be interpreted in the uniform (resp. strong, resp. weak) sense we call  $S$  a *uniform* (resp. *strong*, resp. *weak*) *solution*. Unless one puts restrictive conditions on the initial value  $S_0$  (see e.g., [14]) it is pointless to look for uniform solutions. We note that even the very simple equation  $S' = AS$ ,  $S(0) = I$ , does not have a uniform solution unless  $A$  is bounded.

Existence theorems for infinite-dimensional Riccati equations with unbounded coefficients have been obtained by various authors such as Da Prato, Lions, Lukes, Russell and others (e.g., [2], [3], [8]). The earliest work seems to be due to Lions [7] who via a theorem on the existence of an optimal control was led to an existence result for what essentially are distributional solutions of (1). Tartar [13] extended these results to more general equations of the form

$$(2) \quad \begin{aligned} S' &= AS + SB + \Phi(S), \\ S(0) &= S_0 \end{aligned}$$

and also obtained many qualitative results including certain regularity results, such as strong continuity from the right, and a priori estimates. At about the same time Curtain and Pritchard [1] obtained existence of weak solutions for (1). In another vein Temam [14], by restricting himself to the space  $\mathcal{H}_s$  of Hilbert-Schmidt operators with its natural Hilbert space structure, was able to obtain existence of distributional solutions by employing a constructive approximation procedure. He also obtained regularity results, for example showing that  $S' \in L_\infty((0, T), \mathcal{H}_s)$ .

\* Received by the editors August 21, 1978 and in revised form January 30, 1979.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85281.

‡ Department of Mathematics, University of Houston, Houston, Texas 77002.

Our first result will concern uniform solutions of (2) within the subspace  $\mathcal{H} \subset \mathcal{L}(\mathcal{H})$  of compact linear operators, thus extending certain results of Temam. We then proceed to prove existence of strong solutions for (1). This strengthens the results of Curtain and Pritchard. Since our method of proof seems to blend very nicely with the physical aspects of certain problems in transport theory we feel it is important for us to say a few words about this topic. For more information the reader is referred to [9], [11], [15], [16].

**2. Transport theory.** Consider a slab of material which lies perpendicular to the  $\xi$ -axis and extends from  $\xi = x$  to  $\xi = y > x$ . Suppose this slab is subjected to an input (e.g., a flux of radiation)  $I_+(x)$  on the left and an input  $I_-(y)$  on the right. This results in an output  $I_-(x)$  on the left and an output  $I_+(y)$  on the right. We assume that  $I_{\pm}(x)$  and  $I_{\pm}(y)$  can be regarded as members of some Hilbert space  $\mathcal{H}$ , and that the transport properties of the slab can be described by an operator  $S(x, y) \in \mathcal{L}(\mathcal{H} \oplus \mathcal{H})$ :

$$\begin{bmatrix} I_+(y) \\ I_-(x) \end{bmatrix} = S(x, y) \begin{bmatrix} I_+(x) \\ I_-(y) \end{bmatrix} \equiv \begin{bmatrix} t(x, y) & \rho(x, y) \\ r(x, y) & \tau(x, y) \end{bmatrix} \begin{bmatrix} I_+(x) \\ I_-(y) \end{bmatrix}.$$

Here  $t(x, y), \tau(x, y), \rho(x, y), r(x, y) \in \mathcal{L}(\mathcal{H})$  and are called, respectively, forward transmission, backward transmission, forward reflection and backward reflection operators. One can show that for  $x < y < z$ , *cascading* of  $S(x, y)$  and  $S(y, z)$  (i.e., putting two slabs into physical contact at  $\xi = y$ ) yields:

$$S(y, z) = S(x, y) * S(y, z),$$

where  $*$  is a product introduced by Redheffer [10] and is defined by

$$\begin{bmatrix} t_1 & \rho_1 \\ r_1 & \tau_1 \end{bmatrix} * \begin{bmatrix} t_2 & \rho_2 \\ r_2 & \tau_2 \end{bmatrix} = \begin{bmatrix} t_2(I - \rho_1 r_2)^{-1} t_1 & \rho_2 + t_2 \rho_1 (I - r_2 \rho_1)^{-1} \tau_2 \\ r_1 + \tau_1 r_2 (I - \rho_1 r_2)^{-1} t_1 & \tau_1 (I - r_2 \rho_1)^{-1} \tau_2 \end{bmatrix}.$$

Clearly this product does not always exist. The physical interpretation for this is that the juxtaposition of two slabs may produce an amount of material in excess of “critical mass” or “critical length”.

Assuming  $S(x, y)$  to be differentiable with respect to  $y$  in some sense one can obtain the equation

$$(3) \quad S_y(x, y) = A(y)S(x, y) + S(x, y)B(y) + S(x, y)C(y)S(x, y) + D(y),$$

where  $S(x, x) = I$ , and

$$A = \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ R & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & P \\ 0 & 0 \end{bmatrix}$$

and

$$\begin{aligned} T(y) &= t_y(x, y)|_{x=y}, & R(y) &= r_y(x, y)|_{x=y}, \\ Q(y) &= \tau_y(x, y)|_{x=y}, & P(y) &= \rho_y(x, y)|_{x=y}. \end{aligned}$$

There are various equivalent ways to linearize this problem, one of which is to essentially carry out the optimal control theoretic methods (this is intimately related to the definition of solution in [13]). Another method, which applies in the above problem, is to transform  $S(x, y)$  into a new operator  $\tilde{S}(x, y)$  which takes the left input and output and maps them into the right input and output. Of course  $\tilde{S}$  is not always well defined,

but when it is its graph is obtained from that of  $S$  merely by interchanging second and fourth entries. We have ([9], [11]),  $\overline{S_1 * S_2} = \overline{S_2} \overline{S_1}$ ,  $\overline{\overline{S}} = S$ , and

$$\overline{S}_y(x, y) = \begin{bmatrix} T(y) & P(y) \\ R(y) & Q(y) \end{bmatrix} \overline{S}(x, y).$$

Under certain reasonable assumptions  $\overline{S}(x, y)$  is, for fixed  $x$ , a strongly continuous semigroup of linear operators. However, since the transformation  $\rightarrow$  does not preserve strong continuity or strong differentiability it is of no use in obtaining strong solutions for (1). We however note that our results can be rephrased to say that  $S(x, y)$ , for fixed  $x$ , is locally a strongly continuous semigroup of operators with respect to the  $*$ -product.

There is an interesting connection between our method of handling the existence question and certain physical situations (e.g., a slab geometry for equiflux surfaces in the transport medium) where the transport operators  $S(t)$  can be decomposed,

$$S = S_1 + S_2 = \begin{bmatrix} t_1 & 0 \\ 0 & \tau_1 \end{bmatrix} + \begin{bmatrix} t_2 & \rho \\ r & \tau_2 \end{bmatrix},$$

into their *specular*,  $S_1$ , and *diffuse*,  $S_2$ , parts ([11], [15], [16]). An investigation of the  $*$ -product shows that the specular part of  $S(x, z)$  is simply the ordinary composition of the linear operators  $S_1(x, y)$  and  $S_1(y, z)$ , and that  $S_1(x, x)$  is the identity on  $\mathcal{H} \oplus \mathcal{H}$ :

$$(4) \quad S_1(x, z) = S_1(x, y)S_1(y, z), \quad S_1(x, x) = I.$$

Letting

$$A_i = \begin{bmatrix} a_i & 0 \\ 0 & 0 \end{bmatrix}, \quad B_i = \begin{bmatrix} 0 & 0 \\ 0 & b_i \end{bmatrix}, \quad i = 1, 2,$$

where  $a_i(x) = \lim_{h \downarrow 0} (t_i(x, x+h) - I)/h$ ,  $b_i = \lim_{h \downarrow 0} (\tau_i(x, x+h) - I)/h$  we may write (3) as

$$S' = (A_1 + A_2)S + S(B_1 + B_2) + SCS + D, \quad S(0) = I,$$

where  $A_2, B_2, C$  and  $D$  can be expected to be compact (see e.g., [11]). In this situation one can also show that the diffuse part,  $S_2$ , satisfies a Riccati initial value problem of the form

$$S'_2 = E(t) + F(t)S_2 + S_2G(t) + S_2CS_2, \quad S_2(0) = 0,$$

where  $E = D + A_2S_1 + S_1B_2 + S_1CS_1$ ,  $F = A + S_1C$ ,  $G = B + CS_1$  with  $E$  and  $C$  being compact. Now (4) and physical considerations suggest that the operators  $A_1$  and  $B_1$  (and hence also  $A$  and  $B$ ) should be generators of strongly continuous semigroups of contractions on  $\mathcal{H} \oplus \mathcal{H}$ . We note that the hypotheses of our main result, Theorem 8, are in very close agreement with those dictated by the above physical considerations. Finally, our main results, Theorems 5 and 8, show that the solution to the above transport problem is equal to the sum of a strongly differentiable specular part and a uniformly differentiable diffuse part provided the hypotheses of Theorem 8 are satisfied.

**3. Preliminary results.** We recall that the *numerical range* of a linear operator  $L$  is defined to be the set

$$\{z \in \mathbb{C} | z = \langle Lx, x \rangle, x \in \mathcal{D}(L), \|x\| = 1\},$$

where  $\mathcal{D}(\cdot)$  denotes the domain and  $\langle \cdot, \cdot \rangle$  is the inner product on  $\mathcal{H}$ . We shall assume throughout this section that  $A$  and  $B$  are two densely defined closed linear operators on

$\mathcal{X}$  whose numerical ranges are contained in the wedge

$$\Sigma \equiv \left\{ z \in C \mid |\arg(z - \omega)| \geq \frac{\pi}{2} + \delta \right\},$$

where  $\omega$  and  $\delta$  are positive constants. We note that the adjoints  $A^*$  and  $B^*$  are densely defined closed linear operators. For each closed subspace  $\mathcal{E} \subset \mathcal{X}$ ,  $\pi_{\mathcal{E}}$  will denote the orthogonal projection onto  $\mathcal{E}$ . If it happens that  $\mathcal{E} \subset \mathcal{D}(A)$  then, by the closed graph theorem,  $A\pi_{\mathcal{E}}$  is a bounded linear operator on  $\mathcal{X}$ . Similarly if  $\mathcal{F}$  is a closed subspace contained in  $\mathcal{D}(B^*)$  then  $\pi_{\mathcal{F}}B$ , the closure of the operator  $\pi_{\mathcal{F}}B$ , is equal to  $(B^*\pi_{\mathcal{F}})^*$ , and is a bounded linear operator on  $\mathcal{X}$ . On the subspace  $\mathcal{H} \subset \mathcal{L}(\mathcal{X})$  of compact linear operators, we define the linear operator

$$\mathbf{A}: \mathcal{D}_0 \subset \mathcal{H} \rightarrow \mathcal{H}$$

with domain

$$\mathcal{D}_0 = \{S = \pi_{\mathcal{E}}S\pi_{\mathcal{F}} \mid \mathcal{E}, \mathcal{F} \text{ finite dimensional subspaces of } \mathcal{X} \text{ with } \mathcal{E} \subset \mathcal{D}(A), \mathcal{F} \subset \mathcal{D}(B^*)\},$$

by

$$\mathbf{A}(S) = AS + \overline{SB}.$$

LEMMA 1. *The closure of  $\mathcal{D}_0$  in  $\mathcal{L}(\mathcal{X})_u$  is  $\mathcal{H}$ , and  $\mathbf{A}$  is closable.*

*Proof.* Suppose  $\mathcal{E}$  is a finite dimensional subspace with orthonormal basis  $\{e_1, e_2, \dots, e_n\}$ . For  $1 \leq i \leq n$  suppose we have sequences  $\{e_i^k\}_{k=1}^\infty$  converging to  $e_i$ . Let  $\pi_k$  be the orthogonal projection onto the subspace spanned by  $\{e_1^k, e_2^k, \dots, e_n^k\}$ . Then  $\pi_k \rightarrow \pi_{\mathcal{E}}$  in the uniform topology. Let  $\varepsilon > 0$  and  $L$  be an arbitrary member of  $\mathcal{H}$ . There exist two projection operators  $\pi_1$  and  $\pi_2$ , with finite dimensional range, such that  $\|L - \pi_1L\pi_2\| < \varepsilon/3$ . Since  $\mathcal{D}(A)$  and  $\mathcal{D}(B^*)$  are dense in  $\mathcal{X}$  we can find finite dimensional subspaces  $\mathcal{E} \subset \mathcal{D}(A)$  and  $\mathcal{F} \subset \mathcal{D}(B^*)$  such that  $\|\pi_1 - \pi_{\mathcal{E}}\| + \|\pi_2 - \pi_{\mathcal{F}}\| < \varepsilon/3\|L\|$ . Hence we have  $\|L - \pi_{\mathcal{E}}L\pi_{\mathcal{F}}\| < \varepsilon$ . Next we show that  $\mathbf{A}$  is closable. Suppose  $\{L_n\}$  is a sequence in  $\mathcal{D}_0$  with  $\lim_{n \rightarrow \infty} \|L_n\| = 0$  and  $\mathbf{A}(L_n)$  tending to  $T \in \mathcal{H}$ . It suffices to show  $T = 0$ . To this end let  $x$  be an arbitrary element in  $\mathcal{D}(B)$ , then  $\lim_{n \rightarrow \infty} \overline{L_n B}x = \lim_{n \rightarrow \infty} L_n Bx = 0$ . Since  $\mathbf{A}(L_n) = AL_n + \overline{L_n B}$  we see that  $\lim_{n \rightarrow \infty} AL_n x = Tx$ . However since the sequence  $\{L_n x\}$  tends to zero and  $A$  is a closed operator we see that  $Tx = 0$  for an arbitrary  $x \in \mathcal{D}(B)$  and therefore  $T = 0$ .

We shall henceforth use the symbol  $\mathbf{A}$  to denote the closure of the previously defined operator  $\mathbf{A}$ . The domain of the closure will be denoted by  $\overline{\mathcal{D}}$ :

$$\mathbf{A} = \overline{\mathbf{A}}: \overline{\mathcal{D}} \subset \mathcal{H} \rightarrow \mathcal{H}, \quad \overline{\mathcal{D}} \subset \mathcal{H}.$$

*Remark.* When  $K \in \mathcal{D}$  but  $K \notin \mathcal{D}_0$  then it is probably not true in general that  $\mathbf{A}(K) = AK + \overline{KB}$ . We can however show that  $\mathbf{A}(K) = \overline{AK + KB}$ . Let  $\{K_n\}$  be a sequence in  $\mathcal{D}_0$  such that  $K_n \rightarrow K \in \mathcal{D}$  and  $AK_n + K_n B \rightarrow \mathbf{A}(K)$  in  $\mathcal{L}(\mathcal{X})_u$ . If  $x \in \mathcal{D}(B)$ , the domain of  $B$ , then  $AK_n(x)$  converges to  $\mathbf{A}(K)x - KBx$  while  $K_n x$  converges to  $Kx$ . Therefore, since  $A$  is a closed operator,  $Kx \in \mathcal{D}(A)$  and,  $AKx = \mathbf{A}(K)x - KBx$ . Hence  $\mathbf{A}(K)x = \overline{AKx + KBx}$  for all  $x$  in the dense subset  $\mathcal{D}(B)$  of  $\mathcal{X}$ , and consequently  $\mathbf{A}(K) = \overline{AK + KB}$ . In what follows we shall, for the sake of simplicity of notation, omit the overlining, thus leaving the proper interpretation to the reader.

We use  $I$  to denote the identity operator on  $\mathcal{X}$  and  $\mathbf{I}$  to denote the identity operator on  $\mathcal{L}(\mathcal{X})$  or  $\mathcal{H}$ . The next lemma shows that  $2\omega\mathbf{I} - \mathbf{A}$  is an accretive operator on  $\mathcal{H}$ .

LEMMA 2. For any  $S \in \mathcal{K}$  we have

- (i)  $\|S - \lambda \mathbf{A}(S)\| \geq (1 - 2\lambda\omega)\|S\|, \forall \lambda \in [0, 1/2\omega),$
- (ii)  $\|(\lambda \mathbf{I} - [\mathbf{A} - (2\omega + 1)\mathbf{I}])(S)\| \geq \frac{1 + |\lambda|}{C}\|S\|,$  for some constant  $C > 0$  and all complex  $\lambda$  with  $\text{Re } \lambda \geq 0.$

*Proof.* Let  $S = UP$  be the polar decomposition of  $S$  where  $S \in \mathcal{D}_0, P = (S^*S)^{1/2},$  and  $U$  is a partial isometry. We recall that  $U^*U$  is the orthogonal projection on the closure of the range of  $P.$  Since  $S$  is compact so is  $P$  and therefore there exists a vector  $x$  such that  $\|x\| = 1$  and  $Px = px$  where  $p = \|P\|.$  We therefore have  $Sx = pUx$  and  $\|S\| = p.$  Also, since  $x$  lies in the range of  $P,$  it must be true that  $U^*Ux = x.$  We now have

$$\begin{aligned} \|S - \lambda AS - \lambda \overline{SB}\| &\geq |\langle Sx - \lambda ASx - \lambda \overline{SB}x, Ux \rangle| \\ &= |\langle UPx, Ux \rangle - \lambda \langle AUPx, Ux \rangle - \lambda \langle x, B^*PU^*Ux \rangle| \\ &= |\langle U^*Ux, x \rangle - \lambda \langle AUx, Ux \rangle - \lambda \langle x, B^*x \rangle|p \\ &\geq (1 - 2\lambda\omega)\|S\|. \end{aligned}$$

A similar calculation can be carried out for  $\|(\lambda + 2\omega + 1)S - \mathbf{A}(S)\|$  with  $\text{Re } \lambda \geq 0.$  This quantity is then seen to exceed

$$(5) \quad |(\lambda + 2\omega + 1) - \langle AUx, Ux \rangle - \langle Bx, x \rangle| \|S\|.$$

Let

$$\Sigma_1 = \left\{ z \in \mathbb{C} \mid |\arg(z - 2\omega)| \geq \frac{\pi}{2} + \delta \right\},$$

then  $\Sigma + \Sigma = \Sigma_1$  and we deduce from (5) that

$$\|(\lambda + 2\omega + 1)S - \mathbf{A}(S)\| \geq \|S\| \times \text{dist}(\lambda, \Sigma_1 - 2\omega - 1).$$

A straightforward calculation shows that, letting  $\hat{\delta} = \min(\delta, \pi/4),$

$$\text{dist}(\lambda, \Sigma_1 - 2\omega - 1) \geq (1 + |\lambda|) \sin \hat{\delta}$$

when  $\text{Re } \lambda \geq 0.$  Letting  $C = 1/\sin \hat{\delta}$  we are done.

DEFINITION.  $\mathbf{A}_\lambda = \mathbf{A} - (\lambda + 2\omega + 1)\mathbf{I}.$

In order to prove the main result of this section we use some basic analytic semigroup theory (see e.g., [4], [5], or [17]) and the following well-known theorem ([4, p. 626]).

HILLE-YOSIDA THEOREM. Let  $C$  be a densely defined closed linear operator on a Banach space. A necessary and sufficient condition for  $C$  to be the infinitesimal generator of a strongly continuous semigroup of contraction operators is that  $\|(\lambda I - C)^{-1}\| < 1/\lambda$  for all  $\lambda > 0.$

This theorem can be stated differently by replacing the condition by:  $\|(\lambda I - C)^{-1}\| < 1/\text{Re } \lambda$  for all  $\lambda$  with  $\text{Re } \lambda > 0.$  This follows directly from the fact that  $C$  generates a strongly continuous semigroup of contraction operators if and only if  $C + i\mu I$  generates such a semigroup for any real  $\mu.$

LEMMA 3. Whenever  $\text{Re } \lambda \geq 0, \mathbf{A}_\lambda^{-1}$  exists as a bounded linear transformation on the space  $\mathcal{K}$  of compact linear operators. Moreover  $\|\mathbf{A}_\lambda^{-1}\| \leq C(1 + |\lambda|)^{-1}$  for some constant  $C > 0.$

*Proof.* Consider the map  $K \rightarrow AK - \omega K$  with domain  $\mathcal{D}_0.$  By Lemma 1 we can form the closure of this operator in  $\mathcal{K}.$  We denote this closed linear operator by  $\mathbf{A}_A.$  Its

domain is dense in  $\mathcal{H}$ . Since  $A - (\omega + \lambda)I$  is invertible on  $\mathcal{X}$  for each  $\lambda > 0$ , we see that  $\lambda \mathbf{I} - \mathbf{A}_A$  is invertible on  $\mathcal{X}$  and

$$(\lambda \mathbf{I} - \mathbf{A}_A)^{-1}(L) = [(\lambda + \omega)I - A]^{-1}L, \quad \lambda > 0.$$

Using the inequality

$$|\langle \{(\lambda + \omega)I - A\}x, x \rangle| \geq \lambda \langle x, x \rangle, \quad \lambda > 0$$

we see that

$$\|(\lambda \mathbf{I} - \mathbf{A}_A)^{-1}\| = \|[(\lambda + \omega)I - A]^{-1}\| \leq 1/\lambda$$

for all  $\lambda > 0$ . Applying the Hille–Yosida theorem we conclude that  $\mathbf{A}_A$  generates a strongly continuous (with respect to the uniform topology on  $\mathcal{X}$ ) semigroup of contractions. We denote this semigroup by  $\mathcal{T}_A(s)$ . Of course  $A - \omega I$  generates such a semigroup on  $\mathcal{X}$ . We denote it by  $\mathcal{T}_A(s)$ . Consider  $\psi(s) \equiv [\mathcal{T}_A(s)(K)]x - \mathcal{T}_A(s)(Kx)$ . Clearly  $\psi(0) = 0$  and

$$\begin{aligned} d\psi/ds &= \{\mathbf{A}_A \circ \mathcal{T}_A(s)(K)\}x - (A - \omega I)(\mathcal{T}_A(s)x) \\ &= (A - \omega I)\psi(s), \quad s > 0. \end{aligned}$$

Since this initial value problem has a unique solution, namely  $\mathcal{T}_A(s)\psi(0)$ , we see that  $\psi \equiv 0$  and consequently  $\mathcal{T}_A(s)(K) = \mathcal{T}_A(s)K$ . Next let us consider map  $\mathbf{A}_B: K \rightarrow KB - \omega B$  from  $\mathcal{D}_0$  into  $\mathcal{X}$ . Let  $\mathcal{T}_B(s)$  denote the semigroup on  $\mathcal{X}$  generated by  $\mathbf{A}_B$  and  $\mathcal{T}_B(s)$  the semigroup on  $\mathcal{X}$  generated by  $B - \omega I$ . Let  $\phi(s) \equiv [\mathcal{T}_B(s)(K)]^*x - \mathcal{T}_B(s)^*K^*x$ . Although the map  $L \rightarrow L^*$  from  $\mathcal{L}(\mathcal{X})_s$  into itself is not continuous, it is continuous as a map from  $\mathcal{L}(\mathcal{X})_u$  into itself. In fact  $d[\mathcal{T}_B(s)(K)]^*/ds = [\mathbf{A}_B \circ \mathcal{T}_B(s)(K)]^*$ . By a theorem of Phillips (see e.g., [6], [7]) the adjoint of a strongly continuous semigroup on a Hilbert space is another strongly continuous semigroup generated by the adjoint of the infinitesimal generator. This means that  $d\mathcal{T}_B(s)^*K^*x/ds = (B^* - \omega I)\mathcal{T}_B(s)^*K^*x$ . We therefore have

$$\begin{aligned} d\phi/ds &= [\mathbf{A}_B \circ \mathcal{T}_B(s)(K)]^*x - (B^* - \omega I)\mathcal{T}_B(s)^*K^*x \\ &= [\mathcal{T}_B(s)(K)(B - \omega I)]^*x - (B^* - \omega I)\mathcal{T}_B(s)^*K^*x \\ &= (B^* - \omega I)\phi(s), \quad s > 0. \end{aligned}$$

Since  $\phi(0) = 0$  we have  $\phi \equiv 0$ , hence  $\phi^* \equiv 0$ , or more explicitly  $\mathcal{T}_B(s)(K) = K\mathcal{T}_B(s)$ . Next we define

$$\mathcal{T} = \mathcal{T}_A \circ \mathcal{T}_B.$$

We note that

$$\mathcal{T}_A(s) \circ \mathcal{T}_B(s)(K) = \mathcal{T}_A(s)K\mathcal{T}_B(s) = \mathcal{T}_B(s) \circ \mathcal{T}_A(s)(K).$$

This means that  $\mathcal{T}(s)$  is a semigroup of contraction operators on  $\mathcal{X}$ . We show it is strongly continuous. The fact that  $\mathcal{T}_A$  is a strongly continuous semigroup means that  $\|\mathcal{T}_A(s)(K) - K\|$  tends to zero as  $s$  decreases to zero. This says that  $\|\mathcal{T}_A(s)K - K\|$  tends to zero as  $s$  decreases to zero. Similarly we find that  $\|K\mathcal{T}_B(s) - K\|$  tends to zero as  $s$  decreases to zero. Suppose that  $\mathcal{T}(s)$  is not strongly continuous, then there must exist a sequence of positive numbers  $\{s_i\}_{i=1}^\infty$  with  $\lim_{i \rightarrow \infty} s_i = 0$ , and sequences of unit vectors  $\{x_i\}_{i=1}^\infty$  and  $\{y_i\}_{i=1}^\infty$  and an  $\varepsilon > 0$  such that

$$\langle (\mathcal{T}_A(s_i)K\mathcal{T}_B(s_i) - K)x_i, y_i \rangle > 3\varepsilon, \quad \forall i,$$



or writing this differently,

$$(6) \quad \begin{aligned} & \langle (\mathcal{T}_A(s_i) - I)K(\mathcal{T}_B(s_i) - I)x_i, y_i \rangle \\ & + \langle (\mathcal{T}_A(s_i) - I)Kx_i, y_i \rangle \\ & + \langle K(\mathcal{T}_B(s_i) - I)x_i, y_i \rangle > 3\varepsilon, \quad \forall i. \end{aligned}$$

We may assume without loss of generality that the sequences of unit vectors converge weakly in  $\mathcal{X}$ :  $x_i \rightarrow x$  and  $y_i \rightarrow y$ . Therefore  $Kx_i \rightarrow Kx$  (strongly) and hence  $(\mathcal{T}_A(s_i) - I)Kx_i \rightarrow 0$ , so that the middle term in (6) tends to zero. Also  $K^*y_i \rightarrow K^*y$  and  $(\mathcal{T}_B(s_i))^* - I K^*y_i \rightarrow 0$ , which implies that the last term in (6) tends to zero. We may assume, taking a subsequence if necessary, that  $K(\mathcal{T}_B(s_i) - I)x_i$  converges strongly to some element in  $\mathcal{X}$ , so that  $\langle (\mathcal{T}_A(s_i) - I)K(\mathcal{T}_B(s_i) - I)x_i, y_i \rangle \rightarrow 0$ , leading to a contradiction. We have shown that  $\mathcal{T}(s)$  is a strongly continuous semigroup of contractions on  $\mathcal{X}$ . It must have a densely defined infinitesimal generator. But since we know that the derivative of  $\mathcal{T}(s)(K)$  in  $\mathcal{L}(\mathcal{X})_w$  is equal to  $(\mathbf{A} - 2\omega\mathbf{I})(K)$  for any  $K \in \mathcal{D}$ , it follows that this infinitesimal generator must be  $\mathbf{A} - 2\omega\mathbf{I}$ . Applying Hille-Yosida theorem once again, but now in the other direction, we find that  $\mathbf{A} - 2\omega\mathbf{I} - \lambda\mathbf{I}$  is surjective for any  $\lambda$  with  $\text{Re } \lambda > 0$ . Also  $\|(\mathbf{A} - (2\omega + \lambda)\mathbf{I})^{-1}\| < 1/\text{Re } \lambda$  or  $\|\mathbf{A}_\lambda^{-1}\| < (\text{Re } \lambda + 1)^{-1}$ . By Lemma 2 we get the inequality  $\|\mathbf{A}_\lambda^{-1}\| \leq C(1 + |\lambda|)^{-1}$ .

**4. Existence theorems.** We shall first investigate the existence of uniform solutions to the generalized Riccati equation

$$(7) \quad \begin{aligned} U' &= A(t, U(t))U(t) + U(t)B(t, U(t)) + \mathcal{F}(t, U(t)), \\ U(0) &= S_0 \in \mathcal{X}. \end{aligned}$$

We assume  $\mathcal{F}: \mathcal{X} \rightarrow \mathcal{X}$  and

(I) For each  $t \in [0, t_0)$  and  $U \in \mathcal{X}$  with  $\|U\| \leq r$ , the linear operators  $A(t, U)$  and  $B(t, U)$  are closed linear operators which have their numerical ranges contained in  $\Sigma$ . By Lemma 1 we therefore have a corresponding closed linear map  $\mathbf{A}(t, U)$  on a domain  $\mathcal{D}(t, U)$  which is dense in  $\mathcal{X}$ . Letting  $\mathbf{A}_\lambda(t, U) = \mathbf{A}(t, U) - (\lambda + 2\omega + 1)\mathbf{I}$  we have by Lemma 3 that  $\mathbf{A}_\lambda(t, U)^{-1}$  exists as bounded linear transformation on  $\mathcal{X}$  provided  $\text{Re } \lambda \geq 0$ . Moreover we have

$$(8) \quad \|\mathbf{A}_\lambda(t, U)^{-1}\| \leq C(1 + |\lambda|)^{-1}.$$

The dependence of  $\mathbf{A}(t, U)$  and  $\mathcal{F}(t, U)$  on  $t$  and  $U$  will be required to conform to the following conditions:

(II) For some  $r > 0$  there exist constants  $C(r)$  and  $\sigma > 0$  so that for all  $0 \leq t, \tau \leq t_0$  and all  $K_1, K_2 \in \mathcal{X}$  with  $\|K_1\| \leq r, \|K_2\| \leq r$ :

- (i)  $\|\mathcal{F}(t, K_1) - \mathcal{F}(\tau, K_2)\| \leq C(r)\{|t - \tau|^\sigma + \|K_1 - K_2\|\}$ ,
- (ii) The intersection of the domains of  $\mathbf{A}_0(t, K_1)$  and  $\mathbf{A}_0(\tau, K_2)$  is dense in  $\mathcal{X}$  and  $\|[\mathbf{A}_0(t, K_1) - \mathbf{A}_0(\tau, K_2)]\mathbf{A}_0(\tau, K_2)^{-1}\| \leq C(r)\{|t - \tau|^\sigma + \|K_1 - K_2\|\}$ ,
- (iii)  $\| [A(t, K_1) - A(\tau, K_2)][A(0, 0) - (\omega + \frac{1}{2})I]^{-1} \| + \| [B(0, 0) - (\omega + \frac{1}{2})I]^{-1} [B(t, K_1) - B(\tau, K_2)] \| \leq C(r)[|t - \tau|^\sigma + \|K_1 - K_2\|]$ .

We note that (ii) and (iii) are satisfied if  $A(t, U) = A(0, 0) + \alpha(t, U)$  and  $B(t, U) = B(0, 0) + \beta(t, U)$  where  $\alpha$  and  $\beta$  are members of  $C([0, t_0] \times \mathcal{X}, \mathcal{X})$  and are locally Lipschitz, with respect to the uniform topology on  $\mathcal{X}$ . Moreover it can be shown that (ii) and (iii) are still satisfied if we replace the arguments  $(\tau, K_2)$  in  $\mathbf{A}_0(\tau, K_2)^{-1}$  by  $(s, K)$  for any  $s \in [0, t_0]$  and  $K \in \mathcal{X}$  with  $\|K\| < r$  ([5], [12]).

Letting  $S = U \exp -(2\omega + 1)t$ , (7) becomes

$$(9) \quad dS/dt = \tilde{\mathbf{A}}_0(t, S)(S) + \tilde{\mathcal{F}}(t, S),$$

where  $\tilde{\mathcal{F}}(t, s) \equiv [\exp - (2\omega + 1)t] \mathcal{F}(t, [\exp (2\omega + 1)t] S)$  and  $\tilde{\mathbf{A}}_0(t, S) \equiv \mathbf{A}_0(t, [\exp (2\omega + 1)t] S)$  are operators which also satisfy (I) and (II). We assume a somewhat more restrictive initial condition than was stated above, namely

$$(10) \quad \mathbf{S}(0) = U(0) = S_0 \in \mathcal{D} \equiv \mathcal{D}(0, 0) \subset \mathcal{X}.$$

LEMMA 4. *If (I) and (II) are satisfied we also have:*

(III) *For each  $t \in [0, t_0)$  and  $K \in \mathcal{K}$  with  $\|K\| \leq r e^{-(2\omega+1)t_0}$ ,  $\mathbf{A}_0(t, K)$  is a closed linear operator in  $\mathcal{X}$  with a domain  $\mathcal{D}$  which is dense in  $\mathcal{X}$  and independent of  $t$  and  $K$ . Moreover,  $\|(\tilde{\mathbf{A}}_0(t, K) - \lambda \mathbf{I})^{-1}\| \leq C(1 + |\lambda|)^{-1}$  for all  $\lambda$  with  $\text{Re } \lambda \geq 0$ .*

*Proof.* For each fixed  $t$  and  $K$ ,  $\mathbf{A}_0$  satisfies the hypotheses of the previous section which implies that it is a closed linear operator which satisfies the inequality in the statement of this lemma. To show that the domain is independent of  $t$  and  $K$  we merely note that by (II-ii)  $\tilde{\mathbf{A}}_0(t, K_1) \tilde{\mathbf{A}}_0(\tau, K_2)^{-1}$  is a densely defined bounded linear operator. Using this and the fact that  $\tilde{\mathbf{A}}_0(t, K_1)$  and  $\tilde{\mathbf{A}}_0(\tau, K_2)$  are closed, one easily verifies that  $\tilde{\mathbf{A}}_0(\tau, K_2) \subset \tilde{\mathbf{A}}_0(t, K_1)$ . The reverse inclusion follows similarly. We now apply a result due to Sobolevskii (see [12, § 5] or [5, p. 170]) which states that (9)–(10) has a unique local solution provided (II-i), (II-ii) and (III) hold and  $\|S_0\| < r e^{-(2\omega+1)t_0}$ . If (II) and (III) are satisfied for arbitrarily large  $r > 0$  we can obtain the existence of a global solution. These results are contained in the following theorem.

THEOREM 5. *Suppose (I) and (II) are satisfied for some  $r > 0$ . Then (7) has a unique solution  $U \in C^1([0, t_1), \mathcal{X})$  for some  $t_1 > 0$ , provided  $S_0 \in \mathcal{D}$  with  $\|S_0\| < r e^{-(2\omega+1)t_0}$ . If (I) and (II) are satisfied for all  $r > 0$ , then for each  $S_0 \in \mathcal{D}$  there exists a  $t^*$ ,  $0 < t^* \leq t_0$ , such that (7) has a unique solution  $U \in C^1([0, t^*), \mathcal{X})$  with  $t^* = t_0 \Leftrightarrow \infty$  or  $\limsup_{t \uparrow t^*} \gamma(t) = \infty$ , where*

$$\gamma(t) = \|\mathbf{A}_0(t, U(t))U(t) + U(t)\mathbf{B}_0(t, U(t))\|,$$

*i.e.,  $[0, t^*)$  is the maximal right open interval on which one has a uniform solution of class  $C^1$ . (See [3] where stronger regularity is obtained for a similar problem with  $A(t, U) = B(t, U)^* = A$ , independent of  $t$  and  $U$ ).*

*Proof.* The local existence follows from the remarks above. Let  $[0, t^*)$  be the largest right open interval on which one has a unique  $C^1$ -solution. If  $t^* < t_0$  and  $\gamma(t)$  remains uniformly bounded as  $t$  increases to  $t^*$  then the proof of Sobolevskii's result shows that the solution can be extended to  $[0, t^*]$  (see Thm. 16.5, p. 175 of [5]). However, by the local existence result, one can extend the solution to a somewhat larger interval  $[0, t^{**})$  with  $t^* < t^{**} < t_0$ , contradicting maximality.

Next consider the Riccati equation

$$(11) \quad \begin{aligned} dS/dt &= \mathbf{A}_0(t)(S) + SC(t)S + D(t), \\ S(0) &= S_0 \in \mathcal{L}(\mathcal{X}). \end{aligned}$$

We assume  $C$  and  $D$  are Hölder continuous:

$$(IV) \quad \begin{aligned} C &\in C^\sigma([0, t_0), \mathcal{L}(\mathcal{X})), \\ D &\in C^\sigma([0, t_0), \mathcal{X}) \quad \text{for some } 0 < \sigma < 1. \end{aligned}$$

Let  $A_0(t) = A(t) - (\omega + 1/2)I$ ,  $B_0(t) = B(t) - (\omega + 1/2)I$ .

$$dS/dt = A_0(t)S + SB_0(t) + SC(t)S + D(t).$$

By the above theorem (11) has a uniform solution  $T(t)$  on  $[0, t^*)$  corresponding to an

initial condition  $T(0) = 0$ . Let  $R \equiv S - T$ , then

$$dR/dt = [A_0(t) + T(t)C(t)]R + R[B_0(t) + C(t)T(t)] + RC(t)R,$$

which we write as

$$dR/dt = A_1(t)R + RB_1(t) + RC(t)R.$$

We may assume without loss of generality that  $A_1(t)$  and  $B_1(t)$  have numerical range in  $\Sigma - \omega - \frac{1}{2}$  and hence we can find propagation operators for the infinitesimal generators  $A_1$  and  $B_1^*$  on  $\mathcal{X}$  ([5], [12]) which are strongly differentiable in the region  $t \geq s$ :

$$\frac{\partial}{\partial t} U(t, s)x = A_1(t)U(t, s)x,$$

$$\frac{\partial}{\partial t} V(t, s)x = B_1^*(t)V(t, s)x$$

and  $V(s, s) = U(s, s) = I$ .

LEMMA 6.  $(\partial/\partial t), V(t, s)^*x = \overline{V(t, s)^*B(t)x}$  for every  $x \in \mathcal{X}$ , and  $t > s$ .

Proof. When  $t \geq s$  and  $t + h \geq s$  we have

$$\begin{aligned} V(t+h, s) - V(t, s) &= \int_t^{t+h} B_1^*(\tau)V(\tau, s) d\tau \\ &= \int_t^{t+h} [B_1^*(\tau) - B_1^*(t)]B_1^*(\tau)^{-1}[B_1^*(\tau)V(\tau, s)] d\tau \\ &\quad + \int_t^{t+h} [B_1^*(t)B_1^*(\tau)^{-1}][B_1^*(\tau)V(\tau, s)] d\tau \\ &= B_1^*(t)^* \int_t^{t+h} V(\tau, s) d\tau + O(|h|^{1+\sigma}), \end{aligned}$$

where we have used the following facts:

- (i)  $\|[B_1^*(\tau) - B_1^*(t)]B_1^*(\tau)^{-1}\| \leq \text{const.} \times |t - \tau|^\sigma$  (see II),
- (ii)  $B_1^*(t)B_1^*(\tau)^{-1}$  is uniformly bounded (see (i)),
- (iii)  $B_1^*(t)$  is a closed operator and hence commutes with integration.

We note that the integrals converge in  $\mathcal{L}(\mathcal{X})$ , while  $O(|h|^{1+\sigma})$  refers to the magnitude of the norm. We obtain for each  $x \in \mathcal{D}(B_1(t))$

$$\frac{1}{h}[V^*(t+h, s) - V^*(t, s)]x = \frac{1}{h} \left[ \int_t^{t+h} V(\tau, s) d\tau \right]^* B_1(t)x + O(|h|^\sigma).$$

Since, for  $s \leq \tau \leq t$ ,  $\|V(t, s) - V(\tau, s)\| \leq \sup_{\tau \leq \sigma \leq t} \|B_1^*(\sigma)V(\sigma, s)\|(t - \tau) \leq \text{const.} \times (t - \tau)(\tau - s)^{-1}$  [5, p. 127], we see that the operators on the left hand side are uniformly bounded for sufficiently small  $h$ . Taking the limit as  $h \rightarrow 0$  we obtain the desired result on  $\mathcal{D}(B_1(t)) = \mathcal{D}(B_1(0))$ . Since  $\overline{V(t, s)^*B_1(t)} = [B_1^*V(t, s)]^*$ , a bounded linear operator, we obtain the desired result on all of  $\mathcal{X}$  after taking the closure.

Next we define

$$P(t) = \int_0^t V(\tau, 0)^* C(\tau) U(\tau, 0) d\tau,$$

$$R(t) = \sum_{j=0}^{\infty} U(t, 0) S_0 [P(t) S_0]^j V(t, 0)^*,$$

where  $[P(t)S_0]^0$  is defined to be  $I$  for all  $t \geq 0$ . We will show that  $R$  solves (12) on some interval  $[0, t_1)$ . First we note that the series converges uniformly (in  $t$ ) in  $\mathcal{L}(\mathcal{X})_u$  for small  $t$ . This means  $R$  is strongly continuous and  $R(0) = S_0$ . If we formally differentiate we obtain

$$\begin{aligned} dR/dt &= \sum_{j=0}^{\infty} A_1(t) U(t, 0) S_0 [P(t) S_0]^j V(t, 0)^* \\ &\quad + \sum_{j=0}^{\infty} U(t, 0) S_0 [P(t) S_0]^j V(t, 0)^* B_1(t) \\ &\quad + \sum_{j=0}^{\infty} \sum_{i=1}^j U(t, 0) S_0 [P(t) S_0]^{j-i} V(t, 0)^* \\ &\quad \cdot C(t) U(t, 0) S_0 [P(t) S_0]^{i-1} V(t, 0)^* \\ &= A_1 R + R B_1 + R C R. \end{aligned}$$

We now make this calculation rigorous. First, since  $A_1$  and  $B_1$  are closed operators we can indeed interchange them with the summations, as we have done above, when evaluating  $A_1 R$  and  $R B_1$  (in the strong topology). For  $t > 0$  the differentiation can be carried out term-wise because the series for  $R$  and the formal series for  $dR/dt$  both converge uniformly (in  $t$ ) in  $\mathcal{L}(\mathcal{X})_s$ . To justify the product rule for differentiating individual terms we prove the following little lemma.

LEMMA 7. *Let  $\alpha$  and  $\beta$  be strongly differentiable maps from  $[0, t_0)$  into  $\mathcal{L}(\mathcal{X})_s$ , and suppose they are continuous with respect to the uniform topology. Then, for each  $x \in \mathcal{X}$  we have*

$$\frac{d}{dt} (\alpha\beta x) = \frac{d\alpha}{dt} \beta x + \alpha \frac{d\beta}{dt} x$$

*Proof.*

$$\begin{aligned} &h^{-1}[\alpha(t+h)\beta(t+h) - \alpha(t)\beta(t)]x \\ &= h^{-1}[\alpha(t+h) - \alpha(t)]\beta(t)x + \alpha(t)h^{-1}[\beta(t+h) - \beta(t)]x \\ &\quad + [\alpha(t+h) - \alpha(t)]\{h^{-1}[\beta(t+h) - \beta(t)] - \beta'(t)\}x + [\alpha(t+h) - \alpha(t)]\beta'(t)x. \end{aligned}$$

As  $h$  tends to zero the first two terms tend to

$$(\frac{d\alpha}{dt})\beta x + \alpha(\frac{d\beta}{dt})x$$

while the last term tends to zero. The third term tends to zero because  $\|\alpha(t)\|$  is locally bounded. It should be remarked that this is still true even if we remove the continuity hypothesis since strong continuity implies local boundedness of the norm (see proof of Theorem 8).

Since  $U(t, 0)$  and  $V(t, 0)$  satisfy the hypothesis of this lemma the above differentiations are indeed justified. We therefore have the following.

**THEOREM 8.** *Suppose hypotheses (I), (II), and (IV) are satisfied. Then, for some  $0 < t^* \leq t_0$  there exists a unique strongly differentiable solution of (1) on the interval  $[0, t^*]$ . Moreover the solution is in  $C([0, t^*], \mathcal{L}(\mathcal{X})_u)$ .*

*Proof.* Let  $S = R + T$  where  $T$  is the uniform solution of (1) corresponding to a zero initial condition and  $R$  is given by (13). In order to prove uniqueness we must first show that any strongly differentiable solution  $S$  for (1) must be uniformly bounded on compact sets in  $[0, t^*]$ . Let  $0 < T < t^*$ , then  $\langle S(t)x, y \rangle$  is obviously bounded on  $[0, T]$ . Applying the uniform boundedness theorem to the collection of linear operators on  $\mathcal{X} \otimes \mathcal{X}$  indexed by  $t \in [0, T]$  and defined by  $x \otimes y \rightarrow \langle S(t)x, y \rangle$ , we see that  $\|S(t)\|$  is uniformly bounded on  $[0, T]$ . Let  $S_1$  and  $S_2$  be two solutions for (1). Subtracting the two equations for  $S_1$  and  $S_2$  from each other we obtain the following equations for  $W \equiv S_1 - S_2$ :

$$dW/dt = (A - S_2C)W + W(B - S_1C),$$

$$W(0) = 0.$$

Let  $z$  be an arbitrary element of  $\mathcal{X}$  and let us solve

$$du/dt = -(B - S_1C)u(t), \quad 0 < t < T,$$

$$u(T) = z.$$

Sobolevskii's results tell us that we do indeed have a solution. Let  $q(t) \equiv W(t)u(t)$ ; then, since  $u$  is differentiable,  $W$  is strongly differentiable and  $\|W(t)\|$  uniformly bounded, it follows that  $q$  is differentiable and

$$dq/dt = (A - S_2C)q(t), \quad 0 < t < T$$

with  $q(0) = 0$ . This means that  $q \equiv 0$  or, more specifically, that  $q(T) = W(T)z = 0$ . Since  $T$  and  $z$  were arbitrary this means  $W \equiv 0$  on  $[0, t^*]$ . Since  $R$  is a series whose terms are in  $C([0, t^*], \mathcal{L}(\mathcal{X})_u)$ , and which converges uniformly on compact subintervals of  $[0, t^*]$ , it follows that  $R$ , and hence  $S$ , must also belong to that class.

Returning to the context of the transport problem discussed earlier, we remark that the decomposition  $S = R + T$  which occurs in the proof of the above theorem does not in general coincide with the specular diffuse decomposition. However, Theorem 8 does tell us that the equation for the specular part has a strong solution. Also, the equation for the diffuse part is of the type treated by Theorem 5 and therefore the diffuse part is a uniform solution.

In conclusion we mention that strong solutions are also solutions in the sense defined by Tartar. Therefore the qualitative results in [14] are applicable here.

#### REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43-57.
- [2] G. DAPRATO, *Équations d'évolution dans des algèbres d'opérateurs et application à des équations quasi-linéaires*, J. Math. Pures Appl., 48 (1969), pp. 59-107.
- [3] ———, *Quelques résultats d'existence, unicité et régularité pour un problème de la théorie du contrôle*, J. Math. Pures Appl., 52 (1973), pp. 353-375.
- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Vol. I*, Academic Press, New York, 1958.
- [5] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
- [6] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, American Mathematics Society, Providence, RI, Vol. 31, 1957.

- [7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1970.
- [8] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, SIAM J. Control Optimization, 7 (1969), pp. 101–121.
- [9] R. REDHEFFER, *On the relation of transmission line theory to scattering and transfers*, J. Mathematical Phys., 41 (1962), pp. 1–41.
- [10] R. REDHEFFER, *Inequalities for a matrix Riccati equation*, J. Math. Mech., 8 (1959), pp. 349–367.
- [11] S. SHEW, *Transport impedance*, Doctoral dissertation, Arizona State University, 1977.
- [12] P. E. SOBOLEVSKII, *Equations of parabolic type in a Banach space*, Trans. Amer. Math. Soc., 2 (1965), pp. 1–62.
- [13] L. TARTAR, *Sur l'étude directe d'équations non linéaires intervenant en théorie du contrôle optimal*, J. Functional Analysis, 6 (1974), pp. 1–47.
- [14] R. TEMAM, *Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie*, J. Functional Analysis, 7 (1971), pp. 85–115.
- [15] A. WANG, *Linear operators and transfer of radiation with spherical symmetry*, J. Mathematical Phys., 14 (1973), pp. 855–862.
- [16] ———, *Nonstationary multiple scattering*, J. Mathematical Phys., 18 (1977), pp. 47–51.
- [17] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1966.

## TURÁNIANS AND WRONSKIANS FOR THE ZEROS OF BESSEL FUNCTIONS\*

LEE LORCH†

**Abstract.** Paul Turán [On the zeros of the polynomials of Legendre, Časopis pro Pěstování Mat. a Fys., 75 (1950), pp. 113–122] proved that the Legendre polynomials satisfy the inequality  $P_n(x)P_{n+2}(x) - [P_{n+1}(x)]^2 < 0$ ,  $-1 < x < 1$ . Here it is shown that the positive zeros of arbitrary real Bessel functions satisfy similar inequalities, even in a more general form. An analogous result is established for the corresponding Wronskian. In § 8, Remark 3, the monotonicity results established in the course of the proofs here are used to complement those derived by Sturm methods in [LEE LORCH, Elementary comparison techniques for certain classes of Sturm–Liouville equations, Proc. Uppsala 1977 Inter. Conf. Diff. Equations, Symposia Univ. Upsaliensis Annum Quingentesimum Celebrantis 7, Acta Univ. Upsaliensis, Uppsala 1977, pp. 125–133].

**1. Background.** Paul Turán established [9], for the Legendre polynomials  $P_n(x)$ , the determinantal inequality

$$(1) \quad \begin{vmatrix} P_n(x) & P_{n+1}(x) \\ P_{n+1}(x) & P_{n+2}(x) \end{vmatrix} < 0, \quad -1 < x < 1,$$

$n = 0, 1, 2, \dots$ . G. Szegő [6; cf. also 7, p. 388, Problem 70] later supplied four different proofs. S. Karlin and G. Szegő [3] studied the oscillatory properties of such determinants (which they named Turánians) of second and higher order and Wronskians. An extensive literature, not enumerated here, has arisen from the search for Turán-type inequalities for other orthogonal polynomials.

In addition, O. Szász [5] established an analogue for Bessel functions of the first kind, namely

$$(2) \quad \begin{vmatrix} J_\nu(x) & J_{\nu+1}(x) \\ J_{\nu+1}(x) & J_{\nu+2}(x) \end{vmatrix} < 0, \quad \nu > -1, -\infty < x < \infty.$$

The corresponding result (in which the sign is reversed) for the modified Bessel function  $K_\nu(x)$  has been established independently by M. E. H. Ismail and M. E. Muldoon [2] and by H. van Haeringen [10].

Here there will be established corresponding results for the positive zeros of the general Bessel function  $\mathcal{C}_\nu(x) = AJ_\nu(x) + BY_\nu(x)$ , where the real numbers  $A, B$  are independent of both  $x$  and  $\nu$ . The precise statements are in § 2. In § 3 the appropriate Wronskian is defined and the corresponding inequality stated. §§ 4, 5, 6, 7 are devoted to the proofs.

Finally, § 8 incorporates various remarks, including some which relate results established here (which are basically on the monotonicity of ratios of Bessel function zeros) with those found in [4].

**2. Turánians for the zeros of Bessel functions.** Replacing the Bessel functions in (2) by the positive zeros  $c_{\nu,k}$  of the general Bessel function  $\mathcal{C}_\nu(x)$  suggests two analogues of (2), namely

$$(3) \quad T_1 = \begin{vmatrix} c_{\nu,k} & c_{\nu,k+1} \\ c_{\nu,k+1} & c_{\nu,k+2} \end{vmatrix} < 0,$$

and

$$(4) \quad T_2 = \begin{vmatrix} c_{\nu,k} & c_{\nu+1,k} \\ c_{\nu+1,k} & c_{\nu+2,k} \end{vmatrix} < 0,$$

\* Received by the editors December 5, 1978. This work was supported in part by the National Research Council of Canada.

† Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3.

where  $\nu \geq 0$ . The restriction  $\nu \geq 0$  is purely formal. It can be omitted, provided one adopts the convention suggested by [11, pp. 508–9] so that  $c_{\nu,k}$  remains an analytic function of  $\nu$ . When  $c_{\nu,k} = j_{\nu,k} [y_{\nu,k}]$ , the  $k$ th positive zero of  $J_\nu(x) [Y_\nu(x)]$ , then the results are valid for  $\nu > -1$  (respectively,  $\nu > -\frac{1}{2}$ ).

Both (3) and (4) are valid. Indeed, they are special cases of the inequality (proved in §§ 6, 7)

$$(5) \quad T = \begin{vmatrix} c_{\nu,k} & c_{\nu+\delta,k+h} \\ c_{\nu+\varepsilon,k+r} & c_{\nu+\delta+\varepsilon,k+h+r} \end{vmatrix} < 0,$$

for  $\nu \geq 0; \varepsilon \geq 0, \delta \geq 0; h, r = 0, 1, 2, \dots, \varepsilon + r > 0, h + \delta > 0$ .

The inequality (3) is the special case  $\varepsilon = \delta = 0, h = r = 1$ , (4), the special case  $\varepsilon = \delta = 1, h = r = 0$ . Together, (3) and (4) are analogues of the duality between  $x$  and  $n$  mentioned in [3, p. 4], with  $\nu$  now in the role of  $n$ .

**3. The Wronskian for the zeros of Bessel functions.** Here the notation is arranged so that  $\gamma_{\nu,k} < c_{\nu,m}$ , where  $\gamma_{\nu,k}$  is the  $k$ th positive zero of a Bessel function of order  $\nu$ ,  $c_{\nu,m}$ , the  $m$ th positive zero of a Bessel function of the same order  $\nu$ , not necessarily linearly independent of the first. The derivative with respect to  $\nu$  of  $\gamma_{\nu,k}$  is written as  $\gamma_{\nu,k}^{(1)}$ , of  $c_{\nu,m}$ , as  $c_{\nu,m}^{(1)}$ . In this notation, the Wronskian is defined to be

$$(6) \quad W(\gamma_{\nu,k}, c_{\nu,m}) = \begin{vmatrix} \gamma_{\nu,k} & c_{\nu,m} \\ \gamma_{\nu,k}^{(1)} & c_{\nu,m}^{(1)} \end{vmatrix}.$$

It will be shown that

$$(7) \quad W(\gamma_{\nu,k}, c_{\nu,m}) < 0, \quad \nu \geq 0,$$

so that, in particular, still for  $\nu \geq 0$ ,

$$(8) \quad W(c_{\nu,k}, c_{\nu,k+m}) < 0, \quad k, m = 1, 2, \dots$$

**4. Preliminary remark to the proofs of (5) and (7).** Common to these proofs is the formula given by G. N. Watson [11, (3), p. 508]

$$(9) \quad \frac{dc}{d\nu} = 2c \int_0^\infty K_0(2c \sinh t) \exp(-2\nu t) dt,$$

where  $c$  is any positive zero of  $\mathcal{C}_\nu(x)$  of fixed rank and  $K_0(x)$  is the customary modified Bessel function of order zero.

**5. Proof of (7).** The result for the Wronskian is now immediate, since (9) implies that

$$\frac{1}{2}W(\gamma_{\nu,k}, c_{\nu,m}) = \gamma_{\nu,k}c_{\nu,m} \int_0^\infty [K_0(2c_{\nu,m} \sinh t) - K_0(2\gamma_{\nu,k} \sinh t)] \exp(-2\nu t) dt,$$

and this is negative because  $K_0(x)$  decreases as  $x$  increases.

**6. Proof of (5). Preliminaries.**

LEMMA 1. For  $\nu \geq 0$  and fixed,

$$(10) \quad \frac{c_{\nu,k+m}}{c_{\nu,k}} \downarrow 1 \quad \text{as } k \text{ increases, } m = 1, 2, \dots$$

*Proof.* That the ratio decreases follows readily from [4, (4.5), p. 130]. To see that



the limit is one, we note, e.g. [11, p. 517]

$$0 < \frac{c_{\nu,k+1}}{c_{\nu,k}} - 1 = \frac{c_{\nu,k+1} - c_{\nu,k}}{c_{\nu,k}} = \frac{\pi + o(1)}{c_{\nu,k}} = o(1), \quad k \rightarrow \infty,$$

and the proof is complete.

LEMMA 2. *If  $c_{\nu+r,m} > \gamma_{\nu,k}$ ,  $k, r, m$  fixed, then*

$$(11) \quad \frac{c_{\nu+r,m}}{\gamma_{\nu,k}} \downarrow 1, \quad \text{as } \nu \rightarrow \infty.$$

*Proof.* Denoting differentiation with respect to  $\nu$  by  $D_\nu$ , we have

$$\begin{aligned} D_\nu \left\{ \log \frac{c_{\nu+r,m}}{\gamma_{\nu,k}} \right\} &= \frac{D_\nu c_{\nu+r,m}}{c_{\nu+r,m}} - \frac{D_\nu \gamma_{\nu,k}}{\gamma_{\nu,k}} \\ &= 2 \int_0^\infty [K_0(2c_{\nu+r,m} \sinh t) - K_0(2\gamma_{\nu,k} \sinh t)] \exp(-2\nu t) dt, \end{aligned}$$

and this is negative since  $K_0(x)$  decreases as  $x$  increases.

Thus, the ratio in (11) decreases as  $\nu$  increases. That the limit is one follows from the known asymptotics of the zeros of Bessel functions [8], [11, p. 521].

The special cases useful here are

$$(12) \quad \frac{c_{\nu+r,m}}{c_{\nu,k}} \downarrow 1 \quad \text{as } \nu \uparrow \infty, \text{ for } c_{\nu+r,m} > c_{\nu,k},$$

in particular,

$$(13) \quad \frac{c_{\nu,k+h}}{c_{\nu,k}} \downarrow 1 \quad \text{as } \nu \uparrow \infty, \quad h, k = 1, 2, \dots$$

Moreover, if  $\nu, \delta (\geq 0)$  and  $r = 0, 1, 2, \dots$ , are fixed,  $r + \delta > 0$ , then

$$(14) \quad \frac{c_{\nu+\delta,k+r}}{c_{\nu,k}} \downarrow 1 \quad \text{as } k \uparrow \infty.$$

To prove (14) it suffices to show that

$$\frac{c_{\nu,k+1}}{c_{\nu,k}} > \frac{c_{\nu+\delta,k+r+1}}{c_{\nu+\delta,k+r}}.$$

But, from (12),

$$\frac{c_{\nu,k+1}}{c_{\nu,k}} \geq \frac{c_{\nu+\delta,k+1}}{c_{\nu+\delta,k}},$$

with strict inequality if  $\delta > 0$ , and from Lemma 1 (10),

$$\frac{c_{\nu+\delta,k+1}}{c_{\nu+\delta,k}} \geq \frac{c_{\nu+\delta,k+r+1}}{c_{\nu+\delta,k+r}},$$

with strict inequality if  $r > 0$ .

**7. Proof of (5). Conclusion.** The assertion of (5) can be written equivalently as

$$\frac{c_{\nu+\delta+\varepsilon,k+h+r}}{c_{\nu+\varepsilon,k+r}} < \frac{c_{\nu+\delta,k+h}}{c_{\nu,k}}.$$

The left member, by (12), is less than

$$\frac{c_{\nu+\delta,k+h+r}}{c_{\nu,k+r}}$$

unless  $\varepsilon = 0$ , in which case equality occurs, and this in turn, from (14), is less than

$$\frac{c_{\nu+\delta,k+h}}{c_{\nu,k}}$$

unless  $r = 0$ , in which case there is equality. But  $\varepsilon + r > 0$ , so that strict inequality must occur at least once.

**8. Remarks.**

1. Inequality (3) can be written

$$T_1 = \left| \begin{matrix} c_{\nu,k} & \Delta c_{\nu,k} \\ \Delta c_{\nu,k} & \Delta^2 c_{\nu,k} \end{matrix} \right| < 0.$$

In this form, it is clear that the assertion is trivial when  $\nu \geq \frac{1}{2}$ , for then  $\Delta^2 c_{\nu,k} \leq 0$ , as Sturm pointed out in 1836 [11, p. 517]. However, for  $0 \leq \nu < \frac{1}{2}$ , Sturm's method shows that  $\Delta^2 c_{\nu,k} > 0$  and the inequality (3) acquires more substance.

2. The inequality (5) may be contrasted with the discussion of the function  $\varphi(x)$  defined in [3, (33.2), p. 154]. One possible comparison arises on putting  $h = r = 0$  in  $T$ . The inequality

$$T_3 = \left| \begin{matrix} c_{\nu,k} & c_{\nu+\delta,k} \\ c_{\nu+\varepsilon,k} & c_{\nu+\delta+\varepsilon,k} \end{matrix} \right| < 0, \quad \delta > 0, \quad \varepsilon > 0, \quad k = 1, 2, \dots,$$

results. It is valid for all  $k$ , here corresponding to  $x$  in  $\varphi(x)$ , as well as for all  $\nu \geq 0$ , which here corresponds to  $n$ . Another (dual) analogue is obtained by taking  $\varepsilon = \delta = 0$  in (5), so that

$$T_4 = \left| \begin{matrix} c_{\nu,k} & c_{\nu,k+h} \\ c_{\nu,k+r} & c_{\nu,k+h+r} \end{matrix} \right| < 0, \quad h, k, r = 1, 2, \dots$$

Here  $\nu$  corresponds to  $x$  in  $\varphi(x)$ ,  $k$  to  $n$ .

As is pointed out in [3, p. 155], the inequality  $\varphi(x) < 0$  need not be valid always. They cite, *i.a.*, a counterexample due to A. E. Danese [1].

3. In (13), the case  $h = 1$  justifies the comments found in [4, p. 130] in the paragraph following (4.7). It shows, in particular, that inequalities (4.6) and (4.7) of [4] cannot be improved by making the ranks of the zeros the same throughout those inequalities. Combining (13) with (4.6) and (4.7) gives, respectively, for  $k = 2, 3, \dots$ ,

$$\frac{c_{\nu+\varepsilon,k+1}}{c_{\nu+\varepsilon,k}} < \frac{c_{\nu,k+1}}{c_{\nu,k}} < \frac{c_{\nu+\varepsilon,k}}{c_{\nu+\varepsilon,k-1}}, \quad 0 < \varepsilon \leq 1, \quad \nu \geq 0,$$

and

$$\frac{j_{\nu+\varepsilon,k+1}}{j_{\nu+\varepsilon,k}} < \frac{j_{\nu,k+1}}{j_{\nu,k}} < \frac{j_{\nu+\varepsilon,k}}{j_{\nu+\varepsilon,k-1}}, \quad 0 < \varepsilon \leq 2, \quad \nu \geq 0,$$

where both right hand inequalities remain valid even if  $\varepsilon = 0$ , but become false as  $\varepsilon \rightarrow \infty$ . For example, the foregoing right hand inequality is false for  $\varepsilon = 3$  when, say,  $k = 2, \nu = 0$  and when  $k = 39, \nu = 2$ .

4. In [4] all proofs are based on Sturm comparison methods, unlike here. It may be of interest to note that an alternative proof of the Turán-type inequality (3) is contained

in [4, (4.6), p. 130] on taking  $\varepsilon = 0$ .

5. The inequalities here and in [4] can be used to check tables of zeros of Bessel functions.

**Acknowledgment.** I am indebted to Professor M. E. Muldoon for Reference [10], and to Professor W. A. Al-Salam for remarks (i) and (ii):

(i) (2) is implicit in a formula published by E. C. J. von Lommel in 1879 [cf. 11, p. 135 (11)];

(ii) The analogue for Hermite polynomials  $H_n(x)$ ,  $-\infty < x < \infty$ , of (1) and (2) is an obvious consequence of a formula observed by Hüseyin Demir in 1946 (Problem 4215, Amer. Math. Monthly, 53 (1946), p. 470; Solutions, Ibid., 55 (1948), pp. 34–35, by several authors).

It appears, however, that explicit statements of such inequalities as (1) were not made before P. Turán [9].

#### REFERENCES

- [1] A. E. DANESE, *Some identities and inequalities involving ultraspherical polynomials*, Duke Math. J., 26 (1959), pp. 349–36.
- [2] M. E. H. ISMAIL AND M. E. MULDOON, *Monotonicity of the zeros of a cross-product of Bessel functions*, this Journal, 9 (1978), pp. 759–767.
- [3] S. KARLIN AND G. SZEGÖ, *On certain determinants whose elements are orthogonal polynomials*, J. d'Analyse Math., 8 (1960), pp. 1–157.
- [4] LEE LORCH, *Elementary comparison techniques for certain classes of Sturm–Liouville equations*, Proc. Uppsala 1977 Inter. Conf. Diff. Equations, Symposia Univ. Upsaliensis Annum Quingentesimum Celebrantis 7, Acta Univ. Upsaliensis, Uppsala 1977, pp. 125–133.
- [5] OTTO SZÁSZ, *Inequalities concerning ultraspherical polynomials and Bessel functions*, Proc. Amer. Math. Soc., 1 (1950), pp. 256–267.
- [6] G. SZEGÖ, *On an inequality of P. Turán concerning Legendre polynomials*, Bull. Amer. Math. Soc., 54 (1948), pp. 401–405.
- [7] GÁBOR SZEGÖ, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloquium Publications, vol. 23, Amer. Math. Soc. Providence, RI, 1975.
- [8] F. G. TRICOMI, *Sulle funzioni di Bessel di ordine e argomento pressochè uguali*, Atti d. Accad. Scienze Torino, 83 (1949), pp. 3–20.
- [9] PAUL TURÁN, *On the zeros of the polynomials of Legendre*, Časopis pro Pěstování Mat. a Fys., 75 (1950), pp. 113–122.
- [10] H. VAN HAERINGEN, *Bound states for  $r^{-2}$ -like potentials in one and three dimensions*, J. Mathematical Phys., 19 (1978), pp. 2171–2179.
- [11] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1958.

## ON A CLASS $W[k, s; C_{ij}]$ OF POLYNOMIAL SETS: PART I\*

I. M. SHEFFER†

**Abstract.** A class  $W[k, s; C_{ij}]$  of polynomial sets is introduced which generalizes type zero sets and sets of class  $S^k$ . These latter sets have been studied in some detail, and have been shown to have interesting properties; and the new class should be of similar interest. Sets in  $W[k, s; C_{ij}]$  are characterized in terms of their generating functions.

**Introduction.** A polynomial set [hereafter p.s.] is an infinite sequence  $\{P_n(x)\}$  with degree  $P_n(x) = n$  ( $n = 0, 1, \dots$ ). A class of p.s. of *type zero* was defined and studied in [3]. Gian-Carlo Rota [2] has shown that type zero sets (termed Sheffer polynomials) are important in finite operator calculus. Al-Salam and Verma [1] generalized type zero sets to the class  $S^k$ , defined by

$$(0.1) \quad J[P_n(x)] \equiv \sum_{j=k}^{\infty} b_j P_n^{(j)}(x) = P_{n-k}(x) \quad (b_k \neq 0);$$

and they obtained the generating function  $G(x, t)$  that characterizes sets in  $S^k$ :

$$(0.2) \quad G(x, t) = \sum_{p=0}^{k-1} A_p(t) \exp \{xH(\omega^p t)\},$$

where  $H(t) = \sum_1^{\infty} h_j t^j$  ( $h_1 \neq 0$ ) is the formal inverse of  $J^*(t)$ :

$$(0.3) \quad H(J^*(t)) = J^*(H(t)) = t,$$

$J^*(t)$  is a  $k$ th root of  $J(t) = \sum_{j=k}^{\infty} b_j t^j$ , and  $\omega = \exp \{2\pi i/k\}$ . Further properties of  $S^k$  sets, including a characterization of those  $S^k$  sets that are orthogonal, are given in [4]; and other properties in [5].

In the present work we consider the class of p.s.  $W[k, s; C_{ij}]$  defined by a system of  $s$  equations

$$(0.4) \quad J[P_{sn+j}(x)] = \sum_{m=0}^{\infty} c_{j,m} P_{sn-k+j-m}(x) \quad (j = 0, 1, \dots, s-1; n = 0, 1, 2, \dots)$$

with constant  $\{c_{j,m}\}$ , where  $J$  is the differential operator

$$(0.5) \quad J[y] \equiv \sum_{j=k}^{\infty} b_j y^{(j)}(x) \quad (\{b_j\} \text{ constants; } b_k \neq 0).$$

We propose to characterize p.s. solutions of (0.4) by means of their generating functions. Throughout the work we use formal power series; and  $\omega, \zeta$  will always stand for

$$(0.6) \quad \omega = \exp \{2\pi i/k\}, \quad \zeta = \exp \{2\pi i/s\}.$$

### 1. Preliminary results. One readily proves

LEMMA 1.1. System (0.4) has a p.s. solution  $\{P_n(x)\}$  iff

$$(1.1) \quad \prod_{j=0}^{s-1} c_{j,0} \neq 0.$$

\* Received by the editors November 22, 1978.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

And when (1.1) holds, there is a unique p.s. solution for which (i)  $P_j(x)$  ( $j = 0, \dots, k - 1$ ) are preassigned (with  $P_j$  of degree  $j$ ), and (ii)  $\{p_{n,j}\}$  ( $j = 0, \dots, k - 1; n \geq k$ ) are preassigned. (Here  $P_n(x) = \sum_{j=0}^n p_{n,j}x^j$ .)

Note that if degree  $P_j(x) < j$  for at least one  $j < k$ , then  $\{P_n(x)\}$  will be a solution of (0.4) but not a p.s. solution.

Let  $\{P_n(x)\}$  be any p.s., with generating function

$$(1.2) \quad G(x, t) = \sum_{n=0}^{\infty} P_n(x)t^n = \sum_{n=0}^{\infty} \left\{ \sum_{j=0}^n p_{n,j}x^j \right\} t^n.$$

Writing this as a formal series in  $x$  we have

$$(1.3) \quad G(x, t) = \sum_{n=0}^{\infty} g_n(t)x^n, \quad g_n(t) = \sum_{j=n}^{\infty} p_{j,n}t^j \quad (p_{n,n} \neq 0).$$

It is seen from Lemma 1.1 that we have

LEMMA 1.2. *Let (1.1) hold. There is a unique p.s. satisfying (0.4) for which  $\{g_j(t)\}$  ( $j = 0, \dots, k - 1$ ) are preassigned. And all p.s. solutions are obtained in this way.*

Let  $\{P_n\}$  be any p.s., with generating function [hereafter g.f.] (1.2). Let

$$(1.4) \quad G_i(x, t) = \sum_{n=0}^{\infty} P_{ns+i}(x)t^{ns+i} \quad (i = 0, 1, \dots, s - 1).$$

We call  $G_i$  the  $i$ -th component (mod  $s$ ) of  $G$ . (In general, a formal series  $\sum_{n=0}^{\infty} a_{ns+i}t^{ns+i}$  is an  $i$ -th component (mod  $s$ ).

Let (1.2) define a p.s. solution of (0.4) and let  $\{G_i\}$  be the components (mod  $s$ ) of  $G$ . Multiply the equations of (0.4) by  $t^{sn+i}$  ( $j = 0, \dots, s - 1$ ) respectively and sum on  $n$ . We obtain the system

$$(1.5) \quad J[G_j(x, t)] = t^k \cdot \sum_{h=0}^{s-1} t^h C_{j,h}(t) G_{s-r+j-h}(x, t) \quad (j = 0, 1, \dots, s - 1),$$

where  $k \equiv r \pmod{s}$ ,  $0 \leq r < s$  and

$$(1.6) \quad C_{i,j}(t) = \sum_{n=0}^{\infty} c_{i,j+ns} t^{ns} \quad (i, j = 0, \dots, s - 1),$$

and indices on the  $G$ 's are to be reduced (mod  $s$ ) to the range  $[0, s - 1]$ . E.g., if  $r = 2$  then  $G_{s-r+s-1} = G_{s-3}$  if  $s \geq 3$ . Note that each  $C_{ij}(t)$  is a 0-component (mod  $s$ ).

Conversely, given (1.5), on equating coefficients of like powers of  $t$  we return to (0.4); so we have

LEMMA 1.3. *Systems (0.4), (1.5) are equivalent.*

We find it convenient to deal with (1.5).

Suppose  $s = 1$ . Then (0.4), (1.5) become

$$(1.7) \quad J[P_n(x)] \equiv \sum_{j=k}^{\infty} b_j P_n^{(j)}(x) = \sum_{m=0}^{\infty} c_m P_{n-k-m}(x) \quad (b_k c_0 \neq 0),$$

$$(1.8) \quad J[G(x, t)] = t^k C(t)G(x, t), \quad C(t) = \sum_{j=0}^{\infty} c_j t^j.$$

Let  $J^*(t)$ ,  $C^*(t)$  be arbitrary but fixed  $k$ th roots of  $J(t) = \sum_{j=k}^{\infty} b_j t^j$ ,  $C(t)$ . Now  $J^*$  is a series  $\sum_1^{\infty} j_n t^n$  ( $j_1 \neq 0$ ) so it has a formal inverse

$$(1.9) \quad I(t) = \sum_1^{\infty} e_n t^n \quad (e_1 \neq 0): \quad J^*(I(t)) = I(J^*(t)) = t.$$

THEOREM 1.1. A p.s.  $\{P_n(x)\}$  satisfies (1.7) iff its g.f. has the form

$$(1.10) \quad G(x, t) = \sum_{p=0}^{k-1} A_p(t) \exp \{xI(\omega^p tC^*(t))\},$$

where  $\{A_p(t)\}$  are arbitrary formal power series:  $A_p(t) = \sum_{n=0}^{\infty} a_{p,n}t^n$  such that

$$(1.11) \quad \sum_{p=0}^{k-1} a_{p,0}\omega^{pn} \neq 0 \quad (n = 0, \dots, k-1).$$

*Proof.* If (1.10) is expanded in a series in  $x$  the coefficient of  $x^n/n!$  is  $(e_1^n c_0) (\sum_{p=0}^{k-1} a_{p,0}\omega^{pn})t^n +$  higher powers. So (1.10) will be the g.f. of a p.s. iff (1.11) holds for all  $n$ , hence for all  $n \in [0, k-1]$ , since  $\omega^k = 1$ .

That (1.10) is a solution of (1.8) is seen by direct substitution, using

$$\begin{aligned} \sum_{j=k}^{\infty} b_j [I(\omega^p tC^*(t))]^j &= J(I(\omega^p tC^*(t))) = [J^*(I(\omega^p tC^*(t)))]^k \\ &= [\omega^p tC^*(t)]^k = t^k C(t). \end{aligned}$$

There remains to show that if  $\{P_n(x)\}$  is a p.s. solution of (1.7) then its g.f.  $G(x, t)$  has the form (1.10).  $G$  is given by (1.2), (1.3). By Lemma 1.2 it suffices to show that  $\{A_p(t)\}$  can be chosen so that

$$(1.12) \quad n! g_n(t) = \sum_{p=0}^{k-1} A_p(t) H_p^n(t), \quad (H_p(t) = I(\omega^p tC^*(t))) \quad (n = 0, \dots, k-1).$$

The coefficient determinant  $\Delta(t)$  is a Vandermond determinant in  $\{H_p(t)\}$ . One readily finds that

$$\Delta(t) = D(e_1 c_0^* t)^{(k-1)k/2} + \text{higher powers,}$$

where  $D$  is a nonzero Vandermond determinant in  $1, \omega, \dots, \omega^{k-1}$ .

If we solve (1.12) for  $A_p(t)$  by Cramer's rule, the numerator determinant is seen to have a factor  $t^{(k-1)k/2}$ , so each  $A_p(t)$  is a power series. Hence the g.f. for the given p.s. solution has the form (1.10).

The defining equation (0.1) for sets in class  $S^k$  is a particular case of (0.4), with  $s = 1$  and  $C(t) = 1$  in (1.8). However, a p.s. can be in  $S^k$  even if  $C(t) \neq 1$ :

THEOREM 1.2. Let the p.s.  $\{P_n(x)\}$  satisfy (1.7). If  $C(t)$  is a 0-component (mod  $k$ ) then  $\{P_n(x)\} \in S^k$ .

*Proof.*  $C^*(t)$  is a  $k$ th root of  $C(t)$ , so it also is a 0-component (mod  $k$ ). Let  $H(t) = I(tC^*(t))$ . Then  $H(\omega^p t) = I(\omega^p tC^*(t))$ , so the g.f.  $G$  for  $\{P_n(x)\}$ , which has the form (1.10), can be written as (0.2). Hence  $\{P_n\} \in S^k$ .

*Remark.* The converse of Theorem 1.2 is false for  $k > 1$ . It is easy to set up an example.

**2. General case.** We now consider system (0.4). Let integers  $q, r$  satisfy

$$(2.1) \quad k = qs + r \quad (0 \leq r < s).$$

Multiply the equations of (1.5) by functions (i.e., formal series)  $v_0(t), \dots, v_{s-1}(t)$  and add:

$$(2.2) \quad J[G^*(x, t)] = t^k \sum_{j=0}^{s-1} v_j(t) \left\{ \sum_{m=0}^{s-1} t^m C_{j,m}(t) G_{s-r+j-m}(x, t) \right\},$$

$$(2.3) \quad G^*(x, t) = \sum_{j=0}^{s-1} v_j(t) G_j(x, t).$$

We wish to choose  $\{v_j\}$  so that the right side of (2.2) is  $t^k u(t)G^*$  for a suitably chosen  $u(t)$ . This requires that we satisfy the system

$$(2.4) \quad \sum_{j=0}^{s-1} v_j [t^{j+m} C_{j,j+m}]^* = u(t) v_{s-r-m} \quad (m = 0, \dots, s-1),$$

where the asterisk signifies that wherever  $j+m$  is outside the range  $[0, s-1]$  the corresponding power of  $t$  and second index on  $C$  are to be reduced (mod  $s$ ) to that range. E.g., if  $r+m = s+n$  ( $0 \leq n < s$ ) then  $[t^{j+m} C_{j,j+m}]^* = t^n C_{j,n}$ . Also the  $v$ -index  $s-r-m$  is to be reduced (mod  $s$ ).

After a suitable permutation of the equations of (2.4) the determinant of the coefficients of  $\{v_j\}$  is

$$(2.5) \quad \Delta(t) = \begin{vmatrix} [t^{s-r} C_{0,s-r} - u] & [t^{s-r+1} C_{1,s-r+1}] \cdots [t^{s-r-1} C_{s-1,s-r-1}] \\ [t^{s-r-1} C_{0,s-r-1}] & [t^{s-r} C_{1,s-r} - u] \cdots [t^{s-r-2} C_{s-1,s-r-2}] \\ [t^{s-r+1} C_{0,s-r+1}] & [t^{s-r+2} C_{1,s-r+2}] \cdots [t^{s-r} C_{s-1,s-r} - u] \end{vmatrix}^*$$

where the asterisk has the same meaning as in (2.4).

To have a nontrivial set  $\{v_j\}$  we must have

$$(2.6) \quad \Delta(t) = 0.$$

This equation, which we term the  $u$ -equation, is algebraic of degree  $s$  in  $u(t)$ . Its roots play a central rôle in the theory of system (0.4).

Each root  $u(t)$  determines a solution  $\{v_j\}$  of (2.4), leading to a  $G^*$  of (2.3), and we have

$$(2.7) \quad J[G^*] = t^k u(t)G^*.$$

Now (2.7) is of the form (1.8) with  $C(t) = u(t)$ , so from the case  $s = 1$  we have

$$(2.8) \quad G^*(x, t) = \sum_{p=0}^{k-1} A_p(t) \exp \{xI(\omega^p t u^*(t))\},$$

where  $u^*(t)$  is a  $k$ th root of  $u(t)$ .

Canceling  $(-1)^s$ , (2.6) we can write as

$$(2.9) \quad [u(t)]^s + D_1(t)[u(t)]^{s-1} + \dots + D_s(t)[u(t)]^0 = 0.$$

LEMMA 2.1. Each  $D_i(t)$  in (2.9) is an  $r(s-i)$ -component (reduced (mod  $s$ )), so that

$$(2.10) \quad D_i(\zeta t) = \zeta^{r(s-i)} D_i(t) \quad (i = 1, \dots, s).$$

*Proof.* Let the rows and columns of (2.5) be denoted as the 0th, 1st,  $\dots$ ,  $(s-1)$ th. Each nondiagonal element (and the part of diagonal elements not including the  $u$ ) is of the form  $t^j C_{ij}(t)$ . Now each  $C_{ij}$  is a 0-component (mod  $s$ ); hence each power of  $t$  in series  $t^j C_{ij}(t)$  is congruent to  $j$  (mod  $s$ ). We may therefore ignore the  $C_{ij}$ 's.

Let  $n$  be given,  $0 \leq n < s$ . A typical term in the expansion of  $\Delta(t)$  that has  $u(t)$  as a factor exactly  $n$  times will have  $u(t)$  coming from  $n$  diagonal elements, say in the positions  $(p_1, p_1), \dots, (p_n, p_n)$ . The remaining factors will come from elements in positions  $(q_i, r_i)$  ( $i = 1, \dots, s-n$ ), where  $\{q_i\}, \{r_i\}$  independently fill out the complement of  $p_1, \dots, p_n$  relative to the set  $0, 1, \dots, s-1$ . The power of  $t$  in the place  $(q_i, r_i)$  is  $s-r-q_i+r_i$ , so the above-mentioned typical term has  $t$  to a power congruent (mod  $s$ ) to

$$(2.11) \quad \sum_{i=1}^{s-n} (s-r-q_i+r_i) = (s-r)(s-n) - \sum q_i + \sum r_i.$$

Now

$$\sum_{i=1}^{s-n} r_i + \sum_{j=1}^n p_j = \sum_{m=0}^{s-1} 1 = \sum_{i=1}^{s-n} q_i + \sum_{j=1}^n p_j,$$

so  $\sum q_i = \sum r_i$ . Thus (2.11) has a value  $\equiv rn \pmod{s}$ ; so every power of  $t$  in the series for  $D_{s-n}(t)$  is  $\equiv rn \pmod{s}$ .

From Lemma 2.1 we derive a useful result:

**THEOREM 2.1.** *If  $u(t)$  is any root of the  $u$ -equation then*

$$(2.12) \quad u(t), \zeta^r u(\zeta t), \zeta^{2r} u(\zeta^2 t), \dots, \zeta^{(s-1)r} u(\zeta^{s-1} t)$$

are also roots (not necessarily different).

*Proof.* It suffices to show that  $v(t) = \zeta^r u(\zeta t)$  is a root. In (2.9) replace  $t$  by  $\zeta t$  and use (2.10). We obtain the  $u$ -equation with  $u$  replaced by  $v$ .

**LEMMA 2.2.** *Every root of the  $u$ -equation is nonzero for  $t = 0$ .*

*Proof.* Suppose  $u(0) = 0$  for some root. Taking  $t = 0$  in (2.5) we obtain  $0 = \Delta(0) = \prod_{j=0}^{s-1} c_{j0}$  where  $r = 0$ , and  $0 = \Delta(0) = \pm C_{r,0}(0)C_{r+1,0}(0) \cdots C_{r-1,0}(0) = \prod_{j=0}^{s-1} c_{j0}$  when  $r \neq 0$ . This contradicts (1.1).

Corresponding to a root  $u(t)$  we obtained a solution (2.8) of (2.7). If we suppose that the roots  $u_1, \dots, u_s$  are distinct there will correspond functions  $G_1^*, \dots, G_s^*$  of the form (2.8), and solutions  $\{v_{1j}\}, \dots, \{v_{sj}\}$  of (2.4). If we assume that  $\det |v_{ij}(t)| \neq 0$  we can solve the system obtained from (2.3) for  $\{G_j(x, t)\}$ , and thus obtain the g.f. for a solution of (0.4).

It is thus suggested that solutions of (0.4) may be given by g.f.'s  $G(x, t) = \sum_{i=0}^{s-1} G_i$ , where

$$(2.13) \quad G_i(x, t) = \sum_{m=1}^s \sum_{p=0}^{k-1} A_{i,p,m}(t) \exp \{xI(\omega^p t u_m^*(t))\}$$

( $i = 0, \dots, s-1$ ), where  $\{u_m(t)\}$  are the roots of (2.6) and  $u_m^*(t)$  is a  $k$ th root of  $u_m(t)$ . We shall see that this is so for the case considered in Part I. For other cases, particularly when the  $u$ -equation has multiple roots, (2.13) requires modification.

**3. Case  $(s, k) = 1$ .** From (2.1) we have  $(s, r) = 1$ .

**LEMMA 3.1.** *Let  $(s, k) = 1$ ; then the  $s$  roots of the  $u$ -equation are all different at  $t = 0$ , hence they are distinct.*

*Proof.* Let  $u(t)$  be a root. The  $s$  roots of (2.12) have the values  $\zeta^{jr} u(0) \cdot (j = 0, \dots, s-1)$  at  $t = 0$ . Now  $u(0) \neq 0$  by Lemma 2.2, and  $\{jr\}$  is a complete residue system  $\pmod{s}$ , so no two of  $\{\zeta^{jr} u(0)\}$  are equal.

**LEMMA 3.2.** *Let  $(s, k) = 1$ . If  $u_1, \dots, u_s$  are the roots of (2.6) taken in the order (2.12) so that*

$$(3.1) \quad u_1(t) = u(t), u_2(t) = \zeta^r u(\zeta t), \dots, u_s(t) = \zeta^{(s-1)r} u(\zeta^{s-1} t),$$

then

$$(3.2) \quad \zeta u_i^*(\zeta t) = u_{i+1}^*(t) \quad (i = 1, \dots, s; u_{s+1}^* = u_1^*, u_0^* = u_s^*).$$

Here  $u_1^*(t)$  is a  $k$ -th root of  $u_1(t)$ , and for  $i > 1$ ,  $u_i^*(t)$  is that  $k$ -th root of  $u_i(t)$  given by  $\zeta^{(i-1)r/k} u_1^*(\zeta^{i-1} t) \omega^{(i-1)q}$ .

*Proof.* We have  $\zeta = \omega^{k/s}$ , so

$$(3.3) \quad \zeta u_i^*(\zeta t) = \zeta^{(k-r)/k} \omega^{-q} u_{i+1}^*(t) = u_{i+1}^*(t).$$



We saw earlier that

$$(3.4) \quad G_i(x, t) = \sum_{m=1}^s \sum_{p=0}^{k-1} A_{i,p,m}(t) \exp \{xI(\omega^p tu_m^*(t))\} \quad (i = 0, \dots, s-1)$$

might give a solution of (0.4) [or equally (1.5)]. To determine when it does, substitute (3.4) into (1.5). The left side of (1.5) becomes

$$\sum_{m=1}^s \sum_{p=0}^{k-1} A_{i,p,m}(t) \exp \{xI(\omega^p tu_m^*(t))\} \cdot J(I(\omega^p tu_m^*(t))).$$

Now  $J^{*k}(t) = J(t)$ ,  $J^*(I(t)) = t$ , so

$$J(I(\omega^p tu_m^*(t))) = [\omega^p tu_m^*(t)]^k = t^k u_m(t).$$

Thus

$$J[G_j(x, t)] = t^k \sum_{m=1}^s \sum_{p=0}^{k-1} A_{j,p,m}(t) u_m(t) \exp \{xI(\omega^p tu_m^*(t))\}.$$

The right side of (1.5) becomes

$$t^k \sum_{m=1}^s \sum_{p=0}^{k-1} \sum_{d=0}^{s-1} t^d C_{j,d}(t) A_{s-r+j-d,p,m}(t) \exp \{xI(\omega^p tu_m^*(t))\}.$$

So (3.4) will give a solution of (0.4) (hence of (1.5)) iff the set  $\{A_{i,p,m}(t)\}$  satisfies the system

$$(3.5) \quad \sum_{m=1}^s \sum_{p=0}^{k-1} M_{j,p,m}(t) \exp \{xI(\omega^p tu_m^*(t))\} = 0 \quad (j = 0, \dots, s-1),$$

where

$$(3.6) \quad M_{j,p,m}(t) = A_{j,p,m}(t)u(t) - \sum_{d=0}^{s-1} t^d C_{j,d}(t)A_{s-r+j-d,p,m}(t).$$

LEMMA 3.3. Let  $\{r_m(t)\}$ ,  $\{V_{p,m}(t)\}$  ( $m = 1, \dots, s$ ;  $p = 0, \dots, k-1$ ) and  $I(t) = \sum_{n=1}^{\infty} e_n t^n$  ( $e_1 \neq 0$ ) be formal power series, and suppose

$$\omega^{p_1} r_{m_1}(0) = \omega^{p_2} r_{m_2}(0) \quad (0 \leq p_1, p_2 < k)$$

holds only when  $p_1 = p_2$ ,  $m_1 = m_2$ . If

$$(3.7) \quad \sum_{m=1}^s \sum_{p=0}^{k-1} V_{p,m}(t) \exp \{xI(\omega^p tr_m(t))\} = 0$$

then

$$(3.8) \quad V_{p,m}(t) = 0 \quad (m = 1, \dots, s; p = 0, \dots, k-1).$$

*Proof.* Expand (3.7) in powers of  $x$  and equate to zero the coefficient of  $x^n$ :

$$(3.9) \quad \sum_{m=1}^s \sum_{p=0}^{k-1} V_{p,m}(t) [I(\omega^p tr_m(t))]^n = 0 \quad (n = 0, 1, \dots).$$

Let

$$(3.10) \quad V_{p,m}(t) = \sum_{n=0}^{\infty} v_{p,m,n} t^n, \quad \omega^p r_m(t) = \sum_{n=0}^{\infty} d_{p,m,n} t^n.$$

Cancel  $t^n$  from (3.9) and set  $t = 0$ :

$$(3.11) \quad \sum_{m=1}^s \sum_{p=0}^{k-1} v_{p,m,0} d_{p,m,0}^n = 0 \quad (n = 0, 1, \dots).$$

Take  $n = 0, 1, \dots, sk - 1$ . This gives a homogeneous system of  $sk$  equations in the  $sk$  quantities  $\{v_{p,m,0}\}$ , with a non-zero Vandermond determinant in the  $sk$  distinct numbers  $\{d_{p,m,0}\}$ . Hence  $v_{p,m,0} = 0$  (all  $p, m$ ).

We can now cancel a factor  $t$  from each  $V_{p,m}$  in (3.10), and on repeating the above argument we get  $v_{p,m,1} = 0$ ; then  $v_{p,m,2} = 0$ ; and so on. So (3.8) holds.

Apply Lemma 3.3 to (3.5), with  $V_{p,m} = M_{i,p,m}$  ( $j$  fixed),  $r_m(t) = u_m^*(t)$ . Suppose  $\omega^{p_1} u_{m_1}^*(0) = \omega^{p_2} u_{m_2}^*(0)$ . Raising to the  $k$ th power:  $u_{m_1}(0) = u_{m_2}(0)$ . Since  $(s, k) = 1$ , Lemma 3.1 implies  $m_1 = m_2$ ; hence also  $p_1 = p_2$ . Thus the hypotheses of Lemma 3.3 hold, so  $M_{i,p,m}(t) = 0$  (all  $j, m, p$ ).

We conclude that if (3.4) is a solution of (1.5) then the functions  $\{A_{i,p,m}(t)\}$  satisfy system

$$(3.12) \quad u_m(t) A_{i,p,m}(t) = \sum_{j=0}^{s-1} t^j C_{i,j}(t) A_{s-r+i-j,p,m}(t) \\ (m = 1, \dots, s; i = 0, \dots, s-1; p = 0, \dots, k-1).$$

By convention the first index  $s - r + i - j$  is to be reduced (mod  $s$ ) to the range  $[0, s - 1]$ . Conversely, if  $\{G_i\}$  is given by (3.4), and if (3.12) holds, we can work back to (3.5) and hence to (1.5). To sum up so far;  $\{G_i\}$  given by (3.4) is a solution of (1.5) iff (3.12) holds.

However, we have not yet imposed the condition that  $G_i$  be an  $i$ -component (mod  $s$ ); i.e., that it satisfy

$$(3.13) \quad G_i(x, \zeta t) = \zeta^i G_i(x, t).$$

From (3.4) and (3.2) we get

$$(3.14) \quad G_i(x, \zeta t) = \sum_{m=1}^s \sum_{p=0}^{k-1} A_{i,p,m-1}(\zeta t) \exp \{xI(\omega^p t u_m^*(t))\}.$$

Here  $A_{i,p,0}(t) = A_{i,p,s}(t)$ .

Putting (3.14) into (3.13) we obtain equations like (3.7). Lemma 3.3 again applies, to yield conditions

$$(3.15) \quad A_{i,p,m-1}(\zeta t) = \zeta^i A_{i,p,m}(t) \quad (m = 1, \dots, s; i = 0, \dots, s-1; p = 0, \dots, k-1).$$

Conversely, (3.4) and (3.15) imply (3.13); so if  $\{G_i\}$  is given by (3.4) then (3.13) holds if we have (3.15).

Combining results we have

LEMMA 3.4. Let  $\{G_i\}$  be given by (3.4), with (3.13) holding. Then  $G = \sum_{i=0}^{s-1} G_i$  is a solution of (1.5) iff  $\{A_{i,p,m}(t)\}$  satisfies (3.12) and (3.15).

If in system (3.12) (with  $p, m$  fixed) we transfer all terms to the right and arrange the  $A$ 's in the order  $A_{0,p,m}, A_{1,p,m}, \dots, A_{s-1,p,m}$ , the determinant of the coefficients is  $\Delta^T(t)$ , the transpose of  $\Delta(t)$  in (2.5); hence  $\Delta^T(t) = 0$ . So (3.12) has a nontrivial solution  $\{A_{i,p,m}\}$ .

Let  $a_m = -u_m(0)$ ,  $c_i = c_{i,0}$ . Then

$$(3.16) \quad \Delta^T(0) = \begin{pmatrix} a_m & 0 & \cdots & 0 & c_0 & 0 & \cdots & 0 \\ 0 & a_m & & 0 & 0 & c_1 & \cdots & 0 \\ & & & & & & & c_{r-1} \\ c_r & & & & a_m & & \cdots & 0 \\ 0 & c_{r+1} & & & & a_m & & 0 \\ 0 & & & c_{s-1} & & & & a_m \end{pmatrix}.$$

The elements  $a_m$  form the main diagonal. The other non-zero elements  $\{c_i\}$  lie one in each row and in each column, in the following position:  $c_i (i = 0, 1, \dots, s-1)$  is in the  $i$ th row and  $(s-r+i)$ th column (reduced (mod  $s$ ) to the range  $[0, s-1]$ ).

Let  $N_{i,m}(t)$  be the minor of  $\Delta^T(t)$  of order  $s-1$  obtained by omitting the last row and  $i$ th column ( $i = 0, 1, \dots, s-1$ ).

LEMMA 3.5. Let  $(s, k) = 1$ . Then

$$(3.17) \quad N_{i,m}(0) \neq 0 \quad (i = 0, \dots, s-1; m = 1, \dots, s);$$

so  $N_{i,m}(t) \neq 0$ .

*Proof.* Consider  $i = 0$ . From (3.16) we see that in  $N_{0,m}(0)$  every column except the  $(s-r-1)$ th and the  $(s-1)$ th (counting rows and columns from  $\Delta^T(0)$ ) has two nonzero elements, namely  $a_m$  and a  $c_i$ . Column  $(s-r-1)$  has only the nonzero  $a_m$ , and column  $(s-1)$  only  $c_{r-1}$ . In  $N_{0,m}(0)$ , row 0 has only the non-zero  $c_0$ , so if we expand with respect to this row we get a factor  $c_0$ . The  $(s-r)$ th column contains this  $c_0$ , so the remaining factor is an  $(s-2)$ -order determinant obtained by deleting row 0 and column  $s-r$ . This removes the  $a_m$  from row  $s-r$ , leaving only  $c_{s-r}$  in that row.

So we get a factor  $c_{s-r}$ , leaving an  $(s-3)$ -order determinant obtained by deleting row  $s-r$  and column  $2(s-r)$ . This removes  $a_m$  from row  $2(s-r)$ , leaving only  $c_{2(s-r)}$  in that row. And so on. This process continues until we have the factor

$$(3.18) \quad \pm c_0 c_{s-r} c_{2(s-r)} \cdots c_{(i-1)(s-r)},$$

where  $i$  in  $[0, s-1]$  is such that  $i(s-r) \equiv s-1 \pmod{s}$ . For when we reach the factor (3.18) then  $(i-1)(s-r) \equiv r-1$ , so the last term in (3.18) is  $c_{r-1}$ , which is in column  $s-1$ . Now there is no  $a_m$  in this column, so the process we began stops. If  $i = s-1$  then (3.18) is the value of  $N_{0,m}(0)$ . If  $i < s-1$ , there remains a determinant of order  $s-1-i$ , one of whose columns is the  $(s-r-1)$ th, and  $s-r-1 \equiv (i+1)(s-r)$ . This column contains only an  $a_m$ . Expanding in terms of this column we delete the  $c_{s-r-1}$  from column  $(s-r-1) + (s-r) \equiv (i+2)(s-r) \pmod{s}$ , so we get another factor  $a_m$ . And so on; so we have

$$(3.19) \quad N_{0,m}(0) = \pm a_m^{s-i-1} \cdot \prod_{j=0}^{i-1} c_{j(s-r)}.$$

Hence (3.17) holds for  $i = 0$ . A similar argument applies when  $i = 1, \dots, s-1$ .

Since  $N_{0,m}(t) \neq 0$ , the last equation of (3.12) is linearly dependent on the others, so it can be ignored. We can take  $A_{0,p,m}(t)$  arbitrarily (all  $p, m$ ) and solve for  $A_{i,p,m}(t)$ :

$$(3.20) \quad A_{i,p,m}(t) = (-1)^i N_{i,m}(t) [N_{0,m}(t)]^{-1} A_{0,p,m}(t) \\ (m = 1, \dots, s; i = 0, \dots, s-1; p = 0, \dots, k-1).$$

System (3.20) is equivalent to (3.12), so  $\{G_i\}$  given by (3.4), (3.13) satisfies (1.5) iff  $\{A_{i,p,m}(t)\}$  satisfies (3.15), (3.20).

Let  $\{A_{i,p,m}\}$  satisfy (3.15), (3.20). Substitute (3.20) into (3.15). We get

$$(3.21) \quad \frac{N_{i,m-1}(\zeta t)}{N_{0,m-1}(\zeta t)} \cdot A_{0,p,m-1}(\zeta t) = \zeta^i \frac{N_{i,m}(t)}{N_{0,m}(t)} \cdot A_{0,p,m}(t) \quad (\text{all } i, p, m).$$

From (3.15) with  $i = 0$ :

$$(3.22) \quad A_{0,p,m-1}(\zeta t) = A_{0,p,m}(t) \quad (\text{all } p, m).$$

So the set  $\{A_{0,p,m}(t)\}$  is not completely arbitrary. However, (3.22) will hold if we choose  $A_{0,p,m}(t) = 1$  (all  $p, m$ ), so from (3.21) we get

$$(3.23) \quad \frac{N_{i,m-1}(\zeta t)}{N_{0,m-1}(\zeta t)} = \zeta^i \frac{N_{i,m}(t)}{N_{0,m}(t)} \quad (\text{all } i, m).$$

Since the  $N_{i,j}$  series are independent of the  $A_{i,p,m}$  series, (3.23) holds no matter how we choose  $\{A_{0,p,m}\}$ .

*Remark.* Another proof of (3.23) can be given without use of  $\{A_{i,p,m}\}$ . We sketch it: Expand determinant  $N_{0,m}(\zeta t)$ , using  $u_i(\zeta t) = \zeta^{-r} u_{i+1}(t)$  from (3.1), so that

$$[t^{s-r} C_{i,s-r}(t) - u_m(t)]_{t \rightarrow \zeta t} = \zeta^{s-r} [t^{s-r} C_{i,s-r}(t) - u_{m+1}(t)].$$

The general term  $\pm a_{j_1,1} a_{j_2,2} \cdots a_{j_{s-1},s-1}$  is seen to have the factor  $\zeta$  to the power  $\sum_{i=1}^{s-1} \{s - r + j_i - i\}$ , and this is congruent (mod  $s$ ) to  $r + 1$ , since  $j_1, \dots, j_{s-1}$  is a permutation of  $0, 1, \dots, s - 2$ . Hence

$$(3.24) \quad N_{0,m}(\zeta t) = \zeta^{r+1} N_{0,m+1}(t).$$

By a similar argument we get

$$(3.25) \quad N_{i,m}(\zeta t) = \zeta^{r+i+1} N_{i,m+1}(t).$$

(3.23) follows from (3.24), (3.25).

LEMMA 3.6. *Let  $(s, k) = 1$ . Then (3.20), (3.22) are equivalent to (3.15), (3.20); so  $\{G_i\}$  given by (3.4), (3.13) satisfies (1.5) iff  $\{A_{i,p,m}\}$  satisfies (3.20), (3.22).*

*Proof.* We know that (3.15), (3.20) imply (3.20), (3.22). Now suppose (3.20), (3.22) hold. From (3.20) we get an expression for  $A_{i,p,m-1}(\zeta t)$  which, using (3.23), (3.22) and (3.20) leads to (3.15).

Take  $\{A_{0,p,0}(t)\}$  arbitrarily. From (3.22) we uniquely determine  $\{A_{0,p,-1}(t) = A_{0,p,s-1}(t)\}$ , then  $\{A_{0,p,s-2}\}$ , and so on. So choosing  $\{A_{0,p,0}\}$ , we uniquely determine  $\{A_{0,p,m}\}$  (all  $p, m$ ) to satisfy (3.22); and by means of (3.20) we determine all  $\{A_{i,p,m}\}$ .

THEOREM 3.1. *Let  $(s, k) = 1$ . The most general p.s. solution of (1.5) defined by (3.4) with (3.13) holding is obtained by taking  $\{A_{0,p,0}(t) = \sum_{n=0}^{\infty} a_{p,0,n} t^n\}$  arbitrarily subject to the condition*

$$(3.26) \quad \sum_{p=0}^{k-1} \omega^{pn} a_{p,0,0} \neq 0 \quad (n = 0, 1, \dots, k - 1),$$

and then uniquely determining the remaining  $\{A_{i,p,m}(t)\}$  by means of (3.22), (3.20).

*Proof.* All has been established except (3.26). If  $\{G_i\}$  satisfies (3.4), (3.13) then  $G(x, t) = \sum_{i=0}^{s-1} G_i(x, t) = \sum_0^{\infty} P_n(x)$  need not define a p.s., since we may have degree  $P_n(x) < n$  for some  $n$ . Now

$$(3.27) \quad G(x, t) = \sum_{m=1}^s \sum_{p=0}^{k-1} \sum_{i=0}^{s-1} A_{i,p,m}(t) \exp \{xI(\omega^p t u_m^*(t))\}$$

follows from (3.4). Using (3.20) this becomes

$$(3.28) \quad G(x, t) = \sum_{m=1}^s \sum_{p=0}^{k-1} R_m(t) A_{0,p,m}(t) \exp \{x I_{p,m}(t)\}$$

where

$$(3.29) \quad I_{p,m}(t) = I(\omega^p t u_m^*(t)), \quad R_m(t) = [N_{0,m}(t)]^{-1} \cdot \sum_{i=0}^{s-1} (-1)^i N_{i,m}(t).$$

Expand (3.28) in an  $x$ -series; we get

$$(3.30) \quad n! g_n(t) = \sum_{m=1}^s \sum_{p=0}^{k-1} R_m(t) A_{0,p,m}(t) [I_{p,m}(t)]^n.$$

Now  $[I_{p,m}(t)]^n = [e_1 \omega^p u_m^*(0)]^n t^n + \text{higher powers}$ , so if

$$(3.31) \quad R_m(t) = \sum_{j=0}^{\infty} r_{m,j} t^j, \quad A_{0,p,m}(t) = \sum_{j=0}^{\infty} a_{p,m,j} t^j$$

then

$$n! g_n(t) = \left\{ e_1^n \sum_{m=1}^s \sum_{p=0}^{k-1} r_{m,0} a_{p,m,0} \omega^{pn} [u_m^*(0)]^n \right\} t^n + \text{higher powers}.$$

We are to determine when the brace is non-zero.

Take  $k$ th roots in (3.1) and set  $t = 0$ :

$$(3.32) \quad u_m^*(0) = \zeta^{(m-1)r/k} u^*(0) \omega^{(m-1)q} = \zeta^{m-1} u^*(0).$$

By Lemma 2.2,  $u_m^*(0) \neq 0$ , so our condition becomes

$$(3.33) \quad \sum_{m=1}^s \sum_{p=0}^{k-1} r_{m,0} a_{p,m,0} \omega^{pn} \zeta^{(m-1)n} \neq 0 \quad (\text{all } n).$$

From (3.15) with  $i = 0$  we deduce

$$(3.34) \quad A_{0,p,m}(t) = A_{0,p,0}(\zeta^m t),$$

and taking  $t = 0$ :

$$(3.35) \quad a_{p,m,0} = a_{p,0,0}.$$

Hence

$$\sum_{p=0}^{k-1} a_{p,m,0} \omega^{pn} = \sum_{p=0}^{k-1} a_{p,0,0} \omega^{pn}.$$

So we are to have

$$(3.36) \quad \sum_{m=1}^s r_{m,0} \zeta^{mn} \left( \sum_{p=0}^{k-1} \omega^{pn} a_{p,0,0} \right) \neq 0.$$

Let

$$(3.37) \quad (-1)^i N_{i,m}(t) [N_{0,m}(t)]^{-1} = \sum_{j=0}^{\infty} d_{i,m,j} t^j,$$

so that

$$r_{m,0} = \sum_{i=0}^{s-1} d_{i,m,0}.$$

From (3.23) we have

$$d_{i,m-1,0} = \zeta^i d_{i,m,0}, \quad d_{i,m,0} = \zeta^{-mi} d_{i,s,0},$$

so

$$(3.38) \quad r_{m,0} = \sum_{i=0}^{s-1} \zeta^{-mi} d_{i,s,0}.$$

We are therefore to have

$$\sum_{i=0}^{s-1} d_{i,s,0} \left( \sum_{m=1}^s \zeta^{m(n-i)} \right) \left( \sum_{p=0}^{k-1} \omega^{pn} a_{p,0,0} \right) \neq 0.$$

The sum on  $m$  is  $s$  if  $i \equiv n \pmod{s}$  and 0 otherwise, so we must have

$$(3.39) \quad (sd_{n,s,0}) \sum_{p=0}^{k-1} \omega^{pn} a_{p,0,0} \neq 0.$$

(The  $n$  in  $d_{n,s,0}$  is to be reduced  $\pmod{s}$ .) Now  $d_{i,m,0} \neq 0$  by Lemma 3.5, so we obtain condition (3.26), to hold for all  $n$ ; hence for  $n \in [0, k-1]$ , since  $\omega^k = 1$ .

In (3.28) replace  $A_{0,p,m}(t)$  by its value from (3.34); then use the property  $A_{i,p \pm k,j} = A_{i,p,j}$ . We get

$$G(x, t) = \sum_{m=1}^s \sum_{p=0}^{k-1} R_m(t) A_{0,p,0}(\zeta^m t) \exp \{xI_{p,m}(t)\};$$

so

$$(3.40) \quad G(x, t) = \sum_{p=0}^{k-1} \sum_{m=1}^s \sum_{i=0}^{s-1} (-1)^i \frac{N_{i,m}(t)}{N_{0,m}(t)} \cdot A_{0,p,0}(\zeta^m t) \exp \{xI(\omega^p \zeta^{m-1} t u^*(\zeta^{m-1} t))\}.$$

There remains the question of the existence of p.s. solutions of (0.4) that do not have a g.f. of the form (3.40). On this point we have

**THEOREM 3.2.** *Let  $(s, k) = 1$ . A p.s.  $\{P_n(x)\}$  is a solution of (0.4) iff its g.f. is of the form (3.40), with  $\{A_{0,p,0}(t)\}$  subject to condition (3.26).*

*Proof.* The necessity is given by Theorem 3.1. Now let the p.s.  $\{P_n(x)\}$  satisfy (1.5), with g.f.

$$(3.41) \quad G(x, t) = \sum_0^\infty P_n(x) t^n = \sum_0^\infty g_n(t) x^n.$$

Let  $G(x, t)$  be given by (3.27). Then (3.28) holds iff (3.20) is satisfied. From (3.28) we pass to (3.40) by use of (3.34), which is equivalent to (3.22). We now show that conversely, if (3.28) and (3.40) hold then (3.22) is satisfied.

Define  $E_{p,m}(t)$  by

$$A_{0,p,m}(t) = A_{0,p,0}(\zeta^m t) + E_{p,m}(t).$$

Substitute this into (3.28) and form two double sums. The first one reduces to the right side of (3.40), so since (3.28) and (3.40) both represent  $G(x, t)$ , we have

$$\sum_{m=1}^s \sum_{p=0}^{k-1} R_m(t) E_{p,m}(t) \exp \{xI(\omega^p t u_m^*(t))\} = 0.$$

It follows by Lemma 3.3 that

$$R_m(t) E_{p,m}(t) = 0 \quad (\text{all } p, m).$$

The work of § 2 shows that for each root  $u(t)$  (with  $k$ th root  $u^*(t)$ ) there is a solution of form (2.8). Let  $m$  be fixed. Then there is a solution of (1.5) of the form

$$G(x, t) = \sum_{p=0}^{k-1} A_{p,m}(t) \exp \{xI(\omega^p tu_m^*(t))\}.$$

Hence  $u_m^*(t)$  must be present for all  $m = 1, \dots, s$  in the general solution  $G(x, t)$  of (1.5). Now if  $R_m(t) = 0$  for some  $m$ , say  $m = b$ , then from (3.28) we see that the general solution of form (3.27) does not involve  $u_b^*(t)$ . This contradiction shows that  $R_m(t) \neq 0$  for all  $m$ . So  $E_{p,m}(t) = 0$ , and therefore (3.22) holds.

By Lemma 1.2 it suffices to show that  $\{A_{0,p,0}(t)\}$  can be chosen in (3.40) so that when (3.40) is expanded in a series in  $x$ , then the coefficient of  $x^n$  is  $g_n(t) \cdot (n = 0, 1, \dots, k - 1)$ . The following system must be satisfied:

$$(3.42) \quad n! g_n(t) = \sum_{m=1}^s \sum_{p=0}^{k-1} \sum_{i=0}^{s-1} (-1)^i \frac{N_{i,m}(t)}{N_{0,m}(t)} \cdot A_{0,p,0}(\zeta^m t) [I(\omega^p \zeta^{m-1} tu^*(\zeta^{m-1} t))]^n$$

$(n = 0, 1, \dots, k - 1)$ .

Since  $\{A_{0,p,0}(\zeta^m t)\}$  represents a set of  $sk$  functions, we need  $sk$  equations, which we obtain by replacing  $t$  in (3.42) by  $\zeta t, \dots, \zeta^{s-1} t$ . Using (3.23), this leads to the system

$$(3.43) \quad n! g_n(\zeta^j t) = \sum_m \sum_p \sum_i (-1)^i \zeta^{jii} \frac{N_{i,m}(t)}{N_{0,m}(t)} X_{p,m}(t) [I(\omega^p \zeta^{m-1} tu^*(\zeta^{m-1} t))]^n$$

$(n = 0, \dots, k - 1; j = 0, \dots, s - 1)$ , where we write  $X_{p,m}(t)$  for  $A_{0,p,0}(\zeta^m t)$ . Define  $B_{p,m}(t), g_n^*(t)$  by

$$(3.44) \quad t B_{p,m}(t) = I(\omega^p \zeta^{m-1} tu^*(\zeta^{m-1} t)), \quad t^n g_n^*(t) = g_n(t).$$

We can cancel  $t^n$  from both sides of (3.43) to get

$$(3.45) \quad n! \zeta^{nj} g_n^*(\zeta^j t) = \sum_m \sum_p \sum_i (-1)^i \zeta^{jii} \frac{N_{i,m}(t)}{N_{0,m}(t)} B_{p,m}^n(t) X_{p,m}(t)$$

$(n = 0, \dots, k - 1; j = 0, \dots, s - 1)$ .

If the  $sk$ -order determinant  $\Theta(t)$  of the coefficients of  $\{X_{p,m}\}$  is  $\neq 0$  we can solve for the unknowns. We shall show that  $\Theta(0) \neq 0$ . In (3.45) take the order  $X_{0,1}, X_{0,2}, \dots, X_{0,s}; X_{1,1}, \dots, X_{1,s}; \dots; X_{k-1,1}, \dots, X_{k-1,s}$ . From (3.23) we derive

$$(3.46) \quad \zeta^{jii} \frac{N_{i,m}(t)}{N_{0,m}(t)} = \frac{N_{i,m-j}(\zeta^j t)}{N_{0,m-j}(\zeta^j t)}.$$

Also,  $B_{p,m}^n(t) = [e_1 \omega^p \zeta^{m-1} u^*(0)]^n + \text{higher powers}$ , so the coefficient of  $X_{p,m}$  in (3.45) has for  $t = 0$  the value

$$\left\{ \sum_{i=0}^{s-1} (-1)^i \frac{N_{i,m-j}(0)}{N_{0,m-j}(0)} \right\} [e_1 \omega^p \zeta^{m-1} u^*(0)]^n.$$

This is the general element in  $\Theta(0)$ . Now  $[e_1 u^*(0)]^n$  is a nonzero common factor of all the elements of a row of  $\Theta(0)$ . We drop this factor and denote the new determinant by  $\Theta^*$ . The general element of  $\Theta^*$  is  $R_{m-j} \omega^{pn} \zeta^{n(m-1)}$ , where  $R_j = R_j(0)$  is given by (3.29).

If we choose the rows of  $\Theta^*$  in the order  $(j, n) = (0, 0), (0, 1), \dots, (0, k - 1); (1, 0), \dots, (1, k - 1); \dots; (s - 1, 0), \dots, (s - 1, k - 1)$  then the corresponding elements of  $\Theta^*$  in row  $(j, n)$  are  $R_{1-j}, R_{2-j} \zeta^n, \dots, R_{s-j} \zeta^{n(s-1)}; R_{1-j} \omega^n, R_{2-j} \zeta^n \omega^n, \dots, R_{s-j} \zeta^{n(s-1)} \omega^n; \dots; R_{1-j} \omega^{(k-1)n}, \dots, R_{s-j} \zeta^{n(s-1)} \omega^{(k-1)n}$ . Indices on  $R$ 's are to be reduced (mod  $s$ ).

Determinant  $\Theta^*$  can be described as follows: Divide the matrix of  $\Theta^*$  into blocks,

each of  $k$  rows and  $s$  columns. Denote the upper left block (in rows  $0, 1, \dots, k-1$  and columns  $0, 1, \dots, s-1$ ) by  $[0, 0]$ , the one just below it by  $[0, 1]$  and the one just to the right by  $[1, 0]$ ; and so on. Then block  $[p, i]$  ( $p = 0, \dots, k-1; i = 0, \dots, s-1$ ) is

$$(3.47) \quad [p, i] = \begin{vmatrix} R_{s-i+1} & R_{s-i+2} & \cdots & R_{s-i+s} \\ \frac{R_{s-i+1}\omega^p}{R_{s-i+1}\omega^{(k-1)p}} & \frac{R_{s-i+2}\zeta\omega^p}{R_{s-i+2}\zeta^{k-1}\omega^{(k-1)p}} & \cdots & \frac{R_{s-i+s}\zeta^{s-1}\omega^p}{R_{s-i+s}\zeta^{(s-1)(k-1)}\omega^{(k-1)p}} \end{vmatrix}.$$

To show that  $\Theta^* \neq 0$  we need some lemmas. Let

$$(3.48) \quad R^* = \begin{vmatrix} R_1 & R_2 & \cdots & R_s \\ R_s & R_1 & \cdots & R_{s-1} \\ R_2 & R_3 & \cdots & R_1 \end{vmatrix}.$$

Regarding  $\{R_i\}$  as variables, the following identity is known:

$$(3.49) \quad R^* = \prod_{j=0}^{s-1} \left\{ \sum_{n=0}^{s-1} \zeta^{nj} R_{n+1} \right\}.$$

LEMMA 3.7. *Let  $(s, k) = 1$  and let  $\{R_m = R_m(0)\}$  be given by (3.29). Then*

$$(3.50) \quad R^* \neq 0.$$

*Proof.* From (3.25) we derive

$$N_{i,s-j}(\zeta^i t) = \zeta^{j(r+i+1)} N_{i,s}(t),$$

so

$$\sum_{n=0}^{s-1} \zeta^{nj} R_{n+1} = \zeta^{-j} \sum_{n=0}^{s-1} (-1)^n \left\{ \sum_{m=1}^s \zeta^{m(j-n)} \right\} \frac{N_{n,s}(0)}{N_{0,s}(0)}.$$

The brace is  $s$  if  $n = j$  and  $0$  otherwise, so by Lemma 3.5,  $R^* \neq 0$ .

LEMMA 3.8. *Regarded as a polynomial in the variables  $\{R_i\}$  the determinant  $\Theta^*$  is not identically zero.*

*Proof.* Let  $D_{sk} = \Theta^*$  when  $R_1 = 1, R_i = 0$  ( $i \neq 1$ ). The elements in each row contain  $\zeta$  to the same power, so we may remove  $\zeta$  completely from  $D_{sk}$ , obtaining a new determinant  $D_{sk}^*$ . In rows  $0, 1, \dots, k-1$  all elements in columns other than  $0, s, 2s, \dots, (k-1)s$  are zero; so using the Laplace expansion we express  $D_{sk}^*$  as the product of the nonzero  $k$ th order Vandermond  $V$  in the quantities  $1, \omega, \dots, \omega^{k-1}$ , and a determinant  $D_{(s-1)k}^*$  of order  $(s-1)k$ . Moreover  $D_{(s-1)k}^*$  is like  $D_{sk}^*$  in permitting a Laplace expansion, into the product of the same  $V$  and a determinant  $D_{(s-2)k}^*$  of order  $(s-2)k$ . This latter is likewise expansible, and so on. We finally evaluate  $D_{sk}$  as  $\zeta^a V^b$  ( $a, b$  positive integers). So  $D_{sk} \neq 0$ .

LEMMA 3.9. *Let  $(s, k) = 1$  and let  $R_m = R_m(0)$  be given by (3.29). Then*

$$(3.51) \quad \Theta^* \neq 0.$$

*Proof.* In  $\Theta^*$  subtract column 0 from each of columns  $s, 2s, \dots, (k-1)s$ ; column 1 from each of columns  $s+1, 2s+1, \dots, (k-1)s+1$ ; and so on. Finally subtract column  $s-1$  from each of columns  $2s-1, 3s-1, \dots, ks-1$ . We obtain zero elements in the part of all columns  $s, s+1, \dots, ks-1$  that are in rows  $0, k, 2k, \dots, (s-1)k$ , so by the Laplace expansion we express  $\Theta^*$  as the product of  $R^*$  (given by (3.48)) and a determinant  $E_{s(k-1)}$  of order  $s(k-1)$  whose columns are in the old columns  $s, s+1, \dots, ks-1$ .

$E_{s(k-1)}$  has blocks  $[p, i]$  ( $p = 0, \dots, k-2; i = 0, \dots, s-1$ ) each with  $k-1$  rows



and  $s$  columns:

$$(3.52) \quad [p, i] = \frac{R_{s-i+1}a_{(k-1)p+1} \quad R_{s-i+2}\zeta a_{(k-1)p+1} \quad \cdots \quad R_{s-i+s}\zeta^{s-1}a_{(k-1)p+1}}{R_{s-i+1}a_{(k-1)p+2} \quad R_{s-i+2}\zeta^2 a_{(k-1)p+2} \quad \cdots \quad R_{s-i+s}\zeta^{2(s-1)}a_{(k-1)p+2}} \cdot \frac{R_{s-i+1}a_{(k-1)p+k-1} \quad R_{s-i+2}\zeta^{k-1}a_{(k-1)p+k-1} \quad \cdots \quad R_{s-i+s}\zeta^{(k-1)(s-1)}a_{(k-1)p+k-1}}{R_{s-i+1}a_{(k-1)p+k-1} \quad R_{s-i+2}\zeta^{k-1}a_{(k-1)p+k-1} \quad \cdots \quad R_{s-i+s}\zeta^{(k-1)(s-1)}a_{(k-1)p+k-1}}$$

where the  $a$ 's involve  $\omega$ .

Now  $a_1 = \omega - 1, \neq 0$  for  $k > 1$ . When  $k = 1$  we have  $\Theta^* = \pm R^* \neq 0$  (by Lemma 3.7), so we consider  $k > 1$ . For later purpose we shall not use the above value of  $a_1$ . Rather, we observe that  $a_1, a_2, \dots, a_{k-1}$  cannot all be zero since then a column of  $E_{s(k-1)}$  would consist of zeros, making  $E_{s(k-1)}$ , hence  $\Theta^*$ , identically zero in  $\{R_i\}$ , contrary to Lemma 3.8. So at least one of  $a_1, \dots, a_{k-1}$  is nonzero, and it is no essential restriction to suppose that  $a_1 \neq 0$ .

From columns  $s, 2s, \dots, (k-1)s$  subtract respectively the following multiples of column 0:  $a_k/a_1, a_{2k-1}/a_1, a_{3k-1}/a_1$ , etc. Then from columns  $s+1, 2s+1, \dots, (k-1)s+1$  subtract the same multiples of column 1; and so on. A Laplace expansion gives us a product of  $R^*$  by a power of  $\zeta$ , by a new determinant  $E_{2(k-2)}$ . This latter has blocks  $[p, i]$  ( $p = 0, \dots, k-3; i = 0, \dots, s-1$ ) of  $k-2$  rows and  $s$  columns.  $[p, i]$  can be obtained from (3.52) as follows: Delete the top row, and in the remaining rows replace  $a_{(k-1)p+j}$  by  $b_{(k-1)p+j-1}$  ( $j = 2, \dots, k-1$ ), where the  $b$ 's involve  $\omega$ . Note that if we assume for example that  $a_2 \neq 0$  instead of  $a_1$ , the determinant that we would get for  $E_{s(k-2)}$  would be of the same character as the present  $E_{s(k-2)}$ , so it is not a restriction to take  $a_1 \neq 0$ .

Not all of  $b_1, \dots, b_{k-2}$  are zero (otherwise  $\Theta^* \equiv 0$  in  $\{R_i\}$ ); and generality is not lost in assuming  $b_1 \neq 0$ . We can then continue the process of reducing  $E_{s(k-2)}$ , and so on. We finally get

$$(3.53) \quad \Theta^* = R^{*k} \cdot E$$

where  $E$  involves  $\omega, \zeta$  and is independent of  $\{R_i\}$ . By Lemma 3.8,  $E \neq 0$ ; so  $\Theta^* \neq 0$ .

System (3.45) can then be solved by Cramer's rule, to give a unique solution  $\{X_{p,m}(t)\}$ ; and each  $X_{p,m}$  is a power series. If  $X_{p,m}(t)$  is to be identified with  $A_{0,p,0}(\zeta^m t)$  it should be verified that  $X_{p,m}(\zeta t) = X_{p,m+1}(t)$ . This is done as follows: In (3.45) change  $t$  to  $\zeta t$ . This effects a permutation of (3.45), with  $X_{p,m}(\zeta t)$  in place of  $X_{p,m+1}(t)$ ; so by uniqueness of solution these two functions are equal.

We conclude that the set  $\{A_{0,p,0}(t)\}$  ( $p = 0, \dots, k-1$ ) satisfies (3.42), so by Lemma 1.2 the given p.s. solution  $\{P_n(x)\}$  of (0.4) has a g.f. of the form (3.40). Condition (3.26) holds since  $\{P_n(x)\}$  is a p.s.

*Remark.* In Part II we shall treat the case  $(s, k) > 1$ .

REFERENCES

[1] W. A. AL-SALAM AND A. VERMA, *Generalized Sheffer polynomials*, Duke Math. J., 37 (1970), pp. 361-365.  
 [2] G.-C. ROTA, *Finite Operator Calculus*, Academic Press, Inc., New York, 1975.  
 [3] I. M. SHEFFER, *Some properties of polynomial sets of type zero*, Duke Math. J., 5 (1939), pp. 590-622.  
 [4] ———, *On polynomial sets of class  $S^k$ , and sets of rank  $k$* , Annali di Matematica Pura ed Applicata (Series 4), 118 (1978), pp. 295-324.

## A STRONGER LOGARITHMIC INEQUALITY SUGGESTED BY THE ENTROPY INEQUALITY\*

KENNETH B. STOLARSKY†

**Abstract.** Let  $p_1, \dots, p_n$  be  $n$  probabilities that sum to 1. The classical entropy inequality asserts that  $\sum np_i \log np_i$  is nonnegative. We show that  $np_i$  can be replaced here by  $(np_i)^\theta$  where  $\theta = 1 + (n-1)^{-1} - (\log n)^{-1}$ . This is a stronger result, and nearly best possible. For  $n = 2$  the best possible result follows from the nonnegativity of the coefficients of a certain class of power series.

**1. Introduction.** The classical entropy inequality (see, for example, [1, pp. 14-15], [2, p. 15], or [3, pp. 15-18]) asserts that for  $p_i \geq 0$  and

$$\sum_{i=1}^n p_i = 1$$

we have

$$(1.1) \quad - \sum_{i=1}^n p_i \log p_i \leq \log n.$$

If we introduce a new parameter  $\theta$ , this can be written as

$$(1.2) \quad S(\theta) = \sum_{i=1}^n (np_i)^\theta \log np_i \geq 0$$

where  $\theta = 1$ .

**THEOREM 1.** *The inequality (1.2) is true for*

$$(1.3) \quad \theta = \theta(n) \geq 1 + \frac{1}{n-1} - \frac{1}{\log n}.$$

Examination of  $S'(\theta)$  shows that the left side of (1.2) is *increasing* in  $\theta$ , so the theorem is stronger than the classical inequality. For  $n = 2$  inequality (1.2) is even true for  $\theta = \frac{1}{2}$ ; this is best possible, and is deduced in § 4 from Theorem 2, a result that asserts the nonnegativity of the coefficients of certain power series. For  $n \geq 3$  the optimal  $\theta$  is probably a transcendental number. It is possible (and suggested by what follows) that for large  $n$  and optimal  $\theta$ , there are cases of equality other than the case  $p_i = 1/n$  for all  $i$  (the sole case of equality in the classical entropy inequality).

For  $n$  large the general form of Theorem 1 cannot be improved too much. To see this, set

$$(1.4) \quad p_1 = \frac{1}{n} + \frac{\sqrt{n-1}}{n}, \quad p_i = \frac{1}{n} - \frac{1}{n\sqrt{n-1}} \quad (2 \leq i \leq n),$$

and

$$(1.5) \quad \theta = 1 - \frac{2 \log \log n}{\log \sqrt{2n}}.$$

Here again

$$(1.6) \quad \theta(n) \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

\* Received by the editors December 12, 1978, and in revised form April 23, 1979.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

However, for  $n$  large,

$$(1.7) \quad \begin{aligned} S(\theta) &= (1 + \sqrt{n-1})^\theta \log(1 + \sqrt{n-1}) + (n-1) \left(1 - \frac{1}{\sqrt{n-1}}\right)^\theta \log\left(1 - \frac{1}{\sqrt{n-1}}\right) \\ &\cong \sqrt{2n} (\log n)^{-2} \log \sqrt{2n} - .5\sqrt{n-1} < 0. \end{aligned}$$

For brevity we set

$$(1.8) \quad h(x) = 1 + \frac{1}{x-1} - \frac{1}{\log x},$$

so the expression on the right of (1.3) is  $h(n)$ . Note that we can remove the singularity of  $h(x)$  at  $x = 1$  by letting  $h(1) = \frac{1}{2}$ , and that since

$$(1.9) \quad h'(x) = x^{-1}(\log x)^{-2} - (x-1)^{-2} \geq 0$$

the function  $h(x)$  increases monotonically to 1 as  $x \rightarrow \infty$ .

**2. Reduction to a two variable problem.** Let  $x_i = np_i$  and  $x = (x_1, \dots, x_n)$ . For any  $n$ -vector  $u = (u_1, \dots, u_n)$ , the inequality  $u \geq 0$  shall mean  $u_i \geq 0$  for  $1 \leq i \leq n$ . We also define

$$G(u) = \sum_{i=1}^n u_i.$$

Theorem 1 now asserts that

$$(2.1) \quad F(x) = F_\theta(x) = \sum_{i=1}^n x_i^\theta \log x_i \geq 0$$

provided that

$$(2.2) \quad x \geq 0, \quad G(x) = n, \quad \text{and} \quad \theta = h(n).$$

Suppose Theorem 1 is false. Then there would be an  $n$ -vector  $u$  such that

$$(2.3) \quad u \geq 0, \quad G(u) = n, \quad \text{and} \quad F(u) < 0.$$

Clearly, for  $t > 1$  sufficiently large,

$$(2.4) \quad F(tu) > 0.$$

Hence by continuity there would be a number  $t = t(\theta) > 1$  such that

$$(2.5) \quad tu \geq 0, \quad G(tu) > n, \quad \text{and} \quad F(tu) = 0.$$

We conclude that it suffices to show the following.

**THEOREM A.** *Let  $\theta \cong h(n)$ . If*

$$(2.6) \quad x \geq 0 \quad \text{and} \quad F_\theta(x) = 0$$

then

$$(2.7) \quad G(x) \leq n.$$

*Remark.* If  $c > 0$  and

$$k(t) = t^c \log t,$$

then

$$-(ec)^{-1} \leq k(t) \leq 0, \quad 0 \leq t \leq 1;$$

in fact,  $k(t)$  is monotonically decreasing or increasing depending upon whether or not  $t \leq \exp(-c^{-1})$ . Thus the point set  $M$  defined by (2.6) is bounded. Aside from the isolated point  $0 = (0, 0, \dots, 0)$ , each point in  $M$  has some co-ordinate no less than 1, and some co-ordinate no greater than 1. In fact,  $M - \{0\}$  is a smooth compact manifold whose boundary is contained in the boundary of the first octant of  $n$ -space.

We shall prove Theorem A by induction on  $n$ . For  $n = 1$  it is trivial. Assume true for positive integers less than  $n$ . We proceed to establish it for  $n$ . First, if  $x_j = 0$  for some  $j$ , then

$$(2.8) \quad F(x) = \sum_{i \neq j} x_i^\theta \log x_i = 0.$$

Also, by the remark after (1.8),

$$(2.9) \quad h(n - 1) \leq \theta.$$

Thus, by the induction hypothesis,

$$(2.10) \quad G(x) = \sum_{i \neq j} x_i \leq n - 1 < n.$$

But since  $F(1, 1, \dots, 1) = 0$  and  $G(1, 1, \dots, 1) = n$ , the maximum value of  $G(x)$  must occur at an interior point. At each such point we have the Lagrange multiplier condition

$$(2.11) \quad \nabla F(x) = \mu \nabla G(x).$$

Thus

$$(2.12) \quad \mu = \theta x_i^{\theta-1} \log x_i + x_i^{\theta-1}, \quad (1 \leq i \leq n).$$

Define

$$(2.13) \quad f(t) = \theta t^{\theta-1} \log t + t^{\theta-1}.$$

Then

$$(2.14) \quad f'(t) = t^{\theta-2}[\theta(\theta - 1) \log t + (2\theta - 1)].$$

Hence  $f'(t)$  is positive for  $t$  small, negative for  $t$  large, and  $f'(t_0) = 0$  for a unique  $t_0$ . Thus a horizontal line cuts the graph of  $f(t)$  at most twice, and by (2.12) the  $x_i$  can assume at most two distinct values. Call these values  $x_0$  and  $y_0$ , and their respective multiplicities  $n_1$  and  $n_2$ . Thus

$$(2.15) \quad n_1 + n_2 = n \quad \text{and} \quad G(x) = n_1 x_0 + n_2 y_0.$$

If  $n_1$  or  $n_2$  vanishes (say for example  $n_2 = 0$ ) then

$$(2.16) \quad F(x) = n x_0^\theta \log x_0 = 0$$

and  $x_0 = 0$  or 1. This implies  $G(x) \leq n$ . If  $x_0$  or  $y_0$  vanishes (say for example  $y_0 = 0$ ) then we see similarly that  $x_0 = 0$  or 1. Since this implies  $G(x) \leq n_1 \leq n$ , we see that the difficulty lies in the case where  $n_1, n_2, x_0$  and  $y_0$  are all positive. Set  $\lambda = n_1/n$ , so  $1 - \lambda = n_2/n$  and

$$(2.17) \quad 0 < 1/n \leq \lambda \leq 1 - (1/n) < 1.$$

By dividing  $F(x)$  and  $G(x)$  by  $n$ , we see that it suffices to prove the following result.

**THEOREM B.** *If  $x_0 \geq 0, y_0 \geq 0, \theta \geq h(n)$ ,*

$$(2.18) \quad \lambda x_0^\theta \log x_0 + (1 - \lambda) y_0^\theta \log y_0 = 0,$$

and  $\lambda$  satisfies (2.17), then

$$(2.19) \quad \lambda x_0 + (1 - \lambda)y_0 \leq 1.$$

Henceforth we shall drop the subscripts from  $x_0$  and  $y_0$ . Thus Theorem B says that the graph  $\Gamma = \Gamma(\lambda)$  of

$$(2.20) \quad \lambda x^\theta \log x + (1 - \lambda)y^\theta \log y = 0$$

lies in the region bounded by  $x \geq 0, y \geq 0$ , and

$$(2.21) \quad \lambda x + (1 - \lambda)y \leq 1.$$

It is important to note that aside from the isolated point  $(x, y) = (0, 0)$ , the graph  $\Gamma$  lies in the union of the strips  $0 \leq y \leq 1 \leq x$  and  $0 \leq x \leq 1 \leq y$ , and is, moreover, *connected*. [To see the latter, first consider the part of  $\Gamma$  in  $y \leq 1 \leq x$ . Let  $y^*$  be the unique value of  $y$ , satisfying  $0 < y < 1$ , such that  $-y^\theta \log y$  is maximal, and choose  $x^*$  so that  $(x^*, y^*) \in \Gamma$ . Then as  $y$  increases continuously and monotonically from 0 to  $y^*$ , the value of  $x$  (uniquely determined by  $y$ ) increases continuously and monotonically from 0 to  $x^*$ . As  $y$  increases from  $y^*$  to 1, the value of  $x$  (again uniquely determined by  $y$ ) decreases continuously and monotonically from  $x^*$  to 1. The same considerations hold for  $x \leq 1 \leq y$ , and these two parts of  $\Gamma$  are joined at  $(1, 1)$ .]

**3. The proof.** We shall show it suffices to prove the following.

**THEOREM C.** *If  $x \geq 0, y \geq 0, \theta \geq h(n)$ , and*

$$(3.1) \quad \lambda x + (1 - \lambda)y = 1$$

*then (i) for  $(x, y) \neq (1, 1), (1, 0), (0, 1)$ , or  $(0, 0)$  we have the strict inequality*

$$(3.2) \quad F(x, y) = \lambda x^\theta \log x + (1 - \lambda)y^\theta \log y > 0,$$

*and (ii) there is a neighborhood of  $(1, 1)$  such that*

$$(3.3) \quad \lambda x + (1 - \lambda)y \geq 1$$

*implies*

$$(3.4) \quad F(x, y) = \lambda x^\theta \log x + (1 - \lambda)y^\theta \log y \geq 0$$

*with equality only for  $(x, y) = (1, 1)$ .*

To see this sufficiency, note that if Theorem B were false, then by Theorem C the graph of  $\Gamma$  would not be connected. We now prove Theorem C. Part (ii) is easily established for  $\theta > \frac{1}{2}$  since the Taylor series expansion of  $F(x, y)$  about  $(1, 1)$  is

$$(3.5) \quad F(x, y) = (\lambda x + (1 - \lambda)y - 1) + (\theta - \frac{1}{2})\{\lambda(x - 1)^2 + (1 - \lambda)(y - 1)^2\} + O[(x - 1)^3] + O[(y - 1)^3].$$

For part (i) we can assume, without loss of generality, that

$$(3.6) \quad 0 < y < 1 < x.$$

From (3.1) we also see that

$$(3.7) \quad x < 1/\lambda \leq n.$$

Since

$$(3.8) \quad \lambda = (1 - y)/(x - y),$$

inequality (3.2) becomes

$$(3.9) \quad (1 - y)x^\theta \log x + (x - 1)y^\theta \log y > 0$$

or

$$(3.10) \quad g(y) < g(x)$$

where, for all real  $x$ , the function  $g(x)$  is defined by

$$(3.11) \quad g(x) = \frac{x^\theta \log x}{x - 1}.$$

Note that  $x = 1$  is a removable singularity for  $g(x)$ . Now

$$(3.12) \quad g'(x) = \frac{x^{\theta-1}}{(x-1)^2} [(x-1)(\theta \log x + 1) - x \log x].$$

If  $g'(x) = 0$ , then clearly  $\theta = h(x)$ . However, by (3.7) and the remarks following (1.8), we see that

$$(3.13) \quad h(x) < h(n) \leq \theta.$$

Hence  $g'(x)$  does not change sign for  $0 \leq x < n$ ; an examination of small values of  $x$  shows that  $g'(x) > 0$  here. Hence the strict inequality (3.2) is valid, and Theorems 1, A, B, and C are proved.

**4. Some nonnegative power series.** We shall show that if  $x, y > 0$  and  $x + y = 2$ , then

$$(4.1) \quad x^{1/2} \log x + y^{1/2} \log y \geq 0,$$

and that equality holds only for  $x = y = 1$ . This is equivalent to

$$(4.2) \quad F(z) = F_\theta(z) = (1 + z)^\theta \log(1 + z) + (1 - z)^\theta \log(1 - z) \geq 0$$

for

$$(4.3) \quad \theta = \frac{1}{2} \quad \text{and} \quad 0 \leq z \leq 1.$$

By examining the first few terms in the power series expansion of  $F_\theta(z)$ , we easily see that (4.2) is false for  $\theta < \frac{1}{2}$ .

We shall in fact establish the following stronger result.

**THEOREM 2.** *Every coefficient in the power series expansion of  $F_{1/2}(z)$  about  $z = 0$  is nonnegative.*

Before establishing this, we obtain some preliminary results.

**DEFINITION.** Call the sequence  $a_i$ , where  $i = 1, 2, 3, \dots$ , *convolution nonincreasing* if the associated sequence

$$(4.4) \quad b_k = \sum_{i=1}^{k-1} a_i a_{k-i}, \quad k = 2, 3, 4, \dots,$$

is nonincreasing for  $k \geq 2$ .

**LEMMA.** *The sequence of reciprocals  $1/i$ , for  $i = 1, 2, 3, \dots$  is convolution nonincreasing.*

*Proof.* By partial fractions, the inequality  $b_{k+1} \leq b_k$  is in this case equivalent to

$$(4.5) \quad \sum_{i=1}^k \frac{1}{i} + \sum_{i=1}^k \frac{1}{k+1-i} \leq \left(1 + \frac{1}{k}\right) \sum_{i=1}^{k-1} \left[\frac{1}{i} + \frac{1}{k-i}\right],$$

which is in turn equivalent to the obvious fact that

$$(4.6) \quad \frac{2}{k} \leq \frac{1}{k} \sum_{i=1}^{k-1} \left[ \frac{1}{i} + \frac{1}{k-i} \right].$$

Next, let  $\omega^\alpha$  denote the principal value of the  $\alpha$ th power, so that  $(1+x)^\alpha$  is given by the usual binomial expansion for  $|x| < 1$ . We write  $p(z) \gg 0$  to indicate that the power series expansion of  $p(z)$  about  $z = 0$  has nonnegative coefficients.

**THEOREM 3.** *Let  $0 \leq \alpha \leq 1$ . Suppose that*

$$(4.7) \quad 0 \leq c_{i+1} \leq c_i \leq 1, \quad 1 \leq i < \infty.$$

For  $|z| < 1$  define

$$(4.8) \quad g(z) = 1 + \sum_{n=1}^{\infty} (-1)^n c_n z^n$$

and

$$(4.9) \quad h(z) = g(z)^\alpha.$$

Then all power series coefficients of

$$(4.10) \quad P(z) = z(1+z)^\alpha h(z) + (-z)(1-z)^\alpha h(-z)$$

are nonnegative.

*Proof.* We have

$$(4.11) \quad \begin{aligned} (1-z)^\alpha h(-z) &= \{(1-z)(1+c_1z+c_2z^2+\dots)\}^\alpha \\ &= \{1+(c_1-1)z+(c_2-c_1)z^2+\dots\}^\alpha = \{1-p(z)\}^\alpha \\ &= 1-\alpha p(z) + \frac{\alpha(\alpha-1)}{2!} p^2(z) - \dots = 1-q(z) \end{aligned}$$

where  $p(z) \gg 0$  and hence  $q(z) \gg 0$ . Thus

$$(4.12) \quad P(z) = z[q(z) - q(-z)] \gg 0.$$

To prove Theorem 2 take  $\alpha = \frac{1}{2}$  and

$$(4.13) \quad g(z) = [z^{-1} \log(1+z)]^2;$$

that the  $c_i$  satisfy (4.7) here follows from the lemma. Thus

$$(4.14) \quad zh(z) = \log(1+z),$$

so  $P(z) = F_{1/2}(z)$  and the result follows.

REFERENCES

[1] L. BRILLOUIN, *Science and Information Theory*, 2nd ed., Academic Press, New York, 1962.  
 [2] A. FEINSTEIN, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.  
 [3] R. J. MCELIECE, *The Theory of Information and Coding*, Encyclopedia of Mathematics and its Applications, Addison-Wesley, Reading, MA, 1977.

## ON DIRICHLET'S PROBLEM FOR ELLIPTIC EQUATIONS IN SECTIONALLY SMOOTH $n$ -DIMENSIONAL DOMAINS\*

A. AZZAM†

**Abstract.** This paper is concerned with the first boundary value problem for linear second order elliptic equations in a domain  $\Omega \in R^n (n \geq 2)$  with edges on its boundary. Conditions sufficient for the solution  $u$  to be in  $C_\nu(\bar{\Omega})$ ,  $1 < \nu < 2$ , are given. Further statements concern the nature of singularities which the second partial derivatives of the solution may have at the edges.

For smooth domains, smoothness properties of solutions of general boundary value problems for linear equations have been thoroughly investigated; cf. [1]. For domains with piecewise smooth boundary and for general second-order elliptic equations, however, very little is known about the smoothness (up to the boundary) of the solution; cf. [2]–[6], [8]. In this paper we consider linear elliptic equations in sectionally smooth  $n$ -dimensional domains ( $n \geq 2$ ), especially near an edge.

*Notations.*  $u_{|i}$  denotes  $\partial u / \partial x_i$ , and  $u_{|ij}$  denotes  $\partial^2 u / \partial x_i \partial x_j$ . We also use the summation convention, that is, we sum over an index that appears twice (e.g.,  $a_i u_{|i} = \sum_{i=1}^n a_i (\partial u / \partial x_i)$ ). Furthermore,  $x = x_1, \dots, x_n$ ,  $|x|^2 = x_1^2 + \dots + x_n^2$ . The distance between points  $P$  and  $Q$  will be denoted by  $PQ$ .

In this paper we consider the Dirichlet problem for the uniformly elliptic equation

$$(1) \quad Lu \equiv a_{ij}(x)u_{|ij} + a_i(x)u_{|i} + a(x)u = f(x)$$

in a domain  $\Omega$  the boundary  $\Gamma$  of which consists of  $(n-1)$ -dimensional surfaces  $\Gamma_1, \Gamma_2, \dots, \Gamma_k$  belonging to  $C_{2+\alpha}$ ,  $0 < \alpha < 1$ . We assume that the surface  $\Gamma_i$  intersects only with  $\Gamma_{i-1}$  and  $\Gamma_{i+1}$  along  $(n-2)$ -dimensional manifolds  $S_{i-1}$  and  $S_i$ . We study in detail the case  $k=2$ , the behavior of the solution in the neighborhood of other manifolds may be similarly studied. Let  $\Gamma = \Gamma_1 \cup \Gamma_2$ ,  $\Gamma_1 \cap \Gamma_2 = S$  and  $P \in S$ . Let  $R_1$  and  $R_2$  be the planes which touch  $\Gamma_1$  and  $\Gamma_2$  at  $P$  making an angle  $\gamma(P)$ . We transform the equation

$$(2) \quad a_{ij}(P)u_{|ij} = 0$$

to canonical form. This equation is an equation with constant coefficients since the point  $P$  is fixed. After the transformation, the planes  $R_1$  and  $R_2$  will be transformed to other planes with angle  $\omega(P)$  between them. It is clear that  $\omega(P)$  does not depend upon the way used to transform (2) to canonical form. From [1] it follows that if the right hand side and the coefficients of (1) belong to  $C_\alpha(\bar{\Omega})$ , and if the boundary value  $\psi$  of  $u(x)$  is continuous on  $\Gamma$  and belongs to  $C_{2+\alpha}(\Gamma \setminus S)$  then  $u(x) \in C_{2+\alpha}(\bar{\Omega} \setminus S) \cap C_0(\bar{\Omega})$ . We prove the following

**THEOREM 1.** *If for any  $P \in S$ ,  $\omega(P) < \pi$ , then there exists a number  $\nu$ ,  $1 < \nu < 2$  such that  $u(x) \in C_\nu(\bar{\Omega})$ .*

We first prove this theorem in a special setting.

Consider the two hyperplanes  $x_1 = x_2 \tan \beta$  and  $x_1 = x_2 \tan(\omega + \beta)$ ,  $x_1 \geq 0$  intersecting at an  $(n-2)$ -dimensional space  $S_0$  with the angle  $\omega < \pi$ , where  $\pi/2 < \omega + 2\beta < \pi$ . By  $G_{r_0}$  we denote the part of the sphere centered at the origin with radius  $r_0 > 0$  which is included between the two hyperplanes. By  $\Gamma_{r_0}$  we denote the part of the boundary of  $G_{r_0}$  lying on the hyperplanes. Let  $S_{r_0} = S_0 \cap \Gamma_{r_0}$ .

\* Received by the editors October 23, 1978, and in revised form April 19, 1979.

† Department of Mathematics, University of Windsor, Windsor, Ontario, Canada NGB 3P4. This work is part of the Ph.D. thesis of the author done at Moscow State University under the supervision of Prof. V. A. Kondrat'ev.



**THEOREM 2.** *Suppose that in  $G_d(d < 1)$  the function  $u(x)$  satisfies the uniformly elliptic equation*

$$(3) \quad L_0 u \equiv a_{ij}^0(x)u_{|j} + a_i^0(x)u_{|i} + a^0(x)u = f^0(x),$$

where

- (i)  $a_{ij}^0(0) = \delta_{ij}$ ,  $i, j = 1, \dots, n$ .  $\delta_{ij}$  is the Kronecker delta.
- (ii)  $a_{ij}^0, a_i^0, a^0$  and  $f^0$  belong to  $C_\alpha(\bar{G}_d)$ ,  $0 < \alpha < 1$ .
- (iii)  $u_{|\Gamma_d} = \psi_0 \in C_{2+\alpha}(\Gamma_d \setminus S_d) \cap C_0(\Gamma_d)$ .
- (iv)  $\psi_0$  vanishes on  $S_d$  together with its first derivatives in the directions  $\theta = \beta$  and  $\theta = \omega + \beta$ .

Then there exists a number  $\nu$ ,  $1 < \nu < 2$ , such that  $u(x) \in C_\nu(\bar{G}_{r_0})$ , provided that  $4r_0 < d$ .

To prove this theorem we need two lemmas and we will make use of the following well known a priori estimate [1].

In the  $n$ -dimensional domain  $\Omega$  with boundary  $\bar{\Omega}$  we consider a bounded solution of the elliptic equation  $Lu = F(x)$  (cf. (1)) which coincides on  $\Gamma \subset \bar{\Omega}$  with a given function  $\Phi$ . Consider a subdomain  $\Omega_1$  of  $\Omega$  with the property that  $\bar{\Omega}_1 \cap \bar{\Omega}$  lies in the interior of  $\Gamma$ . If the coefficients of  $L$  and  $F(x)$  belongs to  $C_\alpha(\bar{\Omega})$  and if  $\Gamma \in C_{2+\alpha}$  and  $\Phi \in C_{2+\alpha}(\Gamma)$  then

$$\|u\|_{2+\alpha}^{\Omega_1} \leq c_0[\|u\|_0^\Omega + \|F\|_\alpha^\Omega + \|\Phi\|_{2+\alpha}^\Gamma],$$

where the constant  $c_0$  is independent of  $u$ .

**LEMMA 1.** *The solution  $u(x)$  of (3) satisfies the inequality  $|u(x)| \leq Mr^\nu$  in  $G_{2r_0}$ , where  $1 < \nu < 2$ ,  $r^2 = x_1^2 + x_2^2$  and  $M$  is independent of  $x$ .*

*Proof.* Consider the function  $\zeta(|x|) \in C_3$  in  $G_{4r_0}$ , where

$$\zeta(|x|) \equiv 1 \quad \text{if } 0 \leq |x| \leq 2r_0,$$

and

$$\zeta(|x|) \equiv 0 \quad \text{if } 3r_0 \leq |x| \leq 4r_0.$$

The function  $W(x) = \zeta u$  is defined for  $|x| \leq 4r_0$ ,  $x \in G_d$  and satisfies there the elliptic equation

$$(4) \quad L_0 W = F \equiv \zeta f^0 - 2a_{ij}^0 \zeta_{|i} u_{|j} - a_{ij}^0 \zeta_{|ij} u - a_i^0 \zeta_{|i} u.$$

On  $\Gamma_{4r_0}$ ,  $W(x)$  coincides with a function  $\psi_0$  satisfying conditions (iii) and (iv) of Theorem 2, with  $d$  replaced by  $4r_0$ . Consider the function

$$V = -Mr^\nu \sin \lambda \theta,$$

where  $r^2 = x_1^2 + x_2^2$ ,  $\theta = \arctan x_2/x_1$ ,  $M > 0$  and  $1 < \nu < \lambda = \pi/(\omega + 2\beta) < 2$ . Then

$$\begin{aligned} L_0 V &= V_{111} + V_{122} + \sum_{i,j=1}^2 (a_{ij}^0 - \delta_{ij})V_{|ij} + a_1^0 V_{|1} + a_2^0 V_{|2} + a^0 V \\ &= M(\lambda^2 - \nu^2)r^{\nu-2} \sin \lambda \theta + M \sum_{i,j=1}^2 H_{ij}(a_{ij}^0 - \delta_{ij})r^{\nu-2} + Mh_1 r^{\nu-1} + Mh_2 r^\nu, \end{aligned}$$

where  $H_{ij}$  and  $h_i$  are bounded functions of  $\theta, \lambda$  and  $\nu$ . We put

$$\sum_{i,j=1}^2 |H_{ij}| + |h_1| + |h_2| \leq A, \quad A > 0.$$

Since  $a_{ij}^0(x)$  are continuous functions and  $a_{ij}^0(0) = \delta_{ij}$ , for any  $\varepsilon > 0$  we can find  $r_0 > 0$  such that for  $|x| < 4r_0$  we have

$$|a_{ij}^0(x) - \delta_{ij}| < \varepsilon/4A, \quad i, j = 1, 2.$$

Thus in  $G_{4r_0}$  we have

$$L_0V \cong M[(\lambda^2 - \nu^2) \sin \lambda\beta - \varepsilon]r^{\nu-2} - MAr^{\nu-1} - MAr^\nu.$$

Choosing  $\varepsilon < (\lambda^2 - \nu^2) \sin \lambda\beta$  and  $r_0$  sufficiently small, we obtain  $L_0V \cong |F(x)|$ , hence  $L_0(W - V) \leq 0$ , in  $G_{4r_0}$ . We now prove that on the boundary of  $G_{4r_0}$ ;  $W - V \geq 0$  if  $M$  is chosen sufficiently large. On  $\Gamma_{4r_0}$  we have

$$W - V = \psi_0(x) + Mr^\nu \sin \lambda\beta.$$

At any point  $x = x_1, \dots, x_n$  on  $\Gamma_{4r_0}$  we have

$$\psi'_0(x) = \int_{(0,0,x_3,\dots,x_n)}^x \psi''_0 dt,$$

where  $t$  is the direction from  $(0, 0, x_3, \dots, x_n)$  to  $x$  and  $\psi'_0$  and  $\psi''_0$  are the derivatives of  $\psi_0$  in this direction. Since  $\psi''_0$  is bounded, say,  $|\psi''_0(x)| \leq 2k$ , we have  $|\psi'_0(x)| \leq 2kr$ . Similarly  $|\psi_0(x)| \leq kr^2$ . Thus on  $\Gamma_{4r_0}$  we obtain

$$W - V \geq -kr^2 + Mr^\nu \sin \lambda\beta \geq (M \sin \lambda\beta - k)r^\nu.$$

We choose  $M \geq k/\sin \lambda\beta$ . On  $|x| = 4r_0$ ,  $x \in G_d$  we have

$$W - V = Mr^\nu \sin \lambda\theta \geq 0.$$

Thus  $L(W - V) \leq 0$  inside  $G_{4r_0}$  while  $W - V \geq 0$  on the boundary. If we choose  $r_0$  sufficiently small, we may apply the maximum principle for domains with small diameter [7] in  $G_{4r_0}$ . Thus  $W - V \geq 0$ , or  $W \geq -Mr^\nu \sin \lambda\theta \geq -Mr^\nu$  in  $G_{4r_0}$ . Taking  $r_0$  sufficiently small and  $M$  sufficiently large we prove similarly that  $W \leq Mr^\nu$  in  $G_{4r_0}$ . This proves the lemma since  $W \equiv u$  in  $G_{2r_0}$ .

LEMMA 2. At any point  $x \in G_{r_0}$ ,  $|\partial u/\partial x_i| \leq M_1 r^{\nu-1}$ ,  $i = 1, \dots, n$ , where  $M_1$  is independent of  $x$ .

Proof. Consider the following domains in  $G_{2r_0}$ :

$$D_p = \left\{ x, x \in G_{2r_0} \frac{r_0}{2^{p+2}} \leq r \leq \frac{r_0}{2^{p+1}}, |x_i| < \frac{r_0}{2^p}, i > 2 \right\},$$

$$D'_p = D_{p-1} \cup D_p \cup D_{p+1}.$$

Here  $r^2 = x_1^2 + x_2^2$ . By  $\Gamma'_p$  we denote the part of the boundary of  $D'_p$  lying on the hyperplanes. Consider the transformation

$$(5) \quad x_i = x'_i/2^p, \quad i = 1, 2, \dots, n.$$

This transformation transforms  $D_p$  and  $D'_p$  onto  $D_0$  and  $D'_0$ , respectively. In  $D'_0$  the function  $V(x') = u(x'/2^p)$  satisfies the elliptic equation

$$(6) \quad b_{ij}(x')V_{|ij} + \frac{1}{2^p} b_i(x')V_{|i} + \frac{1}{2^{2p}} b(x')V = \frac{1}{2^{2p}} g(x'),$$

where  $b_{ij}(x') = a_{ij}^0(x'/2^p)$  and  $b_i, b$  and  $g$  are defined similarly. On  $\Gamma'_0$  the values of  $V(x'), \phi(x') = \psi_0(x'/2^p)$  belong to  $C_{2+\alpha}$ . Applying the a priori estimates in  $D_0$  and  $D'_0$

we get

$$(7) \quad \|V\|_{2+\alpha}^{D_0} \leq c_0 \left[ \|V\|_0^{D_0} + \frac{1}{2^{2p}} \|g\|_{\alpha}^{D_0} + \|\phi\|_{2+\alpha}^{\Gamma_0} \right].$$

As before we can show that  $\|\phi\|_{2+\alpha}^{\Gamma_0} \leq k_2(1/2^p)^2$ . Also since  $\|V\|_0^{D_0} = \|u\|_0^{D_0} \leq k_0(1/2^p)^\nu$  and since  $\|g\|_{\alpha}^{D_0} \leq k_1$ , we obtain from (7)  $\|V\|_{2+\alpha}^{D_0} \leq k_3(1/2^p)^\nu$ . Returning to the  $x$ -coordinates, and noticing that  $|\partial V/\partial x'_i| \leq \|v\|_{2+\alpha}$  and that  $\partial V/\partial x'_i = (1/2^p)(\partial u/\partial x_i)$  we get  $|\partial u/\partial x_i| \leq M_1 r^{\nu-1}$ .

*Remark 1.* Similarly we can prove that  $|\partial^2 u/\partial x_i \partial x_j| \leq M_2 r^{\nu-2}$ .

*Proof of Theorem 2.* Consider any two points  $P$  and  $Q$  in  $\bar{G}_{r_0}$  with distances  $r_1$  and  $r_2$  from  $S_0$ , where  $0 \leq r_2 \leq r_1 \leq r_0$ . If  $r_2 \leq \frac{1}{2}r_1$  then  $\overline{PQ} \geq \frac{1}{2}r_1$  and  $|u'(P) - u'(Q)|/\overline{PQ}^{\nu-1} \leq 2M_1 r_1^{\nu-1}/(\frac{1}{2}r_1)^{\nu-1} = H$  where  $u'$  is any of the derivatives  $\partial u/\partial x_i$ . If  $r_2 > \frac{1}{2}r_1$  we consider the domain

$$D_P = \left\{ x \in G_{r_0}, \frac{r_1}{2} \leq r \leq r_1, |x_i - x_i^0| \leq \frac{r_1}{2}, i = 3, \dots, n \right\},$$

where  $r^2 = x_1^2 + x_2^2$  and  $(x_1^0, \dots, x_n^0)$  are the coordinates of  $P$ . The transformation

$$x_i = \frac{2r_1 x'_i}{r_0}, \quad i = 1, 2,$$

$$x_i - x_i^0 = \frac{2r_1}{r_0}(x'_i - x_i^0) \quad i > 2,$$

transforms  $D_P$  into  $D'_P$  where

$$D'_P = \left\{ \frac{r_0}{4} \leq r' \leq \frac{r_0}{2}, |x'_i - x_i^0| \leq \frac{r_0}{4}, i > 2 \right\},$$

$r'^2 = x_1'^2 + x_2'^2$ . In  $D'_P$  the new function  $V(x')$  satisfies the elliptic equation

$$c_{ij}(x')v_{|ij} + \frac{2r_1}{r_0}c_i(x')v_{|i} + \left(\frac{2r_1}{r_0}\right)^2 c(x')v = \left(\frac{2r_1}{r_0}\right)^2 h(x'),$$

$c_{ij}$ ,  $c_i$ ,  $c$ ,  $h$  and  $\psi_1$  are the transformed functions  $a_{ij}^0$ ,  $a_i^0$ ,  $a^0$ ,  $f^0$  and  $\psi_0$ . Consider

$$D''_P = \left\{ \frac{r_0}{8} \leq r' \leq r_0, |x'_i - x_i^0| \leq \frac{r_0}{4}, i > 2 \right\}.$$

In  $D'_P$  and  $D''_P$  we apply the a priori estimate again, finding

$$\|v\|_{2+\alpha}^{D'_P} \leq C_0 \left[ \|v\|_0^{D''_P} + \left(\frac{2r_1}{r_0}\right)^2 \|h\|_{\alpha}^{D''_P} + \|\psi_1\|_{2+\alpha}^{\Gamma''_P} \right],$$

where  $\Gamma''_P$  is the part of the boundary of  $D''_P$  which lies on the hyperplanes. As before we can show that

$$\|v\|_{2+\alpha}^{D'_P} \leq M_0 r_1^\nu.$$

Noting that  $\|v\|_v^{D'_P} \leq \gamma \|v\|_{2+\alpha}^{D'_P}$ ,  $v' = (2r_1/r_0)u'$  and that  $H_{\nu-1}^{D'_P}(v') = (2r_1/r_0)^\nu H_{\nu-1}^{D_P}(u')$ , we get  $H_{\nu-1}^{D'_P}(u') \leq M'_0$  or equivalently  $u(x) \in C_\nu(D_P)$ .

Now consider the case  $r_2 > \frac{1}{2}r_1$ . In addition to  $P$  and  $Q$  we consider the point  $P_1$  lying on the normal from  $Q$  to  $S_0$  with distance  $r_1$  from  $S_0$ . If  $PP_1 \leq \frac{1}{2}r_1$ , then  $Q \in D_P$

where  $u \in C_\nu$ . If  $PP_1 > \frac{1}{2}r_1$ , then  $\overline{PQ} \cong \overline{PP_1} > \frac{1}{2}r_1$  and  $\overline{PQ} \cong \overline{P_1Q}$  and

$$\begin{aligned} \frac{|u'(P) - u'(Q)|}{\overline{PQ}^{\nu-1}} &\leq \frac{|u'(P) - u'(P_1)|}{\overline{PP_1}^{\nu-1}} + \frac{|u'(P_1) - u'(Q)|}{\overline{P_1Q}^{\nu-1}} \\ &\leq \frac{2M_1r_1^{\nu-1}}{(\frac{1}{2}r_1)^{\nu-1}} + H_1 = H_2 \end{aligned}$$

since  $Q \in D_{P_1}$ . This proves the theorem.

*Remark 2.* Similarly we can prove that  $r^\tau u'' \in C_{\tau_0}(\overline{G}_{r_0})$  where  $\tau$  and  $\tau_0$  satisfy  $0 < \tau_0 = \tau - 2 + \nu < 1$ .

*Proof of Theorem 1.* It is sufficient to prove that  $u \in C_\nu(\overline{G}_{P,\rho_0})$ , where  $G_{P,\rho_0}$  is the intersection of the domain  $\overline{\Omega}$  with a sphere of radius  $\rho_0$  centered at any point  $P \in S$ . Let  $P = (x_1^0, \dots, x_n^0)$  and let the two surfaces intersecting at  $S$  have the representations  $x_1 = g(x_2, \dots, x_n)$  and  $x_2 = h(x_1, x_3, \dots, x_n)$  in the neighborhood of the point  $P$ , where  $g$  and  $h$  belong to  $C_{2+\alpha}$ . The transformation

$$(8) \quad \begin{aligned} y_1 &= x_1 - g, \\ y_2 &= x_2 - h, \\ y_i &= x_i - x_i^0, \quad i > 2, \end{aligned}$$

takes the point  $P$  to the new origin. The two surfaces  $x_1 = g$  and  $x_2 = h$  will be transformed to the planes  $y_1 = 0$  and  $y_2 = 0$ . Equation (1) will be transformed to another elliptic equation. Suppose that the transformation

$$(9) \quad z_i = \sum_{j=1}^n \alpha_{ij} y_j, \quad i = 1, \dots, n,$$

transforms this last equation to the equation

$$(10) \quad d_{ij}(z)u_{|ij} + d_i(z)u_{|i} + d(z)u = t(z),$$

where  $d_{ij}(0) = \delta_{ij}$ ,  $i, j = 1, 2, \dots, n$ .

This transformation always exists and its Jacobian is different from zero. The two planes  $y_1 = 0$  and  $y_2 = 0$  will be transformed to others with angle  $\omega = \omega(P)$  between them,  $\omega < \pi$ . We finally use a transformation such that the two hyperplanes have the equations  $\eta_2 = \eta_1 \tan \beta$  and  $\eta_2 = \eta_1 \tan(\omega + \beta)$ , where  $\beta > 0$ , and  $\pi/2 < \omega + 2\beta < \pi$ . Any subdomain  $G_{P,\rho_1} \subset \Omega$  will be transformed into  $G'_{0,\rho_2}$  lying between the two hyperplanes. In  $G_{0,d} = G_d \subset G'_{0,\rho_2}$  the new function  $u_1(\eta) = u(x)$  will satisfy an elliptic equation of the type (3), satisfying the conditions (i)–(iii). Let the boundary values of  $u_1(\eta)$  on the hyperplanes be denoted by  $\phi(\eta)$ . Consider the function

$$\begin{aligned} q(\eta) &= \phi(0, 0, \eta_3, \dots, \eta_n) + (\eta_1 \cos \beta + \eta_2 \sin \beta)\phi_\beta(0, 0, \eta_3, \dots, \eta_n) \\ &\quad + \frac{1}{\sin \omega}(-\eta_1 \sin \beta + \eta_2 \cos \beta)[\phi_{\omega+\beta}(0, 0, \eta_3, \dots, \eta_n) \\ &\quad - \phi_\beta(0, 0, \eta_3, \dots, \eta_n)\cos \omega], \end{aligned}$$

where  $\phi_\beta$  and  $\phi_{\omega+\beta}$  are the first derivatives of  $\phi$  in the two directions  $\theta = \beta$  and  $\theta = \omega + \beta$  normal to  $S_0$ . The function  $u_2(\eta) = u_1(\eta) - q(\eta)$  satisfies in  $G_d$  an elliptic equation of type (3) and coincides on  $\Gamma_d$  with a function  $\psi_0(\eta)$ , satisfying all the conditions of Theorem 2. Thus  $u_2(\eta) \in C_\nu(\overline{G}_{r_0})$ ,  $r_0 < d$  and consequently  $u_1(\eta) \in$

$C_\nu(\bar{G}_{r_0})$ . Returning to the original  $x$ -coordinate we find that in some  $G_{P,\rho_0}$ ,  $\rho_0 < \rho_1$ ;  $u(x) \in C_\nu(\bar{G}_{P,\rho_0})$ . This proves the theorem.

*Remark 3.* From Remarks 1 and 2, we can also prove the following

**THEOREM 1'.** *There exist numbers  $\tau$  and  $\tau_0$ ,  $0 < \tau, \tau_0 < 1$  such that  $\rho^\tau(\partial^2 u / \partial x_i \partial x_j) \in C_{\tau_0}(\bar{\Omega})$ , where  $\rho(x)$  is a differentiable function coinciding near  $S$  with the distance from  $x$  to  $S$ .*

#### REFERENCES

- [1] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary condition, I*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.
- [2] T. CARLEMAN, *Über das Neumann–Poincarésche Problem für ein Gebiet mit Ecken*, Dissertation, Uppsala, 1916.
- [3] G. DZIUK, *Das Verhalten von Lösungen semilinearer elliptischer Systeme an Ecken eines Gebietes*, Math. Z., 159 (1978), pp. 89–100.
- [4] V. A. KONDRAT'EV, *Boundary value problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.
- [5] ———, *The smoothness of a solution of Dirichlet's problem for second order elliptic equations in a region with a peicewise smooth boundary*, Differentsial'nye Uravneneya, 10 (1970), pp. 1831–1843.
- [6] ———, *Singularities of a solution of Dirichlet's problem for a second order elliptic equation in the neighborhood of an edge*, Differentsial'nye Uravneneya, 13 (1977), pp. 2026–2033.
- [7] M. H. PROTTER AND H. F. WEINBERGER, *Maximum principles in differential equations*, Prentice-Hall, Englewood Cliffs, 1967.
- [8] N. M. WIGLEY, *Mixed boundary value problems in plane domains with corners*, Math. Z., 115 (1970), pp. 33–52.

## FREE BOUNDARY PROBLEMS IN SOLIDIFICATION OF ALLOYS\*

VASILIOS ALEXIADES† AND JOHN R. CANNON‡

**Abstract.** Multidimensional three-phase free boundary problems for semilinear diffusion equations are studied as models for the solidification (or melting) of alloys. The intermediate phase represents the “mushy zone” in which the freezing of the remaining liquid provides a heat generation effect. The conditions on the two interfaces are of Stefan-type on one of them and of fast-chemical-reaction-type on the other. The existence, uniqueness and regularity of appropriately defined weak solutions are established when either the temperature or the heat flux is prescribed on the fixed boundary.

**Introduction.** Multidimensional three-phase free boundary problems for diffusion equations of the form

$$\alpha(u) \frac{\partial u}{\partial t} = \operatorname{div} [k(u) \operatorname{grad} u]$$

are studied as models for the solidification (or melting) of alloys. These problems (described precisely in § 1) differ from the usual Stefan problems in that a Stefan-type condition [10], [4], [6] is imposed on one of the free boundaries and a fast-chemical-reaction-type condition [2], [3] is imposed on the other.

Solidification (or melting) of pure substances can be modelled by (two-phase) Stefan problems, but most actual solidification processes involve alloys rather than pure metals. Contrary to the isothermal freezing of a pure metal, the liquid alloy freezes partially and gradually until its temperature drops to a eutectic temperature and then the remaining liquid freezes isothermally at that temperature [11], [5]. Thus, the liquid and the solid are separated by a “mushy zone” between two isothermal surfaces at the liquidus and solidus temperatures respectively. At the solidus temperature, latent heat is released due to the freezing of the remaining liquid and thus imposes a Stefan-type condition across the solidus interface. On the other hand, there is no latent heat being generated at the liquidus temperature which dictates a fast-chemical-reaction-type condition on the liquidus interface. On the fixed boundary either the temperature or the heat flux can be prescribed and so we consider two separate problems: Problem I with Dirichlet boundary conditions and Problem II with Neumann conditions on the fixed boundary.

The precise mathematical problems are described in § 1. As is well known, there are considerable difficulties in multidimensional free boundary problems, and naturally we look for weak solutions which are defined in § 2. Problem I is studied in § 3 by means of monotonicity methods developed by Brezis [1] (see also Lions [9]). For problem II we employ the compactness methods of Kamin (Kamenomostkaja) which appear in [7], [8], [6], [2], [3] to obtain existence and uniqueness in § 4. These methods could also be used to treat Problem I. The regularity results of Ladyzenskaja–Solonnikov–Ural’ceva [8] are used to show that the weak solution of either problem is actually Holder continuous in certain subdomains.

The formulation and methods used here generalize naturally to multiphase problems with any combination of the two kinds of interface conditions considered here.

**1. Classical formulation of the problem.** Let  $G$  be a bounded domain in  $\mathbb{R}^n$  with smooth boundary  $\partial G$ . Set  $G(t) := G \times \{t\}$  and  $\partial G(t) := \partial G \times \{t\}$ . For any  $T$ ,  $0 < T \leq \infty$ , let

\* Received by the editors December 27, 1978.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916.

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

$\Omega \equiv \Omega^T := \cup_{0 < t < T}$  and  $S \equiv S^T := \cup_{0 < t < T} \partial G(t)$ .  $\Omega$  is divided into three parts,  $\Omega_s, \Omega_m, \Omega_l$  (corresponding to the solid, mushy and liquid zones) by the free (unknown) boundaries  $\Gamma_s := \cup_{0 \leq t \leq T} \Gamma_s(t)$  and  $\Gamma_l := \cup_{0 \leq t \leq T} \Gamma_l(t)$ , the solidus and liquidus fronts respectively. Here  $\Gamma_s(t) := \{(x, t) \in \overline{G}(t) : \Phi(x, t) = \Phi_s\}$ ,  $\Gamma_l(t) := \{(x, t) \in \overline{G}(t) : \Phi(x, t) = \Phi_l\}$  are hypersurfaces in  $\overline{G}(t)$  described by a function  $\Phi \in C^1(\overline{\Omega})$  with  $\nabla_x \Phi(x, t)|_{\Gamma_s, \Gamma_l} \neq 0$ , and  $\Phi_s, \Phi_l$  are constants. We assume that  $\Phi < \Phi_s$  in  $\Omega_s$ ,  $\Phi_s < \Phi < \Phi_l$  in  $\Omega_m$  and  $\Phi_l < \Phi$  in  $\Omega_l$ . The hypersurfaces  $\Gamma_s(t)$  and  $\Gamma_l(t)$  divide  $G(t)$  into three parts,  $G_s(t), G_m(t)$  and  $G_l(t)$ , so that  $\Omega_i = \cup_{0 < t < T} G_i(t)$ ,  $i = s, m, l$ . In particular,  $\Gamma_s(0)$  and  $\Gamma_l(0)$  subdivide  $G(0)$  into  $G_s(0), G_m(0), G_l(0)$  some of which may be empty. Finally, let  $S_i := \overline{\Omega}_i \cap S$ ,  $i = s, m, l$ , and note that some of them may be empty.

In the mushy region  $\Omega_m$  the relative amount of the solid present at any temperature is given by a (known) function  $f(u)$  called the solid fraction. At the liquidus temperature  $\theta_l$  there is no solid present and so  $f(\theta_l) = 0$ , whereas for  $\theta_l > u > \theta_s$ , it is  $0 \leq f(u) < 1$ , where  $\theta_s$  is the solidus temperature. Thus  $f(u)$  is decreasing and its time rate of change in the mushy region provides a heat generation effect which adds the term  $\alpha(\partial f/\partial t)$  to the heat condition equation in  $\Omega_m$ . At the solidus front  $\Gamma_s$ , where the temperature is  $\theta_s$ , the freezing of the remaining liquid  $1 - f(\theta_s)$  provides a latent heat effect and creates an interface condition of Stefan-type. On the fixed boundary  $S$ , either the temperature is prescribed and this we will call Problem I, or the heat flux is prescribed and this will be referred to as Problem II. These physical considerations lead to the following problem.

Find  $u^s, u^m, u^l$  and  $\Phi$  satisfying the equations

$$(1.1)_s \quad \alpha_s(u^s) \frac{\partial u^s}{\partial t} = \text{div} [k_s(u^s) \nabla_x u^s] \quad \text{in } \Omega_s,$$

$$(1.1)_m \quad \alpha_m(u^m) \frac{\partial u^m}{\partial t} = \text{div} [k_m(u^m) \nabla_x u^m] + \alpha \frac{\partial f(u^m)}{\partial t} \quad \text{in } \Omega_m,$$

$$(1.1)_l \quad \alpha_l(u^l) \frac{\partial u^l}{\partial t} = \text{div} [k_l(u^l) \nabla_x u^l] \quad \text{in } \Omega_l;$$

the interface conditions (of continuity and heat balance)

$$(1.2)_s \quad u^s = u^m = \theta_s \quad \text{on } \Gamma_s,$$

$$(1.2)_l \quad u^m = u^l = \theta_l \quad \text{on } \Gamma_l,$$

$$(1.3)_s \quad [k_m(\theta_s) \nabla_x u^m - k_s(\theta_s) \nabla_x u^s] \cdot \nabla_x \Phi = \alpha [1 - f(\theta_s)] \frac{\partial \Phi}{\partial t} \quad \text{on } \Gamma_s,$$

$$(1.3)_l \quad [k_m(\theta_l) \nabla_x u^m - k_l(\theta_l) \nabla_x u^l] \cdot \nabla_x \Phi = 0 \quad \text{on } \Gamma_l;$$

the initial conditions

$$(1.4)_s \quad u^s(x, 0) = h_s(x) \quad \text{in } G_s(0),$$

$$(1.4)_m \quad u^m(x, 0) = h_m(x) \quad \text{in } G_m(0),$$

$$(1.4)_l \quad u^l(x, 0) = h_l(x) \quad \text{in } G_l(0),$$

where  $h_s < \theta_s < h_m < \theta_l < h_l$ , and either boundary condition of Dirichlet type

$$(1.5) \quad u^i = g_i(x, t) \quad \text{on } S, \quad i = s, m, l \quad (\text{Problem I}),$$

where  $g_s < \theta_s < g_m < \theta_l < g_l$ , or the boundary conditions of Neumann type

$$(1.6) \quad k_i(u^i) \frac{\partial u^i}{\partial n} = g_i(x, t) \quad \text{on } S, \quad i = s, m, l \quad (\text{Problem II}).$$

Here  $\partial/\partial n$  denotes differentiation in the direction of the outer normal to  $S$ ,  $\alpha$  is a positive constant (=density  $\times$  latent heat per unit mass) and

$$(1.7) \quad \alpha_i(u), k_i(u), i = s, m, l, \text{ are continuous functions satisfying } 0 < \gamma_0 \leq \alpha_i(u) \leq \gamma_1, \\ 0 < \gamma_0 \leq k_i(u) \leq \gamma_1, i = s, m, l, \text{ and } f(u) \text{ is a differentiable decreasing function} \\ \text{satisfying } 0 \leq f(u) \leq 1, 0 \leq -f'(u) \leq \gamma_2,$$

for some positive constants  $\gamma_0, \gamma_1, \gamma_2$ . We also assume that

$$(1.8) \quad G \text{ is bounded and } \partial G \in C^{2+\lambda} \quad (\lambda > 0).$$

The function  $\Phi$  is not uniquely determined; it is only one possible parametrization of the surfaces  $\Gamma_s, \Gamma_l$ . The interface conditions of continuity (1.2) help in describing  $\Gamma_s, \Gamma_l$  as level surfaces of the solution  $\{u^s, u^m, u^l\}$ .

By a classical solution of Problem I [of Problem II] we mean a solution  $\{u^s, u^m, u^l, \Phi\}$  of (1.1)–(1.5) [of (1.1)–(1.4), (1.6)] such that  $u^i \in C(\bar{\Omega}_i), \nabla_x u^i \in C(\bar{\Omega}_i \setminus S)$  [ $u^i, \nabla_x u^i \in C(\bar{\Omega}_i)$ ] and  $D_x^2 u^i, D_t u^i \in C(\Omega_i), i = s, m, l$ .

**2. Generalized formulations of Problems I and II.** We introduce the quantities

$$A_s(u) := \int_u^{\theta_s} \alpha_s(\xi) d\xi + \int_{\theta_s}^{\theta_l} \alpha_m(\xi) d\xi, \\ A_m(u) := \int_u^{\theta_l} \alpha_m(\xi) d\xi, \quad A_l(u) := \int_u^{\theta_l} \alpha_l(\xi) d\xi; \\ K_s(u) := \int_u^{\theta_s} k_s(\xi) d\xi + \int_{\theta_s}^{\theta_l} k_m(\xi) d\xi, \\ K_m(u) := \int_u^{\theta_l} k_m(\xi) d\xi, \quad K_l(u) := \int_u^{\theta_l} k_l(\xi) d\xi; \\ u = \begin{cases} u^s, & \text{in } \Omega_s, \\ u^m, & \text{in } \Omega_m, \\ u^l & \text{in } \Omega_l, \end{cases} \quad h = \begin{cases} h_s, & \text{in } G_s(0), \\ h_m, & \text{in } G_m(0), \\ h_l, & \text{in } G_l(0), \end{cases} \quad g = \begin{cases} g_s, & \text{in } S_s, \\ g_m, & \text{in } S_m, \\ g_l, & \text{in } S_l, \end{cases}$$

and write the problems in the form

$$(2.1)_s \quad \frac{\partial}{\partial t} A_s(u) = \text{div} [\nabla_x K_s(u)] \quad \text{in } \Omega_s,$$

$$(2.1)_m \quad \frac{\partial}{\partial t} [A_m(u) + \alpha f(u)] = \text{div} [\nabla_x K_m(u)] \quad \text{in } \Omega_m,$$

$$(2.1)_l \quad \frac{\partial}{\partial t} A_l(u) = \text{div} [\nabla_x K_l(u)] \quad \text{in } \Omega_l;$$

$$(2.2) \quad u = \theta_s \quad \text{on } \Gamma_s, \quad u = \theta_l \quad \text{on } \Gamma_l,$$

$$(2.3)_s \quad [\nabla_x K_s(u) - \nabla_x K_m(u)] \cdot \nabla_x \Phi = \alpha [1 - f(\theta_s)] \frac{\partial \Phi}{\partial t} \quad \text{on } \Gamma_s,$$

$$(2.3)_l \quad [\nabla_x K_l(u) - \nabla_x K_m(u)] \cdot \nabla_x \Phi = 0 \quad \text{on } \Gamma_l;$$

$$(2.4) \quad u = h \quad \text{in } G(0);$$

$$(2.5) \quad u = g \quad \text{on } S \quad (\text{for Problem I});$$



$$(2.6) \quad -\frac{\partial}{\partial n}K_i(u) = g_i(x, t) \quad \text{on } S_i, \quad i = s, m, l \quad (\text{for Problem II}).$$

Let  $\varphi$  be a smooth test function in  $\mathbb{R}^{n+1}$  such that  $\varphi(x, T) = 0$  and  $\varphi|_S = 0$  for Problem I, whereas  $\partial\varphi/\partial n|_S = 0$  for Problem II. Write (2.1)<sub>s</sub> equivalently as  $(\partial/\partial t)[A_s(u) + \alpha] = \text{div} [\nabla_x K_s(u)]$ , multiply by  $\varphi$  and integrate over  $\Omega_s$ ; thanks to (2.2), (2.4) and (2.5) [resp. (2.6)] we obtain

$$(2.7) \quad \begin{aligned} & \iint_{\Omega_s} \varphi_t [A_s(u) + \alpha] dx dt + \int_{G_s(0)} \varphi(x, 0) [A_s(h) + \alpha] dx + \int_{\Gamma_s} \varphi [A_s(\theta_s) + \alpha] \Phi_t |\nabla\Phi|^{-1} d\Gamma_s \\ & = \iint_{\Omega_s} \nabla_x K_s(u) \cdot \nabla_x \varphi dx dt + \int_{\Gamma_s} \varphi \frac{\partial K_s(u)}{\partial n_x} d\Gamma_s + 0 \quad \left[ \text{resp.} + \int_{S_s} \varphi g_s dS \right]. \end{aligned}$$

Next, multiply (2.1)<sub>m</sub> and (2.1)<sub>l</sub> by  $\varphi$  and integrate to obtain similarly (note that  $A_m(\theta_l) = A_l(\theta_l) = 0, f(\theta_l) = 0$ )

$$(2.7)_m \quad \begin{aligned} & \iint_{\Omega_m} \varphi_t [A_m(u) + \alpha f(u)] dx dt + \int_{G_m(0)} \varphi(x, 0) [A_m(h) + \alpha f(h)] dx \\ & \quad - \int_{\Gamma_s} \varphi [A_m(\theta_s) + \alpha f(\theta_s)] \Phi_t |\nabla\Phi|^{-1} d\Gamma_s \\ & = \iint_{\Omega_m} \nabla_x K_m(u) \cdot \nabla_x \varphi dx dt - \int_{\Gamma_s} \varphi \frac{\partial K_m(u)}{\partial n_x} d\Gamma_s + \int_{\Gamma_l} \varphi \frac{\partial K_m(u)}{\partial n_x} d\Gamma_l + 0 \\ & \quad \left[ \text{resp.} + \int_{S_m} \varphi g_m dS \right], \end{aligned}$$

and

$$\begin{aligned} & \iint_{\Omega_l} \varphi_t A_l(u) dx dt + \int_{G_l(0)} \varphi(x, 0) A_l(h) dx \\ & = \iint_{\Omega_l} \nabla_x K_l(u) \cdot \nabla_x \varphi dx dt - \int_{\Gamma_l} \varphi \frac{\partial K_l(u)}{\partial n_x} d\Gamma_l + 0 \quad \left[ \text{resp.} + \int_{S_l} \varphi g_l dS \right]. \end{aligned}$$

Now we introduce the quantities

$$(2.8) \quad a(u) := \begin{cases} -A_s(u) - \alpha, \\ -A_m(u) - \alpha f(u), \\ -A_l(u) \end{cases} \quad \text{and} \quad k(u) := \begin{cases} -K_s(u) & \text{for } u < \theta_s, \\ -K_m(u) & \text{for } \theta_s < u < \theta_l, \\ -K_l(u) & \text{for } \theta_l \leq u, \end{cases}$$

let us note that they are increasing and continuous except that  $a(u)$  has a jump at  $\theta_s$ , so it could be considered as being multivalued there. We add relations (2.7) together and observing that  $A_s(\theta_s) = A_m(\theta_s)$ , that  $(\partial/\partial n_x)K_s(u) - (\partial/\partial n_x)K_m(u) = [\nabla_x K_s(u) - \nabla_x K_m(u)] \cdot (\nabla_x \Phi / |\nabla\Phi|) = \alpha [1 - f(\theta_s)] (\Phi_t / |\nabla\Phi|)$  on  $\Gamma_s$  by (2.3)<sub>s</sub>, and that  $(\partial/\partial n_x)K_m(u) - (\partial/\partial n_x)K_l(u) = 0$  on  $\Gamma_l$  by (2.3)<sub>l</sub>, we find

$$(2.9) \quad \begin{aligned} & \iint_{\Omega} \varphi_t a(u) dx dt + \int_{G(0)} \varphi(x, 0) a(h) dx = \iint_{\Omega} \nabla_x k(u) \cdot \nabla_x \varphi dx dt + 0 \\ & \quad \left[ \text{resp.} - \int_S \varphi g ds \right], \end{aligned}$$

(the integral  $\int_{\Gamma_s} \varphi \{ \alpha [1 - f(\theta_s)] \} \Phi_t |\nabla \Phi|^{-1} d\Gamma_s$  cancels out since it appears on both sides).

For a classical solution  $u$ , the jump of  $a(u)$  at  $u = \theta_s$  does not affect the integral since  $u = \theta_s$  on an  $(n - 1)$ -dimensional surface  $\Gamma_s$  which has measure zero. However, if we do not want to postulate a priori that even a weak solution must take the value  $\theta_s$  on a set of measure zero, then the meaning of the first integral in (2.9) is ambiguous. We handle this difficulty in two different ways for the two problems.

Consider  $a(\cdot)$  as a multivalued mapping by defining  $a(\theta_s) := [a(\theta_s - 0), a(\theta_s + 0)] \equiv [-\int_{\theta_s}^{\theta_s} \alpha_m(\xi) d\xi - \alpha, -\int_{\theta_s}^{\theta_s} \alpha_m(\xi) d\xi - \alpha f(\theta_s)]$ . Then,  $a(\cdot)$  being continuous and increasing has a continuous, increasing and single-valued inverse  $a^{-1}(\cdot)$ . Set  $A(\cdot) := k(a^{-1}(\cdot))$ , which is also continuous (in fact Lipschitz), increasing and single-valued, and introduce the new unknown

$$(2.10) \quad v \in a(u), \quad \text{i.e., } u = a^{-1}(v)$$

so that  $k(u) = A(v)$ . Expression (2.10) means  $v(x, t) = a(u(x, t))$  if  $u(x, t) \neq \theta_s$  and  $v(x, t) \in a(\theta_s) = [a(\theta_s - 0), a(\theta_s + 0)]$  if  $u(x, t) = \theta_s$ , and thus  $v(x, t)$  is a function (single-valued). This leads to the following

**DEFINITION (Weak solution for Problem I).** A function  $u(x, t)$  is a weak solution of Problem I (i.e. of (1.1)–(1.5)) if  $u \in L^2(0, T; H^1(G))$ ,  $u = g$  on  $S$  and  $u = a^{-1}(v)$  for some function  $v \in L^2(0, T; H^1(G))$  satisfying

$$(2.11) \quad \int_0^T (v, \varphi_t) dt - \int_0^T (\nabla_x A(v), \nabla_x \varphi) dt + (a(h), \varphi(x, 0)) = 0$$

for any smooth  $\varphi(x, t)$  such that  $\varphi(x, T) \equiv 0$ ,  $\varphi|_S = 0$ . Here (and below)  $(\cdot, \cdot)$  is the  $L^2(G)$  inner product and  $H^1(G)$  denotes the usual Sobolev space. The weak free boundaries  $\Gamma_s$  and  $\Gamma_l$  are the sets where  $\{u = \theta_s\} \equiv \{a(\theta_s - 0) \leq v \leq a(\theta_s + 0)\}$  and  $\{u = \theta_l\} \equiv \{v = a(\theta_l) = 0\}$ .

For Problem II we proceed differently. Since  $k(\cdot)$  is continuous and strictly increasing, so is its inverse  $k^{-1}(\cdot)$ . Consider the multivalued mapping

$$(2.12) \quad b(\rho) = \begin{cases} a(k^{-1}(\rho)), & \text{if } \rho \neq k(\theta_s) =: \kappa, \\ [a(\theta_s - 0), a(\theta_s + 0)], & \text{if } \rho = k(\theta_s) =: \kappa, \end{cases}$$

and introduce the new unknown

$$(2.13) \quad v = k(u), \quad \text{i.e., } u = k^{-1}(v).$$

Let  $B(v)$  denote any function such that  $B(v) \subset b(v)$  in the sense of graphs, in other words,  $B(v(x, t)) = b(v(x, t)) = a(k^{-1}(v(x, t)))$  if  $v(x, t) \neq \kappa$  and  $B(v(x, t)) \in b(\kappa) \equiv [a(\theta_s - 0), a(\theta_s + 0)]$  if  $v(x, t) = \kappa$  ( $\kappa = k(\theta_s)$ ).

**DEFINITION (Weak solution for Problem II).** By a weak solution of Problem II (i.e. of (1.1)–(1.4) and (1.6)) we mean a bounded measurable function  $u(x, t)$  such that the bounded measurable function  $v = k(u)$  satisfies

$$(2.14) \quad \iint_{\Omega} \{B(v)\varphi_t + v\Delta\varphi\} dx dt + \int_{G(0)} B(k(h))\varphi(x, 0) dx + \int_S g\varphi dS = 0$$

for some function  $B(v)$  as above, and for any smooth  $\varphi(x, t)$  such that  $\varphi(x, T) \equiv 0$ ,  $\partial\varphi/\partial n|_S = 0$ . The weak versions of  $\Gamma_s$  and  $\Gamma_l$  are the sets where  $\{u = \theta_s\} \equiv \{v = k(\theta_s)\}$  and  $\{u = \theta_l\} \equiv \{u = k(\theta_l) = 0\}$ . Note that since  $h \neq \theta_s$  a.e.,  $B(k(h)) = a(h)$  for any  $B$  as above.

We summarize the discussion up to now in the following

**THEOREM 1.** A classical solution of Problem I [Problem II] is also a weak solution. Conversely one can easily show (similarly to [6, p. 54], [2, p. 435]).

**THEOREM 2.** *A sufficiently regular weak solution is a classical solution.*

In closing this section we collect some easily obtainable (from (1.7)) properties of  $a(\cdot)$ ,  $k(\cdot)$ ,  $A(\cdot)$  and  $b(\cdot)$ , which will be needed later.

**LEMMA.** (i)  $a(r)$  is strictly increasing, and continuous except at  $r = \theta_s$  where it has a jump of magnitude  $J := \alpha[1 - f(\theta_s)]$ ; for any  $r > \bar{r}$  in  $\mathbb{R}$ ,  $\gamma_0(r - \bar{r}) \leq a(r) - a(\bar{r}) \leq \gamma_3(r - \bar{r}) + J$  with  $\gamma_3 := \gamma_1 + \alpha\gamma_2$ ; (ii)  $k(r)$  is strictly increasing and continuous; for any  $r > \bar{r}$  in  $\mathbb{R}$ ,  $\gamma_0(r - \bar{r}) \leq k(r) - k(\bar{r}) \leq \gamma_3(r - \bar{r})$ ; (iii) the derivatives  $a'(r)$ ,  $k'(r)$  are (in general) discontinuous at  $r = \theta_s$ ,  $r = \theta_l$  and  $0 < \gamma_0 \leq a'(r) \leq \gamma_3$ ,  $0 < \gamma_0 \leq k'(r) \leq \gamma_3$ ; (iv)  $A(\rho) := k(a^{-1}(\rho))$  is increasing, Lipschitz continuous:  $|A(\rho) - A(\bar{\rho})| \leq \gamma_3/\gamma_0|\rho - \bar{\rho}|$ , and there exist constants  $\lambda_1^\pm$  and  $\lambda_2$ ,  $\lambda_1^\pm > 0$ , such that  $\lambda_1^-\rho + \lambda_2 \leq A(\rho)$ ,  $\rho \leq 0$ ,  $\lambda_1^+\rho + \lambda_2 \leq A(\rho)$ ,  $\rho > 0$ ; (v)  $b(\rho)$  of (2.12) is strictly increasing, continuous, multivalued at  $\rho = \kappa := k(\theta_s)$  and for any  $\rho > \bar{\rho}$

$$(2.15) \quad \frac{\gamma_0}{\gamma_3}(\rho - \bar{\rho}) \leq b(\rho) - b(\bar{\rho}) \leq \frac{\gamma_3}{\gamma_0}(\rho - \bar{\rho}) + J.$$

**3. Existence, uniqueness and regularity for Problem I.** We shall cast the problem in a form for which the abstract existence result of Brezis [1, p. 31] is applicable. The following will be required of the data:

$$(3.1) \quad h \in L^2(G), \quad h \neq \theta_s \quad \text{a.e.},$$

$$(3.2) \quad g \in H^{1/2}(S), \quad g \neq \theta_s \quad \text{a.e.},$$

where  $H^\sigma(S)$  stands for the space  $W_2^\sigma(S)$  of [8, p. 70]. Then  $g$  admits an extension to a function

$$(3.3) \quad \tilde{g} \in H^1(\Omega) \quad \text{such that} \quad g = \tilde{g} \quad \text{on } S \quad \text{and} \quad \tilde{g} \neq \theta_s \quad \text{a.e.}$$

We set

$$(3.4) \quad z_1 := a(\tilde{g}) \in H^1(\Omega), \quad z_2 := k(\tilde{g}) \in H^1(\Omega),$$

(note that  $a(\tilde{g}(x, t))$  is single-valued a.e.), and define

$$(3.5) \quad \tilde{A}(\rho) := A(\rho + z_1) - z_2$$

which has the same properties as  $A(\rho)$  (see Lemma, § 2).

We seek  $v \in L^2(0, T; H^1(G))$  satisfying  $v = a(g) = z_1|_S$  on  $S$  and (2.11). Equivalently, this can be expressed briefly by saying that we seek the solution of

$$(3.6) \quad \begin{aligned} \frac{\partial v}{\partial t} - \Delta A(v) &= 0 && \text{in } \Omega, \\ v(0) &= a(h) && \text{in } G(0), \\ v|_S &= z_1|_S && \text{on } S, \end{aligned}$$

in the sense of the dual of the space  $\dot{W}^{1,1}(\Omega) := \{\varphi \in L^2(0, T; H_0^1(G)) : \varphi(T) = 0, \partial\varphi/\partial t \in L^2(0, T; L^2(G))\}$ , where  $\Delta : H_0^1(G) \rightarrow L^2(G)$  denotes the weak Laplacian. Letting

$$(3.7) \quad v = w + z_1$$

we see that  $w \in L^2(0, T; H_0^1(G))$  must satisfy

$$(3.8) \quad \begin{aligned} \frac{\partial w}{\partial t} - \Delta \tilde{A}(w) &= F \quad \text{in } \Omega, & F &:= -\frac{\partial}{\partial t} z_1 - \Delta z_2, \\ w(0) &= w_0 := a(h) - z_1(0), \\ \tilde{A}(w)|_S &= 0, \end{aligned}$$

in the sense of the dual of  $\dot{W}^{1,1}(\Omega)$ . Next, using the bounded, linear, monotone and self-adjoint operator  $E: L^2(G) \rightarrow H_0^1(G) \subset L^2(G)$  defined as  $E := (-\Delta)^{-1}$  with zero boundary conditions, one can easily put (3.8) in the form of Brezis [1, p. 31], namely

$$(3.9) \quad - \int_0^T \left( w, \frac{\partial}{\partial t} E\varphi \right) dt + \int_0^T (\tilde{A}(w), \varphi) dt = \int_0^T (EF, \varphi) dt + (E^{1/2} w_0, E^{1/2} \varphi(0)),$$

for all  $\varphi \in \dot{W}^{1,1}(\Omega)$ . The hypotheses of Theorem 2 [1, p. 31] are satisfied with the choices  $V = H = L^2(0, T; L^2(G))$  because our operator  $\tilde{A}: V \rightarrow V$  is monotone, hemicontinuous (in fact Lipschitz continuous), bounded and coercive as one easily checks using the lemma at the end of § 2. Therefore there exists  $w \in V = L^2(\Omega)$  solution of (3.9). This is a very weak solution, but it can be shown by the method of [1, p. 35] that  $\tilde{A}(w) \in L^2(0, T; H_0^1(G))$ , so that  $w \in L^2(0, T; H_0^1(G))$  is a solution of (3.8), and  $v := w + z_1 \in L^2(0, T; H^1(G))$  is a solution of (3.6). If  $v_1$  and  $v_2$  are two solutions then (3.6) immediately implies  $A(v_1) = A(v_2)$  and, since  $k$  is strictly increasing, also  $a^{-1}(v_1) = a^{-1}(v_2)$ . Then  $u := a^{-1}(v) \in L^2(0, T; H^1(G))$  is the unique weak solution of Problem I as defined in § 2.

The weak solution  $u$  possesses additional smoothness. Indeed, by an energy estimate one can show that  $w \in L^\infty(0, T; L^2(G))$  (in fact  $t \mapsto \|w(t)\|_{L^2(G)}$  is continuous), which means that  $w$  is an element of the space  $\dot{V}_2(\Omega)$  of [8, p. 6]. Then the method of [8, p. 156–9] can be employed to show that the function  $t \mapsto w(\cdot, t)$  is continuous in  $L^2(G)$ , so that  $w(x, t)$ , hence also  $v(x, t)$  and  $u(x, t)$ , belong to  $C(0, T; L^2(G))$  (to the space  $V_2^{1,0}(\Omega)$  in the terminology of [8]).

We summarize the above results in the following:

**THEOREM 3.** (Existence and uniqueness for Problem I). *Under the assumptions (1.7), (1.8) and (3.1), (3.2), Problem I has unique weak solution  $u \in L^2(0, T; H^1(G)) \cap C(0, T; L^2(G))$ .*

Now we appeal to the regularity results of [8] to prove

**THEOREM 4** (Regularity for Problem I). *Under the assumptions (1.7), (1.8) and*

$$(3.10) \quad h \in L^\infty(G), \quad h \neq \theta_s \quad \text{a.e.,}$$

$$(3.11) \quad g \in H^{1/2}(S) \cap L^\infty(S), \quad g \neq \theta_s \quad \text{a.e.,}$$

*the weak solution  $u$  of Problem I is bounded and Hölder continuous on every compact subdomain  $\Omega'$  of  $\Omega$  in which  $u < \theta_s$  or  $u > \theta_s$  a.e., i.e.,  $u \in L^\infty(\Omega') \cap \mathcal{H}^{\mu, \mu/2}(\Omega')$  for some  $0 < \mu < 1$ . Here, as before,  $H^{1/2}$  denotes the fractional order Sobolev space  $W_2^{1/2}$  of [8, p. 70], while  $\mathcal{H}^{\mu, \mu/2}$  stands for the space  $H^{\mu, \mu/2}$  of [8, p. 8].*

*Proof.* We have already seen that any solution  $u(x, t)$  of (3.6) belongs to the space  $V_2^{1,0}(\Omega)$  of [8, p. 6]. With the help of the lemma of § 2, one easily checks that the hypotheses of Theorems 2.1 and 1.1 [8, pp. 425 and 419] are satisfied for any domain  $\Omega' \subset \Omega$  not intersecting the interface  $\Gamma_s$ . Hence  $v \in \mathcal{H}^{\mu', \mu'/2}(\Omega')$  for some  $\mu' > 0$ . Then  $u = a^{-1}(v) \in \mathcal{H}^{\mu, \mu/2}(\Omega')$  with  $\mu = \min\{\mu', 1\}$ . Q.E.D.

**4. Existence, uniqueness and regularity for Problem II.** We begin by establishing uniqueness with the method of [7] which has also been used in [8], [6], [2], [3]. We outline the method here referring for more details to [2].

**THEOREM 5** (Uniqueness for Problem II). *Under the assumptions (1.7), (1.8), Problem II has at most one weak solution.*

*Proof.* Let  $u_1, u_2$  be two weak solutions. By definition, they are bounded measurable functions such that  $v_i = k(u_i), i = 1, 2$ , which satisfy (2.14) for some functions  $B_1(v_1)$  and  $B_2(v_2)$  respectively. Since  $h \neq \theta_s$  a.e. in  $G$ ,  $a(h)$  is single-valued and therefore  $B_1(k(h)) = B_2(k(h)) = a(h)$ . Subtracting (2.14) for  $v_2$  from (2.14) for  $v_1$  we find

$$(4.1) \quad \iint_{\Omega} \{ [B_1(v_1) - B_2(v_2)] \varphi_t + [v_1 - v_2] \Delta \varphi \} dx dt = 0.$$

The aim is to prove

$$(4.2) \quad \iint_{\Omega} [B_1(v_1) - B_2(v_2)] \psi dx dt = 0 \quad \forall \psi \in C_0^\infty(\Omega),$$

which implies  $B_1(v_1) = B_2(v_2)$ , hence  $v_1 = v_2$  a.e. in  $\Omega$ , and therefore also  $u_1 = u_2$  a.e. in  $\Omega$ .

To prove (4.2), consider the bounded nonnegative measurable function (see the lemma, § 2)

$$(4.3) \quad e(x, t) := \begin{cases} \frac{v_1(x, t) - v_2(x, t)}{B_1(v_1(x, t)) - B_2(v_2(x, t))}, & v_1(x, t) \neq v_2(x, t), \\ 0, & v_1(x, t) = v_2(x, t), \end{cases}$$

and set  $\beta(x, t) := B_1(v_1(x, t)) - B_2(v_2(x, t))$ , noting that  $|\beta(x, t)| \leq C_0 = \text{const.}$  by (2.15) and the boundedness of  $v_i, i = 1, 2$ . Now (4.1) can be written in the form

$$(4.4) \quad \iint_{\Omega} \beta(x, t) \{ \varphi_t + e(x, t) \Delta \varphi \} dx dt = 0.$$

By appropriate mollifications one can construct a sequence  $\bar{e}_m \in C^\infty(\bar{\Omega})$  such that  $0 \leq \bar{e}_m(x, t) \leq \sup_{\bar{\Omega}} e = \gamma_3/\gamma_0 =: C_1$  in  $\Omega$ , and  $\|e_m - e\|_{L^2(\Omega)} \leq 1/m, m = 1, 2, \dots$ . Now set  $e_m := \bar{e}_m + 1/m$  and for any  $\psi \in C_0^\infty(\Omega)$  fixed, let  $\varphi_m(x, t)$  be the solution of the (well-posed) problem

$$(4.5) \quad \begin{aligned} \frac{\partial}{\partial t} \varphi_m + e_m(x, t) \Delta \varphi_m &= \psi && \text{in } \Omega, \\ \varphi_m(x, T) &= 0 && \text{in } G(T), \\ \frac{\partial}{\partial n_x} \varphi_m &= 0 && \text{on } S. \end{aligned}$$

The functions  $\varphi_m$  are acceptable test functions in (4.4) which for  $\varphi = \varphi_m$  implies

$$(4.6) \quad \iint_{\Omega} \beta(x, t) \psi dx dt = \iint_{\Omega} \beta(x, t) [e_m - e] \Delta \varphi_m dx dt, \quad m = 1, 2, \dots$$

One can show (see [2]) that there exist constants  $C_2 = C_2(T; \max_{\bar{\Omega}} |\psi|), C_3, C_4$  independent of  $m$  such that  $\max_{\bar{\Omega}} |\varphi_m| \leq C_2, \|e_m^{1/2} \Delta \varphi_m\|_{L^2(\Omega)} \leq C_3$  and  $\|e/e_m\|_{L^2(\Omega)} \leq C_4$ . These, together with the boundedness of  $\beta(x, t)$ , allow one to prove that the right hand side of (4.6) tends to zero as  $m \rightarrow \infty$ , which establishes (4.2). Q.E.D.

Existence of the solution will be proved first for the case where the data satisfy

$$(4.7) \quad h \in H^1(G) \cap L^\infty(G), \quad h \neq \theta_s \quad \text{a.e.},$$

$$(4.8) \quad g \in L^\infty(S) \quad \text{and} \quad \frac{\partial g}{\partial t} \in L^1(S), \quad g \neq \theta_s \quad \text{a.e.}$$

After a stability (with respect to data) result has been obtained (Theorem 7), these assumptions will be weakened (see Theorem 8).

**THEOREM 6.** *Under the assumptions (1.7), (1.8) and (4.7), (4.8), Problem II admits a weak solution in  $H^1(\Omega)$ .*

*Proof.* We approximate  $b(\rho)$  of (2.12) uniformly on compact subsets of  $\mathbb{R}$  not containing  $\kappa = k(\theta_s)$ , by smooth functions  $b_m(\rho)$  which can be chosen to satisfy  $a(\theta_s - 0) \leq b_m(\kappa) \leq a(\theta_s + 0)$  and

$$(4.9) \quad 0 < \gamma_4 \leq b'_m(\rho) \leq \gamma_5(m) < \infty,$$

for some constants  $\gamma_4$  (independent of  $m$ ) and  $\gamma_5 = \gamma_5(m) \rightarrow \infty$  with  $m$ . Set  $v^0 := k(h) \in H^1(G) \cap L^\infty(G)$ , and approximate it by  $v_m^0 \in C^\infty(\bar{G})$  such that  $\max_{\bar{G}} |v_m^0| \leq \text{ess sup}_G |v^0| \leq \gamma_3 \cdot \text{ess sup}_G |h| =: C_5$  and  $v_m^0 \rightarrow v^0$  in  $H^1(G)$ . Finally we introduce a sequence  $\{g_m(x, t)\}$  of smooth functions defined on  $S$  and satisfying  $\sup_S |g_m| \leq \text{ess sup}_S |g| =: C_6$ ,  $g_m(x, 0) = (\partial/\partial n_x)v_m^0(x)|_{\partial G}$ ,  $g_m \rightarrow g$  in  $L^1(S)$ ,  $(\partial/\partial t)g_m \rightarrow (\partial/\partial t)g$  in  $L^1(S)$ .

Consider the approximating problems

$$(4.10) \quad \begin{aligned} \frac{\partial}{\partial t} b_m(v_m) &= \Delta v_m && \text{in } \Omega, \\ v_m(x, 0) &= v_m^0(x) && \text{in } G(0), \\ \frac{\partial}{\partial n_x} v_m &= g_m && \text{on } S. \end{aligned}$$

Existence and uniqueness of  $v_m$  is established by writing the equation in the form  $(\partial/\partial t)v_m - [b'_m(v_m)]^{-1} \Delta v_m = 0$ , reducing (4.10) to a problem with homogeneous initial condition using the transformation  $w_m = v_m - v_m^0$ , and then applying Theorem 7.4 [8, p. 491] for each  $m = 1, 2, \dots$ .

Now, one easily checks that the hypotheses of Theorem 2.3 [8, p. 16] are satisfied by choosing  $\mu_1 = 1/\gamma_4$ ,  $a_0 = 0$ ,  $\delta = 1$  and  $b_0 = 1$ . Thus the maximum principle [8, p. 17] yields

$$(4.11) \quad \max_{\bar{\Omega}} |v_m| \leq C_7, \quad m = 1, 2, \dots,$$

where the constant  $C_7$  depends on  $C_5, C_6$  but not on  $m$ .

Next, the fundamental  $m$ -independent bound

$$(4.12) \quad \|v_m\|_{H^1(\Omega)} \leq C_8, \quad m = 1, 2, \dots,$$

can be established as follows: Multiply the equation in (4.10) by  $(\partial/\partial t)v_m$  and integrate over  $G(t)$  to get  $\int_{G(t)} b'_m(v_m) |\partial v_m/\partial t|^2 dx + \frac{1}{2} \int_{G(t)} (\partial/\partial t) |\nabla_x v_m|^2 dx = \int_{\partial G(t)} g_m (\partial v_m/\partial t) d\sigma$ . Integrating this over  $[0, t]$  and using (4.9) and (4.11) yields  $\gamma_4 \int_0^t \int_{G(\tau)} |\partial v_m/\partial t|^2 dx d\tau + \frac{1}{2} \int_{G(t)} |\nabla_x v_m|^2 dx \leq \frac{1}{2} \int_{G(0)} |\nabla_x v_m^0|^2 dx + C_6 C_7 \text{meas}(\partial G) + C_7 \|(\partial/\partial t)g_m\|_{L^1(S)}$ . Since  $v_m^0$  and  $(\partial/\partial t)g_m$  converge in  $H^1(G)$  and  $L^1(S)$  respectively, the sequences  $\|\nabla_x v_m^0\|_{L^2(G)}$  and  $\|(\partial/\partial t)g_m\|_{L^1(S)}$  are bounded. It follows that  $\gamma_4 T \|\partial v_m/\partial t\|_{L^2(\Omega)}^2 + \frac{1}{2} \|\nabla_x v_m\|_{L^2(\Omega)}^2 \leq TC_9$  which, together with (4.11), proves (4.12).

The boundedness of  $\{v_m\}$  in  $H^1(\Omega)$  allows us to extract a subsequence (in what follows no distinction will be made in the notation between sequences and subsequences) which converges weakly in  $H^1(\Omega)$  and strongly in  $L^2(\Omega)$  (Rellich's lemma) to some  $v \in H^1(\Omega)$ . In particular,  $v_m \rightarrow v$  in measure, so a subsequence can be chosen to converge to  $v$  a.e. in  $\Omega$ . By (4.11),  $v$  can be taken to be bounded by  $C_7$  in  $\Omega$ . The functions  $v_m$  being classical solutions of (4.10) are also weak solutions, that is

$$(4.13) \quad \iint_{\Omega} \{b_m(v_m)\varphi_t + v_m \Delta\varphi\} dx dt + \int_{G(0)} b_m(v_m^0)\varphi(x, 0) dx + \int_S g_m\varphi dS = 0.$$

Since the sequence  $\{b_m(v_m(x, t))\}$  is a bounded sequence of measurable functions it has a subsequence which converges weakly in  $L^2(\Omega)$  to some  $\tilde{B}(x, t)$ . We claim that  $\tilde{B}(x, t)$  is a function of the type  $B(v(x, t))$ , i.e.,  $\tilde{B}(x, t) \in b(v(x, t))$  a.e. in  $\Omega$  (see § 2), so that by taking  $m \rightarrow \infty$ , (4.13) yields (2.14) which proves that  $u := k^{-1}(v) \in H^1(\Omega)$  is a weak solution. Indeed,  $v_m \rightarrow v$  a.e. in  $\Omega$  implies that for almost all points  $(x, t)$  for which  $v(x, t) \neq \kappa \equiv k(\theta_s)$ ,  $b_m(v_m(x, t)) \rightarrow b(v(x, t)) = a(k^{-1}(v(x, t)))$  because of the uniform convergence of  $b_m$  to  $b$ , hence  $\tilde{B}(x, t) = b(v(x, t))$  at such points. On the other hand, for a.a. points where  $v(x, t) = \kappa$ , we have  $a(\theta_s - 0) \leq \underline{\lim} b_m(v_m(x, t)) \leq \overline{\lim} b_m(v_m(x, t)) \leq a(\theta_s + 0)$ , hence  $\tilde{B}(x, t) \in [a(\theta_s - 0), a(\theta_s + 0)] = b(\kappa)$ , and this completes the proof. Q.E.D.

**THEOREM 7.** *If  $u_1$  and  $u_2$  are weak solutions of Problem II corresponding to data  $h_1, g_1$  and  $h_2, g_2$  which satisfy (4.7), (4.8), then*

$$\|u_1 - u_2\|_{L^2(\Omega)} \leq C_{10} \{ \|h_1 - h_2\|_{L^2(G)} + \|g_1 - g_2\|_{L^1(S)} \}.$$

*Proof.* Let  $v_i = k(u_i)$ ,  $i = 1, 2$ , let  $B_i(v_i)$ ,  $i = 1, 2$  be the functions used in the definition of weak solution (2.14), and set  $\beta(x, t) := B_1(v_1(x, t)) - B_2(v_2(x, t))$ . Next let  $\psi \in C_0^\infty(\Omega)$  and let  $\varphi_m$  be the solution of (4.5) which is an acceptable test function. Using (4.5), (4.3) and (2.14) we have  $\iint_{\Omega} \beta(x, t)\psi dx dt = \iint_{\Omega} \beta(x, t)\{(\partial/\partial t)\varphi_m + e \Delta\varphi_m\} dx dt + I_m = -\int_{G(0)} [B_1(k(h_1)) - B_2(k(h_2))]\varphi_m dx - \int_S [g_1 - g_2]\varphi_m dS + I_m$  where  $I_m := \iint_{\Omega} \beta(x, t)[e_m - e] \Delta\varphi_m dx dt$ . It can be shown [2, p. 490-1] that  $I_m \rightarrow 0$  as  $m \rightarrow \infty$  and that  $\max_{\bar{\Omega}} |\varphi_m| \leq C_2(T; \max |\psi|)$ . Hence

$$(4.14) \quad \left| \iint_{\Omega} \beta(x, t)\psi dx dt \right| \leq C_2 \cdot \{ \gamma_3 \|h_1 - h_2\|_{L^2(G)} + \|g_1 - g_2\|_{L^1(S)} \}$$

because  $h_i \neq \theta_s$  a.e.  $\Rightarrow |B_1(k(h_1)) - B_2(k(h_2))| = |a(h_1) - a(h_2)| \leq \gamma_3 |h_1 - h_2|$ . Now the bounded measurable function  $\beta(x, t)$  can be approximated in  $L^2(\Omega)$  by a sequence  $\psi_j \in C_0^\infty(\Omega)$  such that all the  $\psi_j$  are bounded in  $\Omega$  by the bound on  $\beta(x, t)$ . This bound is estimated by (2.15) and (4.11) in terms of only  $\gamma_0, \gamma_3, T, \text{ess sup}_G |h_i|, \text{ess sup}_S |g_i|, i = 1, 2$ . Thus (4.14) holds for each  $\psi_j$  ( $C_2$  will depend on the quantities just mentioned). Letting  $j \rightarrow \infty$  we obtain  $\iint_{\Omega} |\beta(x, t)|^2 dx dt \leq C_{11} \{ \|h_1 - h_2\|_{L^2(G)} + \|g_1 - g_2\|_{L^1(S)} \}$  whence the result follows because of  $|\beta(x, t)| \geq \gamma_0/\gamma_3 |v_1 - v_2| \geq \gamma_0^2/\gamma_3 |u_1 - u_2|$ . Q.E.D.

This result enables us to obtain the existence of a weak solution when  $h \in L^\infty(G)$ ,  $g \in L^\infty(S)$ . Indeed, given such data let  $h_j, g_j$  be approximations satisfying (4.7), (4.8). For each  $j = 1, 2, \dots$ , Theorem 6 yields the existence of a weak solution  $u_j$  such that  $v_j = k(u_j)$  satisfies (2.14) with data  $h_j, g_j$ . The sequences  $\{h_j\}, \{g_j\}$  are Cauchy, so by Theorem 7,  $\{u_j\}$  (and  $\{v_j\}$ ) is Cauchy in  $L^2(\Omega)$  and therefore it has a bounded limit  $u \in L^2(\Omega)$  such that  $v = k(u)$  satisfies (2.14) because  $v_j, B_j(k(h_j)) = a(h_j)$  and  $g_j$  converge to  $v, a(h)$  and  $g$  in  $L^2(\Omega), L^2(G)$  and  $L^1(S)$  respectively, and (as in the proof of Theorem 6)  $B_j(v_j)$  converges weakly in  $L^2(\Omega)$  to a function of type  $B(v)$ . Thus we have the more general

**THEOREM 8** (Existence and uniqueness for Problem II). *If (1.7), (1.8) hold and if  $h(x)$  and  $g(x, t)$  are bounded measurable functions in  $G$  and on  $S$  respectively, then Problem II has unique weak solution  $u \in L^2(\Omega)$ .*

*Remark.* Similarly one can show that the stability result of Theorem 7 still holds under the weaker hypotheses of Theorem 8.

We mention that the following monotone dependence result can be shown by the method of [6, p. 64], [2, p. 448]:

**THEOREM 9.** *Under the assumptions of Theorem 8, if  $h \cong \hat{h}$  a.e. in  $G$ ,  $g \cong \hat{g}$  a.e. on  $S$ , then the solutions  $u$  and  $\hat{u}$  corresponding to data  $h, g$  and  $\hat{h}, \hat{g}$  respectively satisfy  $u \cong \hat{u}$  a.e. in  $\Omega$ .*

Finally, the Hölder estimates of [8] can be employed as in [8, p. 501–2] to show:

**THEOREM 10** (Regularity for Problem II). *Under the assumptions of Theorem 8, the weak solution  $u$  of Problem II is Hölder continuous in every compact subdomain  $\Omega'$  of  $\Omega$  where  $u < \theta_s$  or  $u > \theta_s$ .*

*Remark.* The methods of § 4 can also be applied to Problem I and then a stability result like Theorem 7 as well as a monotone dependence result like Theorem 9 can also be obtained for that problem.

#### REFERENCES

- [1] H. BREZIS, *On some degenerate nonlinear parabolic equations*, Nonlinear Functional Analysis, F. E. Browder, ed., Proc. Symp. Pure Math. XVIII part 1, American Mathematical Society, Providence, RI, 1970, pp. 28–38.
- [2] J. R. CANNON AND C. D. HILL, *On the movement of a chemical reaction interface*, Indiana Univ. Math. J., 20 (1970), pp. 429–454.
- [3] J. R. CANNON AND A. FASANO, *Boundary value multidimensional problems in fast chemical reactions*, Arch. Rational Mech. Anal., 53 (1973), pp. 1–13.
- [4] J. R. CANNON, *Multiphase parabolic free boundary value problems*, Moving Boundary Problems, D. G. Wilson, A. D. Solomon, P. T. Boggs, eds., Proc. Symp. at Gatlinburg, Tennessee, Sept. 26–28, 1977, Academic Press, New York, 1978, pp. 3–20.
- [5] S. H. CHO AND J. E. SUNDERLAND, *Heat conduction problems with melting or freezing*, J. Heat Transfer, 91C(1969), pp. 421–426.
- [6] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 132 (1968), pp. 51–87.
- [6a] ———, *Erratum*, Ibid., 142 (1969), p. 557.
- [7] S. L. KAMENOMOSTKAJA, *On Stefan's problem*, Mat. Sb. (95), 53 (1961), pp. 489–514.
- [8] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Mono. Vol. 23, American Mathematical Society, Providence, RI, 1968.
- [9] J. L. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites non Lineaires*, Dunod, Paris, 1969.
- [10] L. I. RUBINSTEIN, *The Stefan Problem*, Transl. Math. Mono. Vol. 27, American Mathematical Society, Providence, RI 1971.
- [11] R. H. TIEN AND G. E. GEIGER, *A heat transfer analysis of the solidification of a binary eutectic system*, J. Heat Transfer, 89C(1967), pp. 230–234.



## BOUNDS FOR SOLUTIONS TO A CLASS OF INTEGRODIFFERENTIAL EQUATIONS ASSOCIATED WITH A THEORY OF RIGID NONCONDUCTING MATERIAL DIELECTRICS\*

FREDERICK BLOOM†

**Abstract.** Let  $H, H_+$  be real Hilbert spaces with  $H_+ \subseteq H$  algebraically and topologically and  $H_+$  dense in  $H$ . Let  $H_-$  be the dual of  $H_+$  via the inner product of  $H$  and denote by  $\mathcal{L}_S(H_+, H_-)$  the space of symmetric bounded linear operators from  $H_+$  into  $H_-$ . We prove that the evolution of the electric displacement field in a simple class of holohedral isotropic dielectrics can be modeled by an abstract initial-value problem of the form

$$\mathbf{u}_t - \alpha \mathbf{u}_t - \mathbf{L}\mathbf{u} + \int_0^t \mathbf{M}(t-\tau)\mathbf{u}(\tau) d\tau = \beta(t)\mathbf{u}_0, \quad 0 \leq t \leq T,$$

$$\mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_t(0) = \mathbf{u}_1 \quad (\mathbf{u}_0, \mathbf{u}_1 \in H_+),$$

where  $\mathbf{L} \in \mathcal{L}_S(H_+, H_-)$ ,  $\mathbf{M}(t) \in L^2([0, T]; \mathcal{L}_S(H_+, H_-))$ ,  $\beta(t) \in C^1([0, T])$ , and  $\alpha$  is an arbitrary (nonzero) real number. By employing a logarithmic convexity argument we derive growth estimates for solutions of the above system which lie in uniformly bounded classes of the form

$$\mathcal{N} = \{\mathbf{u} \in C^2([0, T]; H_+) \mid \sup_{[0, T]} \|\mathbf{u}\|_+ \leq N\}$$

for some  $N > 0$ ; our results are derived under a variety of assumptions concerning  $\alpha, \beta(t)$ , and the initial data (without making any definiteness assumptions on the operators  $\mathbf{L}$  or  $\mathbf{M}(t)$ ,  $0 \leq t < T$ ) and are used to obtain growth estimates for the electric displacement field  $\mathbf{D}(\mathbf{x}, t)$  in rigid dielectrics which satisfy constitutive relations of the form

$$\mathbf{D}(\mathbf{x}, t) = a_0 \mathbf{E}(\mathbf{x}, t) + \int_0^t \phi(t-\tau)\mathbf{E}(\mathbf{x}, \tau) d\tau,$$

$$\mathbf{H}(\mathbf{x}, t) = b_0 \mathbf{B}(\mathbf{x}, t) + \int_0^t \psi(t-\tau)\mathbf{B}(\mathbf{x}, \tau) d\tau,$$

where  $\mathbf{E}, \mathbf{H}, \mathbf{B}$  are the usual electromagnetic field variables,  $(\mathbf{x}, t) \in \Omega \times [0, T]$ ,  $\Omega \subseteq R^3$  is bounded region with smooth boundary  $\partial\Omega$ ,  $a_0$  and  $b_0$  are positive constants, and  $\phi, \psi$  are nonnegative monotonically decreasing functions of  $t$ .

**1. Introduction.** In recent work [1]–[4] this author has derived stability and growth estimates for specific classes of solutions to initial-value problems associated with abstract integrodifferential equations of the form

$$(1.1) \quad \mathbf{u}_t - \mathbf{N}\mathbf{u} + \int_{-\infty}^t \mathbf{K}(t-\tau)\mathbf{u}(\tau) d\tau = \mathbf{0}, \quad 0 \leq t < T.$$

In this equation  $\mathbf{u} \in C^2([0, T]; H_+)$  with  $\mathbf{u}_t \in C^1([0, T]; H_+)$ , and  $\mathbf{u}_t \in C([0, T]; H_-)$ , where  $H_+, H_-$  are Hilbert spaces which are defined as follows: Let  $H$  be any real Hilbert space with inner-product  $\langle \cdot, \cdot \rangle$  and let  $H_+ \subseteq H$  (algebraically and topologically) with  $H_+$  dense in  $H$ ; denote the inner-product on  $H_+$  by  $\langle \cdot, \cdot \rangle_+$ . Then  $H_-$  is the completion of  $H$  under the norm

$$(1.2) \quad \|\mathbf{w}\|_- = \sup_{\mathbf{v} \in H_+} \frac{|\langle \mathbf{v}, \mathbf{w} \rangle|}{\|\mathbf{v}\|_+}.$$

If we let  $\mathcal{L}(H_+, H_-)$  denote the space of bounded linear operators from  $H_+$  into  $H_-$  then

\* Received by the editors December 1, 1978 and in revised form May 9, 1979.

† Department of Mathematics and Computer Science, University of South Carolina, Columbia, South Carolina 29208. This research was supported in part by the Air Force Office of Scientific Research under AFOSR Grant 77-3396.

in (1.1) we only require that

- (i)  $\mathbf{N} \in \mathcal{L}(H_+, H_-)$  be symmetric; and
- (ii)  $\mathbf{K}(t), \mathbf{K}_t(t) \in L^2((-\infty, \infty); \mathcal{L}(H_+, H_-))$ ;

where  $\mathbf{K}_t$  denotes the strong operator derivative of  $\mathbf{K}$ ; no definiteness assumptions are placed on  $\mathbf{N}$  and thus the initial-value problem obtained by appending to (1.1) the initial data

$$(1.3a) \quad \mathbf{u}(0) = \mathbf{f}, \quad \mathbf{u}_t(0) = \mathbf{g}, \quad \mathbf{f}, \mathbf{g} \in H_+$$

and the prescription of the past history which is given by

$$(1.3b) \quad \mathbf{u}(\tau) = \mathbf{U}(\tau), \quad -\infty < \tau < 0,$$

is, in general, not well posed. If we restrict our attention to classes of bounded solutions to (1.1)–(1.3) of the form  $\mathcal{N} = \{\mathbf{v} \in C^2([0, T]; H_+) | \sup_{[0, T]} \|\mathbf{v}(t)\|_+ \leq N^2\}$  then it is possible to derive both stability and growth estimates for solutions  $\mathbf{u} \in \mathcal{N}$  under the assumption that  $\mathbf{K}(0)$  satisfies

$$(1.4a) \quad -\langle \mathbf{v}, \mathbf{K}(0)\mathbf{v} \rangle \geq \kappa \|\mathbf{v}\|_+^2, \quad \forall \mathbf{v} \in H_+,$$

where

$$(1.4b) \quad \kappa \geq \omega T \sup_{[0, T]} \|\mathbf{K}_t(t)\|_{\mathcal{L}(H_+, H_-)}$$

with  $\omega$  the embedding constant for the injection  $i: H \rightarrow H_+$ .

The technique used in [1]–[3] is based on a logarithmic convexity argument first employed by Knops and Payne [5] for the abstract wave equation obtained from (1.1) by setting  $\mathbf{K}(t) \equiv 0$ ; a different logarithmic convexity argument was employed by this author in [4] to derive continuous data dependence theorems for the system (1.1), (1.3a), (1.3b). The results obtained in [2]–[4] are applied in those papers to obtain growth, stability, and continuous data dependence theorems for solutions to initial-value problems associated with the equations of motion for linear isothermal viscoelastic materials; the spaces  $H$ ,  $H_+$ , and  $H_-$ , as well as the operators  $\mathbf{N}$  and  $\mathbf{K}(t)$ , are constructed and no definiteness assumptions are made on the initial value of the relaxation tensor. In the case of a one-dimensional homogeneous (isothermal) linear viscoelastic body, it is shown in [3] that the conditions (1.4a), (1.4b) are equivalent to the requirements that

$$(1.5) \quad \mathcal{g}'(0) \leq -\kappa \quad \text{with} \quad \kappa > \omega T \left( \sup_{[0, T]} |\ddot{\mathcal{g}}(t)| \right),$$

where  $\mathcal{g}(t)$  is the relaxation function of the material.

More recently we have turned our attention to the way in which integrodifferential equations arise in the theory of polarized nonconducting material dielectrics, i.e., in [6] we have considered the following problem: Let  $\mathbf{E}$ ,  $\mathbf{B}$ ,  $\mathbf{P}$ , and  $\mathbf{D}$  denote, respectively, the electric field vector, the magnetic flux density, the polarization vector, and the electric displacement in a nonconducting medium; the polarization and electric displacement vectors are related via

$$(1.6) \quad \mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \quad \varepsilon_0 \equiv \text{const.}$$

If  $(x^i, t)$ ,  $i = 1, 2, 3$ , denotes a Lorentz reference frame, with the  $(x^i)$  rectangular Cartesian coordinates and  $t$  the time parameter, then Maxwell's equations have the

local form

$$(1.7) \quad \frac{\partial \mathbf{B}}{\partial t} + \text{curl } \mathbf{E} = 0, \quad \text{div } \mathbf{B} = 0,$$

$$(1.8) \quad \text{curl } \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} = 0, \quad \text{div } \mathbf{D} = 0$$

whenever the density of free current  $\mathbf{J}_F = \mathbf{0}$ , the magnetization  $\mathbf{M} = \mathbf{0}$ , and the density of free charge  $Q_F = 0$ ; in (1.7b),  $\mathbf{H}$  represents the magnetic intensity and is related to the magnetic flux density via  $\mathbf{H} = \mu_0^{-1} \mathbf{B}$  where  $\varepsilon_0 \mu_0 = c^{-2}$ ,  $c$  being the speed of light in a vacuum. A determinate system of equations for the fields appearing in Maxwell's equations is obtained by specifying a set of constitutive relations. For example, in a vacuum  $\mathbf{P} = \mathbf{0}$  so

$$(1.9) \quad \mathbf{D} = \varepsilon_0 \mathbf{E}, \quad \mathbf{H} = \mu_0^{-1} \mathbf{B},$$

while in a rigid, linear, stationary nonconducting dielectric

$$(1.10) \quad \mathbf{D} = \boldsymbol{\varepsilon} \cdot \mathbf{E}, \quad \mathbf{B} = \boldsymbol{\mu} \cdot \mathbf{H},$$

where  $\boldsymbol{\varepsilon}$  and  $\boldsymbol{\mu}$  are constant second order tensors; the constitutive equations (1.10) were given by Maxwell in 1873 [7]. In [6] we considered the set of equations which define the dielectric as being a Maxwell–Hopkinson material, i.e., (1.10<sub>2</sub>) and

$$(1.11) \quad \mathbf{D}(t) = \varepsilon \mathbf{E}(t) + \int_{-\infty}^t \phi(t-\tau) \mathbf{E}(\tau) d\tau,$$

where  $\varepsilon > 0$  and  $\phi(t)$  is a continuous monotonically decreasing function for  $t \geq 0$ ; following a suggestion of Maxwell, Hopkinson [8] employed the constitutive equations (1.10<sub>2</sub>), (1.11) in connection with his studies on the residual charge of the Leyden jar. It was demonstrated in [6] that (1.11) in conjunction with the local Maxwell equations (1.7), (1.8), yield certain integrodifferential equations for the evolution of the electric field and the electric displacement field, respectively, in a nonconducting material dielectric of Maxwell–Hopkinson type.

By introducing suitable Hilbert spaces  $H, H_+, H_-$  and operators  $N \in \mathcal{L}(H_+, H_-)$  and  $K(t) \in L^2((-\infty, \infty); \mathcal{L}(H_+, H_-))$  we were able in [6] to treat the initial-boundary value problem for  $\mathbf{D}$ , as a special case of the abstract initial-value problem (1.1), (1.2) (in [6] we assumed that  $\mathbf{D}(\tau) = \mathbf{0}$ ,  $-\infty < \tau < 0$ ). From the stability and growth estimates derived for the electric displacement field  $\mathbf{D}$ , corresponding estimates were then derived from the electric field  $\mathbf{E}$ <sup>1</sup> by employing the relation

$$(1.12) \quad \mathbf{E}(t) = \varepsilon^{-1} \mathbf{D}(t) + \varepsilon^{-1} \int_0^t \Phi(t-\tau) \mathbf{D}(\tau) d\tau$$

which is obtained by inverting the Maxwell–Hopkinson relation (1.11) via the usual technique of successive approximation.

The constitutive relations associated with the Maxwell–Hopkinson theory, i.e., (1.10<sub>2</sub>) and (1.11), embody three basic simplifying assumptions: They are linear, they effect an a priori separation of electric and magnetic effects, and they do not allow for magnetic memory effects. As early as 1912 Volterra [9] proposed extending the

<sup>1</sup> For an excellent discussion of the qualitative behavior of electromagnetic fields and dielectric constants in dielectrics of Maxwell–Hopkinson type (especially in the presence of an applied time periodic electric field) we refer the reader to the monograph of H. Fröhlich, *Theory of Dielectrics*, Oxford University Press, 1949.

Maxwell–Hopkinson theory to treat the case where the dielectric is anisotropic, nonlinear, and magnetized; his constitutive relations were of the form

$$(1.13a) \quad \mathbf{D}(\mathbf{x}, t) = \boldsymbol{\varepsilon} \cdot \mathbf{E}(\mathbf{x}, t) + \mathcal{D}'_{-\infty}(\mathbf{E}(\mathbf{x}, \tau)),$$

$$(1.13b) \quad \mathbf{B}(\mathbf{x}, t) = \boldsymbol{\mu} \cdot \mathbf{H}(\mathbf{x}, t) + \mathcal{B}'_{-\infty}(\mathbf{H}(\mathbf{x}, \tau))$$

and it can be shown that (1.13a) reduces to (1.11) if the functional  $\mathcal{D}$  is linear and isotropic and the body satisfies various restrictions which follow from considerations of material symmetry. Of course, (1.13a), (1.13b) still effect an a priori separation of electric and magnetic effects and, as pointed out by Toupin and Rivlin [10], such a separation is inadequate with respect to predicting such a phenomena as the Faraday effect in dielectrics. In [10] Toupin and Rivlin postulated constitutive equations of the form

$$(1.14a) \quad \mathbf{D}(t) = \sum_{\nu=0}^n \mathbf{a}_{\nu} \cdot \mathbf{E}^{(\nu)}(t) + \sum_{\nu=0}^n \mathbf{c}_{\nu} \cdot \mathbf{B}^{(\nu)}(t) \\ + \int_{-\infty}^t \boldsymbol{\phi}_1(t, \tau) \cdot \mathbf{E}(\tau) d\tau + \int_{-\infty}^t \boldsymbol{\phi}_2(t, \tau) \cdot \mathbf{B}(\tau) d\tau,$$

$$(1.14b) \quad \mathbf{H}(t) = \sum_{\nu=0}^n \mathbf{d}_{\nu} \cdot \mathbf{E}^{(\nu)}(t) + \sum_{\nu=0}^n \mathbf{b}_{\nu} \cdot \mathbf{B}^{(\nu)}(t) \\ + \int_{-\infty}^t \boldsymbol{\psi}_1(t, \tau) \cdot \mathbf{B}(\tau) d\tau + \int_{-\infty}^t \boldsymbol{\psi}_2(t, \tau) \cdot \mathbf{E}(\tau) d\tau,$$

where  $\mathbf{E}^{(\nu)}(t) = d^{\nu} \mathbf{E}(t) / dt^{\nu}$  and  $\mathbf{a}_{\nu}, \dots, \mathbf{d}_{\nu}$  are constant tensors; the kernels  $\boldsymbol{\phi}_1, \dots, \boldsymbol{\psi}_2$  are taken to be continuous tensor functions of  $t$  and  $\tau$  which satisfy growth conditions of the form

$$\boldsymbol{\phi}_1(t, \tau) < c / (t - \tau)^{1+\rho}, \quad \rho > 0.$$

Toupin and Rivlin [10] also assumed that the dielectric does not exhibit aging and as a consequence it follows that  $\mathbf{D}(t)$  and  $\mathbf{H}(t)$  are periodic functions whenever  $\mathbf{E}(t)$  and  $\mathbf{B}(t)$  are; this latter result, when combined with the hypothesized growth estimates on the kernel functions, and early results of Volterra on the theory of functionals [9], yields the conclusion that  $\boldsymbol{\phi}_1, \dots, \boldsymbol{\psi}_2$  depend on  $t$  and  $\tau$  only through the difference  $t - \tau$  (the converse of this result is also true). Toupin and Rivlin [10] then prove that if the dielectric exhibits holohedral isotropy, i.e., if it admits as its group of material symmetry transformations the full orthogonal group, then  $\mathbf{E}(t)$  may be eliminated from (1.14b) and  $\mathbf{B}(t)$  may be eliminated from (1.14a); for a holohedral isotropic dielectric the constitutive equations (1.14a), (1.14b) reduce to

$$(1.15a) \quad \mathbf{D}(t) = \sum_{\nu=0}^n a_{\nu} \mathbf{E}^{(\nu)}(t) + \int_{-\infty}^t \phi(t - \tau) \mathbf{E}(\tau) d\tau,$$

$$(1.15b) \quad \mathbf{H}(t) = \sum_{\nu=0}^n b_{\nu} \mathbf{B}^{(\nu)}(t) + \int_{-\infty}^t \psi(t - \tau) \mathbf{B}(\tau) d\tau,$$

where  $\phi = \phi_1$ ,  $\psi = \psi_1$  and where (due to the assumption of holohedral isotropy)  $\mathbf{a}_{\nu}$ ,  $\mathbf{b}_{\nu}$ ,  $\boldsymbol{\phi}_1$  and  $\boldsymbol{\psi}_1$  are all proportional to the identity tensor and thus appear as scalars in (1.15a), (1.15b).

In this paper we examine the special case of (1.15a), (1.15b) which corresponds to the assumptions  $a_\nu = 0, b_\nu = 0, \nu \geq 1$  and  $\mathbf{E}(\tau) = \mathbf{0}, \mathbf{B}(\tau) = \mathbf{0}, -\infty < \tau < 0$ , i.e.

$$(1.16a) \quad \mathbf{D}(t) = a_0 \mathbf{E}(t) + \int_0^t \phi(t - \tau) \mathbf{E}(\tau) d\tau,$$

$$(1.16b) \quad \mathbf{H}(t) = b_0 \mathbf{B}(t) + \int_0^t \psi(t - \tau) \mathbf{B}(\tau) d\tau.$$

This special case of a holohedral isotropic nonconducting material dielectric still embodies a separation of electric and magnetic effects in the constitutive theory but generalizes the Maxwell–Hopkinson theory in that magnetic memory effects are taken into account through the presence of the kernel function  $\psi(t)$ . In the next section we will formulate an initial-boundary value problem for the electric displacement field  $\mathbf{D}(t)$  in a holohedral isotropic dielectric; provided  $\psi(0) \neq 0$ ,  $\mathbf{D}(t)$  will be shown to satisfy a (nonhomogeneous) integrodifferential equation. By introducing suitable Hilbert spaces and operators, the initial-boundary value problem for  $\mathbf{D}(t)$  is easily demonstrated to be equivalent to an initial value problem for an abstract integrodifferential equation and growth estimates for specific classes of solutions to this abstract problem are then obtained by employing a suitable logarithmic convexity argument.

**2. Initial-boundary value problems for holohedral isotropic dielectrics.** Let  $(x^i, t)$  be a fixed Lorentz reference frame; the local forms of Maxwell’s equations are then given by (1.7), (1.8). Let  $\Omega \subseteq R^3$  be a bounded region with boundary  $\partial\Omega$  and assume that  $\partial\Omega$  is sufficiently smooth so that the divergence theorem may be applied. Finally, assume that  $\Omega$  is filled with a holohedral isotropic nonconducting dielectric material which is nondeformable and which satisfies the hypotheses of § 1 so that, in  $\Omega$ , the electromagnetic field satisfies constitutive relations of the form (1.16a), (1.16b) where we assume that  $a_0 > 0, b_0 > 0$  and  $\phi(t), \psi(t)$  are monotonically decreasing functions which are (at least) twice continuously differentiable on  $[0, \infty)$  with  $\psi^{(3)}(t)$  a bounded integrable function on  $[0, \infty)$ . The basic result of this section is

**THEOREM 2.1.** *The evolution of the electric displacement field  $\mathbf{D}(x, t)$  in any holohedral isotropic nonconducting material dielectric (which conforms to the constitutive hypotheses (1.16a), (1.16b)) is governed by the system of equations*

$$(2.1) \quad \frac{\partial^2 D_i}{\partial t^2} + \Psi(0) \frac{\partial D_i}{\partial t} - b_0 \dot{\Psi}(0) \left[ c_0 \delta_{ij} \delta_{jl} \frac{\partial^2 D_k}{\partial x_j \partial x_l} - D_i \right] + b_0 \int_0^t \left( \dot{\Psi}(t - \tau) D_i(\tau) - \Phi_0(t - \tau) \delta_{ik} \delta_{jl} \frac{\partial^2 D_k(\tau)}{\partial x_j \partial x_l} \right) d\tau = b_0 \dot{\Psi}(t) D_i(0),$$

where  $c_0 = 1/(a_0 \dot{\Psi}(0))$ ,  $\Phi_0(t) = \Phi(t)/a_0$  and

$$(2.2) \quad \begin{aligned} \Phi(t) &= \sum_{n=1}^{\infty} (-1)^n \phi^n(t), \\ \phi^n(t) &= \int_0^t \phi^1(t - \tau) \phi^{n-1}(\tau) d\tau, \quad n \geq 2, \\ \phi^1(t) &= a_0^{-1} \phi(t) \end{aligned}$$

with an analogous definition for  $\Psi(t)$  in terms of  $\psi(t)$ .

*Proof.* By using successive approximations we may invert the constitutive relations (1.16a) and (1.16b) to obtain, respectively,

$$(2.3a) \quad \mathbf{E}(t) = \frac{1}{a_0} \mathbf{D}(t) + \frac{1}{a_0} \int_0^t \Phi(t-\tau) \mathbf{D}(\tau) \, d\tau,$$

$$(2.3b) \quad \mathbf{B}(t) = \frac{1}{b_0} \mathbf{H}(t) + \frac{1}{b_0} \int_0^t \Psi(t-\tau) \mathbf{H}(\tau) \, d\tau$$

with  $\Phi(t)$  and  $\Psi(t)$  defined in terms of  $\phi(t)$  and  $\psi(t)$  respectively, as indicated in (2.2). From (2.3a) and the second Maxwell relation in (1.8)  $\text{div } \mathbf{E}(t) = 0$  so

$$(2.4) \quad \Delta \mathbf{E}(t) = -\text{curl curl } \mathbf{E}(t).$$

From (2.3b), however, and the first Maxwell relation in (1.7)

$$(2.5) \quad \text{curl } \mathbf{E}(t) = -\mathbf{B}_t = -\frac{1}{b_0} \mathbf{H}_t - \frac{1}{b_0} \Psi(0) \mathbf{H}(t) - \int_0^t \Psi_t(t-\tau) \mathbf{H}(\tau) \, d\tau.$$

Therefore,

$$(2.6) \quad \begin{aligned} & -\text{curl curl } \mathbf{E}(t) \\ &= \frac{1}{b_0} (\text{curl } \mathbf{H})_t + \frac{1}{b_0} \Psi(0) (\text{curl } \mathbf{H}(t)) + \int_0^t \Psi_t(t-\tau) (\text{curl } \mathbf{H}(\tau)) \, d\tau \\ &= \frac{1}{b_0} \mathbf{D}_{tt} + \frac{1}{b_0} \Psi(0) \mathbf{D}_t + \int_0^t \Psi_t(t-\tau) \mathbf{D}_t(\tau) \, d\tau, \end{aligned}$$

where the second relation in (2.6) follows from the first Maxwell equation in (1.7). Combining (2.6<sub>2</sub>) with (2.4) and employing (2.3a) we obtain

$$(2.7) \quad \mathbf{D}_{tt} + \Psi(0) \mathbf{D}_t + b_0 \int_0^t \Psi_t(t-\tau) \mathbf{D}_t(\tau) \, d\tau = \frac{b_0}{a_0} \Delta \mathbf{D}(t) + \frac{b_0}{a_0} \int_0^t \Phi(t-\tau) \Delta \mathbf{D}(\tau) \, d\tau.$$

However,

$$(2.8) \quad \int_0^t \Psi_t(t-\tau) \mathbf{D}_t(\tau) \, d\tau = \dot{\Psi}(0) \mathbf{D}(t) - \dot{\Psi}(t) \mathbf{D}(0) + \int_0^t \Psi_{tt}(t-\tau) \mathbf{D}(\tau) \, d\tau.$$

Substituting (2.8) into (2.7) we have on  $\Omega \times [0, \infty)$ :

$$(2.9) \quad \begin{aligned} & \mathbf{D}_{tt} + \Psi(0) \mathbf{D}_t + b_0 \dot{\Psi}(0) (\mathbf{I} - c_0 \Delta) \mathbf{D}(t) \\ &+ b_0 \int_0^t (\Psi_{tt}(t-\tau) \mathbf{I} - \Phi_0(t-\tau) \Delta) \mathbf{D}(\tau) \, d\tau = b_0 \dot{\Psi}(t) \mathbf{D}(0), \end{aligned}$$

where  $c_0 = 1/(a_0 \dot{\Psi}(0))$  and  $\Phi_0(t) = \Phi(t)/a_0$ . Q.E.D.

In conjunction with the integrodifferential equation (2.9) we consider initial and boundary data of the form

$$(2.10a) \quad \mathbf{D}(\mathbf{x}, 0) = \mathbf{D}_0(\mathbf{x}), \quad \mathbf{D}_t(\mathbf{x}, 0) = \mathbf{D}_1(\mathbf{x}), \quad \mathbf{x} \in \bar{\Omega},$$

$$(2.10b) \quad \mathbf{D}(\mathbf{x}, t) = \mathbf{0}, \quad (\mathbf{x}, t) \in \partial\Omega \times [0, \infty),$$

where  $\mathbf{D}_0, \mathbf{D}_1$  are continuous on  $\bar{\Omega}$ . At this point it is convenient to recast the initial-boundary value problem (2.9), (2.10a), (2.10b) as an initial-value problem for an

integrodifferential equation in Hilbert space.<sup>2</sup> As in [6] we let  $C_0^\infty(\Omega)$  denote the set of three dimensional vector fields with compact support in  $\Omega$  whose components are in  $C_0^\infty(\Omega)$ . We take  $H = L_2(\Omega)$ , i.e., the completion of  $C_0^\infty(\Omega)$  under the norm induced by the inner product

$$(2.11) \quad \langle \mathbf{v}, \mathbf{w} \rangle_{L_2} \equiv \int_{\Omega} v_i w_i \, d\mathbf{x}$$

while the Hilbert space  $H_+$  is taken to be  $H_0^1(\Omega)$  the completion of  $C_0^\infty(\Omega)$  under the norm induced by the inner product

$$(2.12) \quad \langle \mathbf{v}, \mathbf{w} \rangle_{H_0^1} \equiv \int_{\Omega} \frac{\partial v_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, d\mathbf{x}.$$

Finally,  $H_- = H^{-1}(\Omega)$ , the Hilbert space obtained by completing  $C_0^\infty(\Omega)$  under the norm

$$(2.13) \quad \|\mathbf{v}\|_{H^{-1}} \equiv \sup_{\mathbf{w} \in H_0^1} \left[ \left| \int_{\Omega} v_i w_i \, d\mathbf{x} \right| / \left( \int_{\Omega} \frac{\partial w_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, d\mathbf{x} \right)^{1/2} \right].$$

It is known that  $H_0^1(\Omega) \subseteq L_2(\Omega)$  (both topologically and algebraically) and that  $H_0^1$  is dense in  $L_2$ . We denote by  $\omega$  the embedding constant for the inclusion map  $i: H_0^1(\Omega) \rightarrow L_2(\Omega)$ .

Operators  $\mathbf{L} \in \mathcal{L}(H_0^1, H^{-1})$  and  $\mathbf{M}(t) \in L^2((-\infty, \infty); \mathcal{L}(H_0^1, H^{-1}))$  are now defined as follows:

$$(2.14a) \quad (\mathbf{L}\mathbf{v})_i \equiv b_0 \dot{\Psi}(0) \left[ c_0 \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} - \delta_{ij} v_j \right], \quad \mathbf{v} \in H_0^1(\Omega),$$

$$(2.14b) \quad (\mathbf{M}(t)\mathbf{v})_i \equiv b_0 \left[ \ddot{\Psi}(t) \delta_{ij} v_j - \Phi_0(t) \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} \right], \quad v \in H_0^1(\Omega), t \in (-\infty, \infty),$$

where the derivatives are taken in the distribution sense. It follows directly from these definitions and the smoothness assumptions on  $\phi(t)$  and  $\psi(t)$  that

- (i)  $\mathbf{L} \in \mathcal{L}_S(H_0^1, H^{-1})$ ,  $\mathbf{M}(t) \in \mathcal{L}_S(H_0^1, H^{-1})$ ,  $t \in (-\infty, \infty)$ ;
- (ii)  $\mathbf{M}_t(\cdot) \in L^2((-\infty, \infty); \mathcal{L}(H_0^1, H^{-1}))$ ;

where  $\mathcal{L}_S(H_0^1, H^{-1})$  denotes the space of all symmetric bounded linear operators from  $H_0^1$  into  $H^{-1}$  and  $\mathbf{M}_t$  is the strong operator derivative of  $\mathbf{M}(\cdot)$ . Thus the system (2.1), (2.10a), (2.10b) is equivalent to

$$(2.15) \quad \mathbf{D}_t + \Psi(0)\mathbf{D}_t - \mathbf{L}\mathbf{D} + \int_0^t \mathbf{M}(t-\tau)\mathbf{D}(\tau) \, d\tau = b_0 \dot{\Psi}(t)\mathbf{D}_0,$$

$$(2.16) \quad \mathbf{D}(0) = \mathbf{D}_0, \quad \mathbf{D}_t(0) = \mathbf{D}_1,$$

where  $\mathbf{D}_0, \mathbf{D}_1 \in H_0^1$  and  $\mathbf{D} \in C^2([0, \infty); H_0^1)$ . Actually, we shall be interested in solutions of (2.15), (2.16) on finite time intervals of the form  $[0, T)$  where  $T, 0 < T < \infty$ , is an arbitrary real number; this suggests that we examine the following abstract initial-value problem: Let  $H, H_+$  be Hilbert spaces with inner products  $\langle \cdot, \cdot \rangle$  and  $\langle \cdot, \cdot \rangle_+$ , respectively, and assume that  $H_+ \subseteq H$  (algebraically and topologically) with  $H_+$  dense in  $H$ ;

<sup>2</sup> We specify, below, three spaces  $H, H_+$ , and  $H_-$  which are taken to be certain Sobolev spaces in the application and which satisfy certain mild requirements in the general development.

define  $H_-$  as in (1.2). We consider solutions  $\mathbf{u} \in C^2([0, T]; H_+)$  of the system

$$(2.17) \quad \mathbf{u}_t - \alpha \mathbf{u}_t - \mathbf{L}\mathbf{u} + \int_0^t \mathbf{M}(t - \tau)\mathbf{u}(\tau) \, d\tau = \beta(t)\mathbf{u}_0, \quad 0 \leq t < T,$$

$$(2.18) \quad \mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_t(0) = \mathbf{u}_1 \quad (\mathbf{u}_0, \mathbf{u}_1 \in H_+),$$

where  $\alpha \neq 0$  is an arbitrary real constant,  $\beta(t)$  is any real-valued function such that  $\dot{\beta}(t)$  exists a.e. on  $[0, T]$ ,  $L \in \mathcal{L}_S(H_+, H_-)$  and  $\mathbf{M}(\cdot), \mathbf{M}_t(\cdot) \in L^2([0, T]; \mathcal{L}_S(H_+, H_-))$ . We assume that  $\mathbf{u}_t \in C^1([0, T]; H_+)$  and  $\mathbf{u}_t \in C([0, T]; H_-)$ .

In § 3 we derive some growth estimates for solutions  $\mathbf{u}(t)$  of the system (2.17), (2.18), which lie in the set  $\mathcal{N}$ . Our estimates will be obtained under various combinations of the following hypotheses:

$$\alpha \begin{cases} > 0, \\ < 0, \end{cases} \quad \mathbf{u}_0 \begin{cases} = \mathbf{0}, \\ \neq \mathbf{0}, \end{cases} \quad \text{and} \quad \beta(t) \begin{cases} = 0, & 0 \leq t < T, \\ \neq 0, & \text{on } [0, T]. \end{cases}$$

In § 4 we apply our results to the system consisting of (2.1), (2.10a), and (2.10b); at no point in this work do we make any definiteness assumptions on the operators  $\mathbf{L}$  or  $\mathbf{M}(t)$ ,  $t \in [0, T]$ .

**3. Some growth estimates.** Let  $\mathcal{K}(t) = \frac{1}{2}\|\mathbf{u}_t\|^2$  denote the kinetic energy associated with solutions  $\mathbf{u}$  of the system (2.17), (2.18) and  $\mathcal{P}(t) = -\frac{1}{2}\langle \mathbf{u}, N\mathbf{u} \rangle$  the potential energy; then  $\mathcal{E}(t) \equiv \mathcal{K}(t) + \mathcal{P}(t)$  is the total energy. Let  $\gamma$  and  $t_0$  be arbitrary nonnegative real numbers and define

$$(3.1) \quad F(t; \gamma, t_0) \equiv \|\mathbf{u}(t)\|^2 + \gamma(t + t_0)^2, \quad 0 \leq t < T.$$

The growth estimates in this section all follow from the following basic lemma.

LEMMA. Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18). Suppose that

$$(3.2a) \quad -\langle \mathbf{v}, \mathbf{M}(0)\mathbf{v} \rangle \leq \kappa \|\mathbf{v}\|_+^2, \quad \forall \mathbf{v} \in H_+$$

with

$$(3.2b) \quad \kappa \geq \gamma T \sup_{[0, T]} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)}.$$

Then there exists  $\mu > 0$  such that for all  $t$ ,  $0 \leq t < T$

$$(3.3) \quad \begin{aligned} FF'' - F'^2 \geq & -2F(2\mathcal{E}(0) + \mu) + \alpha FF' - 2\alpha F \left( \gamma(t + t_0) + 4 \int_0^t \mathcal{K}(\tau) \, d\tau \right) \\ & + 2F \left( 2 \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}, \mathbf{u}_0 \rangle \, d\tau - \beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle \right) + 4F\beta(0)\|\mathbf{u}_0\|^2. \end{aligned}$$

*Proof.* From the definition of  $F(t; \gamma, t_0)$ , i.e., (3.1), we compute

$$(3.4) \quad F'(t; \gamma, t_0) = 2\langle \mathbf{u}, \mathbf{u}_t \rangle + \gamma(t + t_0)$$

$$(3.5) \quad \begin{aligned} F''(t; \gamma, t_0) = & 2\|\mathbf{u}_t\|^2 + 2\alpha \langle \mathbf{u}, \mathbf{u}_t \rangle + 2\langle \mathbf{u}, \mathbf{L}\mathbf{u} \rangle \\ & - 2 \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t - \tau)\mathbf{u}(\tau) \, d\tau \right\rangle + 2\beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle + 2\gamma, \end{aligned}$$

where we have made use of (2.17) in (3.5). Using the definitions of  $\mathcal{K}(t)$ ,  $\mathcal{E}(t)$ , we may



rewrite (3.5) in the form

$$\begin{aligned}
 F''(t; \gamma, t_0) &= 2\alpha \langle \mathbf{u}, \mathbf{u}_t \rangle + 2\beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle - 2 \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle \\
 (3.6) \qquad \qquad \qquad &+ 4(2\mathcal{H}(t) + \gamma) - 2(2\mathcal{E}(0) + \gamma) - 4(\mathcal{E}(t) - \mathcal{E}(0)).
 \end{aligned}$$

However, for any  $\tau, 0 \leq \tau \leq t < T$ ,

$$(3.7) \quad \mathcal{E}'(\tau) = \langle \mathbf{u}_\tau, \mathbf{u}_{\tau\tau} \rangle - \langle \mathbf{u}_\tau, \mathbf{L}\mathbf{u} \rangle = \alpha \|\mathbf{u}_\tau\|^2 + \beta(\tau) \langle \mathbf{u}_\tau, \mathbf{u}_0 \rangle - \left\langle \mathbf{u}_\tau, \int_0^\tau \mathbf{M}(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle.$$

Therefore,

$$\begin{aligned}
 \mathcal{E}'(\tau) &= 2\alpha \mathcal{H}(\tau) + \beta(\tau) \langle \mathbf{u}_\tau, \mathbf{u}_0 \rangle - \frac{d}{d\tau} \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle \\
 (3.8) \qquad \qquad \qquad &+ \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle + \langle \mathbf{u}(\tau), \mathbf{M}(0)\mathbf{u}(\tau) \rangle.
 \end{aligned}$$

Integrating this last result from zero to  $t$  and substituting for  $\mathcal{E}(t) - \mathcal{E}(0)$  in (3.6) we obtain

$$\begin{aligned}
 F''(t; \gamma, t_0) &= 2\alpha \langle \mathbf{u}, \mathbf{u}_t \rangle + 2\beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle + 2 \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle \\
 (3.9) \qquad \qquad \qquad &+ 4(2\mathcal{H}(t) + \gamma) - 2(2\mathcal{E}(0) + \gamma) - 8\alpha \int_0^t \mathcal{H}(\tau) d\tau - 4 \int_0^t \beta(\tau) \langle \mathbf{u}_\tau, \mathbf{u}_0 \rangle d\tau \\
 &- 4 \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle d\tau - 4 \int_0^t \langle \mathbf{u}(\tau), \mathbf{M}(0)\mathbf{u}(\tau) \rangle d\tau.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 FF'' - F'^2 &= 4F(2\mathcal{H}(t) + \gamma) - F'^2 + 2\alpha F(2\mathcal{E}(0) + \gamma) + 2\alpha F \left( \langle \mathbf{u}, \mathbf{u}_t \rangle - 4 \int_0^t \mathcal{H}(\tau) d\tau \right) \\
 (3.10) \qquad \qquad \qquad &+ 2F \left( \beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle - 2 \int_0^t \beta(\tau) \langle \mathbf{u}_\tau, \mathbf{u}_0 \rangle d\tau \right) + 2F \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle \\
 &+ 4F \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle d\tau - 4F \int_0^t \langle \mathbf{u}(\tau), \mathbf{M}(0)\mathbf{u}(\tau) \rangle d\tau.
 \end{aligned}$$

However, from (3.1), (3.4), the definition of  $\mathcal{H}(t)$ , and the Schwarz inequality it follows that

$$(3.11) \qquad \qquad \qquad G(t; \gamma, t_0) \equiv 4F(t; \gamma, t_0)(2\mathcal{H}(t) + \gamma) - F'^2(t; \gamma, t_0) \geq 0$$

and, therefore, (3.10) yields the inequality

$$\begin{aligned}
 FF'' - F'^2 &\geq -2F(2\mathcal{E}(0) + \gamma) + \alpha F \left( \frac{d}{dt} \|\mathbf{u}\|^2 - 8 \int_0^t \mathcal{H}(\tau) d\tau \right) \\
 &\quad + 2F \left( 2 \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau - \beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle \right) + 4F\beta(0) \|\mathbf{u}_0\|^2 \\
 (3.12) \quad &\quad + 2F \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle - 4F \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle d\tau \\
 &\quad - 4F \int_0^t \langle \mathbf{u}(\tau), \mathbf{M}(0) \mathbf{u}(\tau) \rangle d\tau.
 \end{aligned}$$

If we make note of the fact that

$$\frac{d}{dt} \|\mathbf{u}\|^2 = F'(t; \gamma, t_0) - 2\gamma(t + t_0),$$

then we can rewrite (3.12) in the form

$$\begin{aligned}
 FF'' - F'^2 &\geq -2F(2\mathcal{E}(0) + \gamma) + \alpha FF' - 2\alpha F \left( \gamma(t + t_0) + 4 \int_0^t \mathcal{H}(\tau) d\tau \right) \\
 (3.13) \quad &\quad + 2F \left( 2 \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau - \beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle \right) + 4F\beta(0) \|\mathbf{u}_0\| \\
 &\quad + 2F \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle - 4F \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau-\sigma) \mathbf{u}(\sigma) d\sigma \right\rangle d\tau \\
 &\quad - 4F \int_0^t \langle \mathbf{u}(\tau), \mathbf{M}(0) \mathbf{u}(\tau) \rangle d\tau.
 \end{aligned}$$

In order to complete the proof of the lemma we now use the hypotheses (3.2a), (3.2b) and the fact that  $u \in \mathcal{N}$  to bound, from below, the sum of the last three terms in (3.13), i.e.,

$$\begin{aligned}
 &\left| \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle \right| \\
 (3.14a) \quad &\cong \|\mathbf{u}(t)\| \int_0^t \|\mathbf{M}(t-\tau) \mathbf{u}(\tau)\| d\tau \cong \omega \|\mathbf{u}(t)\|_+ \int_0^t (\|\mathbf{M}(t-\tau)\|_{\mathcal{L}(H_+, H_-)}) \|\mathbf{u}(\tau)\|_+ d\tau \\
 &\cong \omega T \left( \sup_{[0, T]} \|\mathbf{u}\|_+ \right)^2 \sup_{[0, T]} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)} \cong \omega N^2 T \sup_{[0, T]} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}
 \end{aligned}$$

and thus, as  $F(t; \gamma, t_0) \geq 0, 0 \leq t < T$ ,

$$(3.14b) \quad 2F \left\langle \mathbf{u}, \int_0^t \mathbf{M}(t-\tau) \mathbf{u}(\tau) d\tau \right\rangle \geq -2\omega N^2 T \sup_{[0, T]} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)} F(t; \gamma, t_0).$$

Also,

$$\begin{aligned}
 (3.15) \quad &-4F \int_0^t \langle \mathbf{u}(\tau), \mathbf{M}(0) \mathbf{u}(\tau) \rangle d\tau \geq 4\kappa F \int_0^t \|\mathbf{u}(\tau)\|_+^2 d\tau \\
 &\geq 4\omega T \sup_{[0, T]} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)} F \int_0^t \|\mathbf{u}(\tau)\|_+^2 d\tau
 \end{aligned}$$

by virtue of (3.2a) and (3.2b). Finally

$$\begin{aligned}
 \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau - \sigma)\mathbf{u}(\sigma) d\sigma \right\rangle d\tau &\cong \int_0^t \left| \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau - \sigma)\mathbf{u}(\sigma) d\sigma \right\rangle \right| d\tau \\
 &\cong \int_0^t \|\mathbf{u}(\tau)\| \left( \int_0^\tau (\|\mathbf{M}_\tau(\tau - \sigma)\|_{\mathcal{L}(H_+, H_-)}) \|\mathbf{u}(\sigma)\|_+ d\sigma \right) d\tau \\
 (3.16a) \quad &\cong \omega \sup_{[0, T)} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)} \int_0^t \|\mathbf{u}(\tau)\|_+ \left( \int_0^\tau \|\mathbf{u}(\sigma)\|_+ d\sigma \right) d\tau \\
 &\cong \omega \sup_{[0, T)} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)} \left( \int_0^t \|\mathbf{u}(\tau)\|_+ d\tau \right)^2 \\
 &\cong \omega T \sup_{[0, T)} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)} \int_0^t \|\mathbf{u}(\tau)\|_+^2 d\tau
 \end{aligned}$$

from which we easily deduce that

$$\begin{aligned}
 -4F \int_0^t \left\langle \mathbf{u}(\tau), \int_0^\tau \mathbf{M}_\tau(\tau - \sigma)\mathbf{u}(\sigma) d\sigma \right\rangle d\tau \\
 (3.16b) \quad &\cong -4\omega T \sup_{[0, T)} \|\mathbf{M}_t\|_{\mathcal{L}(H_+, H_-)} F \int_0^t \|\mathbf{u}(\tau)\|_+^2 d\tau.
 \end{aligned}$$

Combining (3.13) with the estimates (3.14b), (3.15<sub>2</sub>) and (3.16b) we obtain the estimate (3.3) with

$$(3.17) \quad \mu \cong \gamma + \omega N^2 T_{[0, T)} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}. \quad \text{Q.E.D.}$$

With the preceding lemma as a starting point we now begin our study of the growth behavior of solutions to (2.17), (2.18) which lie in the class  $\mathcal{N}$ ; in each of the cases examined below we assume that  $\mathbf{M}(0)$  satisfies (3.2a) for some  $\kappa > 0$  which satisfies (3.2b).

Case I.  $\mathbf{u}_0 = 0$  and  $\alpha < 0$ . In this case  $\mathcal{E}(0) = \frac{1}{2}\|\mathbf{u}_1\|^2$  and the second expression on the right-hand side of (3.3) is nonnegative; thus

$$(3.18) \quad FF'' - F'^2 \cong -2F(\|\mathbf{u}_1\|^2 + \mu) - |\alpha|FF'$$

for all  $t, 0 \leq t < T$ , where  $\mu$  is given by (3.17). However, for  $\gamma, t_0$  arbitrary nonnegative real numbers,

$$(3.19) \quad \lambda \gamma t_0^2 \leq \lambda \|\mathbf{u}(t)\|^2 + \lambda \gamma (t + t_0)^2 \cong \lambda F(t; \gamma; t_0)$$

for any  $\lambda \geq 0$ . If, in particular, we choose

$$(3.20) \quad \lambda = \lambda(\gamma; t_0) \cong 2(\|\mathbf{u}_1\|^2 + \mu) / \gamma t_0^2$$

then for all  $t, 0 \leq t < T$ , and all  $\gamma, t_0 \geq 0$

$$(3.21) \quad 2(\|\mathbf{u}_1\|^2 + \mu) \leq \lambda(\gamma; t_0)F(t; \gamma, t_0)$$

and (3.18) may be replaced by the estimate

$$(3.22) \quad FF'' - F'^2 \geq -\lambda(\gamma; t_0)F^2 - |\alpha|FF'.$$

The differential inequality (3.22) now forms the basis for the following growth estimate:

**THEOREM 3.1.** *Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) with  $\mathbf{u}_0 = \mathbf{0}$  and  $\alpha < 0$ . Assume that  $\mathbf{M}(0)$  satisfies (3.2a), (3.2b) and that  $T > 1/|\alpha|$ . Then there exists a constant*

$M > 0$  such that

$$(3.23) \quad \|\mathbf{u}(t)\|^2 \leq M^{1-\bar{\delta}} e^{-(\bar{\lambda}/|\alpha|)t}, \quad 0 \leq t < T,$$

where  $\bar{\delta}$  is given by (3.27).

*Proof.* From (3.22) and Jensen’s inequality we obtain the estimate,

$$(3.24) \quad F(t; \gamma, t_0) \leq e^{-(\lambda/|\alpha|)t} [F(t_1; \gamma, t_0) e^{(\lambda/|\alpha|)t_1}]^\delta [F(t_2; \gamma, t_0) e^{(\lambda/|\alpha|)t_2}]^{1-\delta}$$

(valid for  $0 \leq t_1 < t \leq t_2 < T$ ) where

$$(3.25) \quad \delta(t) = (e^{-|\alpha|t} - e^{-|\alpha|t_2}) / (e^{-|\alpha|t_1} - e^{-|\alpha|t_2}).$$

The interval  $[t_1, t_2] \subseteq [0, T)$  is any closed interval such that  $F(t; \gamma, t_0) > 0, t_1 \leq t \leq t_2$ . However, it is a simple consequence of (3.24) and the definition of  $F(t; \gamma, t_0)$  that  $F(t; \gamma, t_0) \equiv 0$  on  $[0, T)$  if  $F(\bar{t}; \gamma, t_0) = 0$  for any  $\bar{t} \in [0, T)$ . Thus, without loss of generality, we may assume that  $F(t; \gamma, t_0) > 0, 0 \leq t < T$ . Taking  $t_1 = 0, t_2 = T$  in (3.14) we obtain

$$(3.26) \quad F(t; \gamma, t_0) \leq e^{-(\lambda/|\alpha|)t} [\gamma t_0^2]^\delta [F(T; \gamma, t_0) e^{(\lambda/|\alpha|)T}]^{1-\delta},$$

where

$$(3.27) \quad \bar{\delta}(t) = (e^{-|\alpha|t} - e^{-|\alpha|T}) / (1 - e^{-|\alpha|T}).$$

We now choose  $\gamma = 1/t_0^2$  and then take the limit in (3.26) as  $t_0 \rightarrow +\infty$ . Clearly, as

$$(3.28) \quad F(t; \gamma 1/t_0^2, t_0) = \|\mathbf{u}(t)\|^2 + \left(\frac{t}{t_0} + 1\right)^2,$$

$$\lim_{t_0 \rightarrow +\infty} F(t; 1/t_0^2, t_0) = \|\mathbf{u}(t)\|^2 + 1$$

for all  $t \in [0, T)$ . Also, as  $\mathbf{u} \in \mathcal{N}$

$$(3.29) \quad \lim_{t_0 \rightarrow +\infty} F(T; 1/t_0^2, t_0) = \lim_{t_0 \rightarrow +\infty} \left( \|\mathbf{u}(T)\|^2 + \left(\frac{T}{t_0} + 1\right)^2 \right) \leq \omega^2 N^2 + 1,$$

$$(3.30) \quad \lim_{t_0 \rightarrow +\infty} \lambda(1/t_0^2; t_0) = \lim_{t_0 \rightarrow +\infty} 2(\|\mathbf{u}_1\|^2 + 1/t_0^2 + \bar{\mu}) = 2(\|\mathbf{u}_1\|^2 + \bar{\mu}) \equiv \bar{\lambda},$$

where  $\bar{\mu} = \omega N^2 T \sup_{[0, T)} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}$ . Thus, with  $\gamma = 1/t_0^2$  and  $t_0 \rightarrow +\infty$  in (3.26), we obtain the estimate

$$(3.31) \quad \|\mathbf{u}(t)\|^2 \leq e^{-(\bar{\lambda}/|\alpha|)t} [(\omega^2 N^2 + 1) e^{(\bar{\lambda}/|\alpha|)T}]^{1-\bar{\delta}}, \quad 0 \leq t < T$$

and the result, which shows that  $\|\mathbf{u}\|^2$  is bounded above by an exponentially decreasing function of  $t$  for all  $t \in [0, T)$ , follows by choosing  $M > 0$  so large that  $\omega^2 N^2 + 1 < M \exp(-\bar{\lambda}/|\alpha|)$ .

In contrast to the result contained in the statement of Theorem 3.1, we have the following theorem concerning lower bounds for solutions  $\mathbf{u} \in \mathcal{N}$  of (2.17), (2.18).

**THEOREM 3.2.** *Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) with  $\mathbf{u}_0 = 0$  and  $\alpha < 0$  and assume that  $\mathbf{M}(0)$  satisfies (3.2a), (3.2b). If  $|\alpha| < 1$  then there exists  $T > 0$  such that  $\|\mathbf{u}\|^2$  is bounded below by a monotonically increasing exponential function of  $t, 0 \leq t < T$ .*

*Proof.* We begin by integrating the differential inequality (3.22) according to the “tangent property” of convex functions—assuming that  $F(t; \gamma, t_0) > 0, 0 \leq t < T$ , where  $T > 0$  is an arbitrary real number; by the “tangent property” for convex functions we

refer to the fact that the graph of a convex function<sup>3</sup> on  $[0, T]$  lies above the tangent line to the curve at any point  $\bar{t} \in [0, T]$ . Thus, we obtain directly from (3.22) the estimate

$$(3.32) \quad F(t; \gamma, t_0) \geq F(0; \gamma; t_0) \exp \left[ \left\{ \frac{F'(0; \gamma, t_0) + (\lambda/|\alpha|)F(0; \gamma, t_0)}{|\alpha|F(0; \gamma, t_0)} \right\} (1 - e^{-|\alpha|t}) - \frac{\lambda}{|\alpha|} t \right].$$

However,  $F(0; \gamma, t_0) = \gamma t_0^2$  and  $F'(0; \gamma, t_0) = 2\gamma t_0$ . Therefore, if we set  $\gamma = 1/t_0^2$  in (3.40) we obtain

$$(3.33) \quad \|u(t)\|^2 + [t/t_0 + 1]^2 \geq \exp [\chi(t; t_0)], \quad 0 \leq t < T,$$

where

$$(3.34) \quad \chi(t; t_0) \equiv \frac{1}{|\alpha|} \left[ \left( \frac{2}{t_0} + \frac{\lambda(1/t_0^2; t_0)}{|\alpha|} \right) (1 - e^{-|\alpha|t}) - \lambda(1/t_0^2; t_0)t \right]$$

and

$$(3.35) \quad \lambda(1/t_0^2; t_0) = 2 \left( \|u_1\|^2 + \frac{1}{t_0^2} + \omega^2 N^2 T \sup_{[0, T]} \|M\|_{\mathcal{L}(H_+, H_-)} \right).$$

We note, in passing, that  $\chi(0; t_0) = 0$ . For the sake of convenience we now set

$$\varepsilon(t_0) = \frac{2}{t_0} + \frac{\lambda(1/t_0^2; t_0)}{|\alpha|}.$$

Then

$$(3.36) \quad \chi'(t; t_0) = \varepsilon(t_0) e^{-|\alpha|t} - \lambda(1/t_0^2; t_0).$$

From (3.36) it follows immediately that  $\chi'(t; t_0) > 0$  for

$$0 < t < \frac{1}{|\alpha|} \ln \left\{ \frac{\varepsilon(t_0)}{\lambda(1/t_0^2; t_0)} \right\}$$

provided  $\varepsilon(t_0) > \lambda(1/t_0^2; t_0)$ . We now take the limit in (3.33) as  $t_0 \rightarrow +\infty$  and obtain

$$(3.37) \quad \|u(t)\|^2 + 1 \geq \exp \left[ \lim_{t_0 \rightarrow +\infty} \chi(t; t_0) \right], \quad 0 \leq t < T.$$

But

$$(3.38) \quad \begin{aligned} \lim_{t_0 \rightarrow +\infty} \chi(t; t_0) &= \frac{1}{|\alpha|} \left[ \lim_{t_0 \rightarrow +\infty} \varepsilon(t_0) (1 - e^{-|\alpha|t}) - \lim_{t_0 \rightarrow +\infty} \lambda(1/t_0^2; t_0) \right] \\ &= \frac{\bar{\lambda}}{|\alpha|^2} (1 - e^{-|\alpha|t}) - \bar{\lambda}t \equiv \bar{\chi}(t), \end{aligned}$$

where  $\bar{\lambda}$  is given by (3.30). Also

$$(3.39) \quad \lim_{t_0 \rightarrow +\infty} \chi'(t; t_0) = \frac{d}{dt} \bar{\chi}(t) = \bar{\lambda} \left( \frac{e^{-|\alpha|t}}{|\alpha|} - 1 \right)$$

and, therefore,

$$(3.40) \quad \bar{\chi}'(t) > 0, \quad 0 \leq t < \frac{1}{|\alpha|} \ln \left( \frac{1}{|\alpha|} \right)$$

<sup>3</sup> The inequality (3.22) and the assumption that  $F(t; \gamma, t_0) > 0$  on  $[0, T]$  imply that  $\ln(F(\sigma; \gamma, t_0) e^{-\lambda/\alpha^2})$  is a convex function of  $\sigma = e^{-|\alpha|t}$  on  $[0, T]$ .

if  $|\alpha| < 1$ . The statement of the theorem now follows with  $T = (1/|\alpha|) \ln(1/|\alpha|)$ , i.e.,

$$(3.41) \quad \|\mathbf{u}(t)\|^2 + 1 \geq \exp(\bar{\chi}(t)), \quad 0 \leq t < \frac{1}{|\alpha|} \ln\left(\frac{1}{|\alpha|}\right),$$

where  $\bar{\chi}(t)$ , as determined by (3.38), is nonnegative and monotonically increasing on  $[0, (1/|\alpha|) \ln(1/|\alpha|)]$ . Q.E.D.

*Case II.*  $\mathbf{u}_0 = \mathbf{0}$  and  $\alpha > 0$ . In this case the expression  $H(t; \gamma; t_0) \equiv -2\alpha F(\gamma(t+t_0) + 4 \int_0^t K(\tau) d\tau)$  can not be dropped from the differential inequality (3.3). As  $t < T$  and  $\alpha > 0$ , (3.3) with  $\mathbf{u}_0 = \mathbf{0}$  implies that

$$(3.42) \quad FF'' - F'^2 \geq -2F(\|\mathbf{u}_1\|^2 + \mu) + \alpha FF' - 2\alpha F\left(\gamma(T+t_0) + 2 \int_0^t \|\mathbf{u}_\tau\|^2 d\tau\right).$$

In order to proceed further we shall need the following lemma.

**LEMMA.** *Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) with  $\mathbf{u}_0 = \mathbf{0}$ . Then there exists a real-valued continuous function  $h_\alpha(t)$ , defined for  $0 \leq t < T$ , such that*

$$(3.43) \quad \frac{1}{2t} \int_0^t \|\mathbf{u}_\tau\|^2 d\tau \leq \|\mathbf{u}_1\|^2 + h_\alpha(T), \quad 0 \leq t < T.$$

*Proof.* From the identity

$$\mathbf{u}_t = \int_0^t \mathbf{u}_{\tau\tau} d\tau + \mathbf{u}_1$$

and (2.17), we obtain

$$(3.44) \quad \mathbf{u}_t = \mathbf{u}_1 + \alpha \mathbf{u} + \int_0^t \mathbf{L}\mathbf{u}(\tau) d\tau - \int_0^t \int_0^\tau \mathbf{M}(\tau - \sigma) \mathbf{u}(\sigma) d\sigma d\tau.$$

Thus,

$$(3.45) \quad \begin{aligned} \|\mathbf{u}_t\| &\leq \|\mathbf{u}_1\| + \alpha \|\mathbf{u}(t)\| + \int_0^t \|\mathbf{L}\|_{\mathcal{L}(H_+, H_-)} \|\mathbf{u}(\tau)\|_+ d\tau \\ &\quad + \int_0^t \int_0^\tau \|\mathbf{M}(t - \sigma)\|_{\mathcal{L}(H_+, H_-)} \|\mathbf{u}(\sigma)\|_+ d\sigma d\tau \\ &\leq \|\mathbf{u}_1\| + \alpha \omega \|\mathbf{u}(t)\|_+ + t \|\mathbf{L}\|_{\mathcal{L}(H_+, H_-)} \sup_{[0, T]} \|\mathbf{u}(\tau)\|_+ \\ &\quad + \frac{t^2}{2} \sup_{[0, T]} \|\mathbf{M}(\tau)\|_{\mathcal{L}(H_+, H_-)} \sup_{[0, T]} \|\mathbf{u}(\tau)\|_+ \\ &\leq \|\mathbf{u}_1\| + p_\alpha(t) \sup_{[0, T]} \|\mathbf{u}(\tau)\|_+, \end{aligned}$$

where

$$(3.46) \quad p_\alpha(t) \equiv \alpha \omega + t \|\mathbf{L}\|_{\mathcal{L}(H_+, H_-)} + \frac{t^2}{2} \sup_{[0, T]} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}.$$

Clearly  $p_\alpha(t) < p_\alpha(T)$ , for all  $t \in [0, T)$  and, as  $\mathbf{u} \in \mathcal{N}$

$$(3.47) \quad \|\mathbf{u}_1\| \leq \|\mathbf{u}_1\| + N p_\alpha(T), \quad 0 \leq t < T.$$

Therefore,

$$(3.48) \quad \int_0^t \|\mathbf{u}_\tau\|^2 d\tau \leq 2t(\|\mathbf{u}_1\|^2 + N^2 p_\alpha^2(T)), \quad 0 \leq t < T$$

and the lemma follows with

$$(3.49) \quad h_\alpha(t) = N^2 p_\alpha^2(t).$$

If we combine (3.42) with (3.43) we obtain

$$(3.50) \quad FF'' - F'^2 \geq -2F(\|\mathbf{u}_1\|^2 + \tilde{\mu}) + \alpha FF',$$

where  $\tilde{\mu} > 0$  is defined by

$$(3.51) \quad \tilde{\mu} = \mu + \alpha[\gamma(T + t_0) + 4T(\|\mathbf{u}_1\|^2 + h_\alpha(T))].$$

Choosing

$$(3.52) \quad \lambda^* = \lambda^*(\gamma; t_0) = \frac{2(\|\mathbf{u}_1\|^2 + \tilde{\mu})}{\gamma t_0^2}$$

we have

$$(3.53) \quad FF'' - F'^2 \geq -\lambda^*(\gamma; t_0)F^2 + \alpha FF', \quad 0 \leq t < T.$$

If we apply Jensen's inequality to (3.53) we obtain

$$(3.54) \quad F(t; \gamma, t_0) \leq e^{\lambda^* t / \alpha} [\gamma t_0^2]^{\delta^*} [F(T; \gamma, t_0) e^{(\lambda^* / \alpha)T}]^{1 - \delta^*}, \quad 0 \leq t < T,$$

where

$$(3.55) \quad \delta^*(t) = (e^{\alpha t} - e^{\alpha T}) / (1 - e^{\alpha T}), \quad 0 \leq t < T.$$

Taking  $\gamma = 1/t_0^2$  in (3.54), extracting the limit as  $t_0 \rightarrow +\infty$ , and then choosing  $\varrho > 0$  so large that  $\omega^2 N^2 + 1 \leq \varrho e^{(\lambda^* / \alpha)T}$  we obtain the estimate

$$(3.56) \quad \|\mathbf{u}(t)\|^2 \leq \varrho^{1 - \delta^*} e^{(\lambda^* / \alpha)t}, \quad 0 \leq t < T.$$

To close out our study of the case  $\mathbf{u}_0 = 0$ ,  $\alpha > 0$  we now integrate the differential inequality (3.53) according to the "tangent property" of convex functions and we obtain

$$(3.57) \quad F(t; \gamma, t_0) \geq \gamma t_0^2 \exp \left[ \left\{ \frac{2\gamma t_0 - (\lambda^* / \alpha)\gamma t_0^2}{-\alpha\gamma t_0^2} \right\} (1 - e^{\alpha t}) + \frac{\lambda^*}{\alpha} t \right]$$

which, with  $\gamma = 1/t_0^2$ ,  $\lambda^* = \lambda^*(1/t_0^2; t_0)$ , reduces to

$$(3.58) \quad \|\mathbf{u}(t)\|^2 + \left( \frac{t}{t_0} + 1 \right)^2 \geq \exp \left[ \left\{ \frac{\lambda^*}{\alpha^2} - \frac{2}{\alpha t_0} \right\} \cdot (1 - e^{\alpha t}) + \frac{\lambda^*}{\alpha} t \right].$$

Were we to follow the arguments previously employed we would, at this point, take the limit in (3.58) as  $t_0 \rightarrow +\infty$ . This procedure, however, does not lead to a viable lower bound for  $\|\mathbf{u}\|^2$  in this case. It is worthwhile, however, to examine the function

$$(3.59) \quad J(t; \gamma, t_0) \equiv \left( \frac{\lambda^*(\gamma; t_0)}{\alpha^2} - \frac{2}{\alpha t_0} \right) \cdot (1 - e^{\alpha t}) + \frac{\lambda^*(\gamma; t_0)t}{\alpha}.$$

Clearly,  $J(0; \gamma, t_0) = 0$  for arbitrary nonnegative constants  $\gamma, t_0$ . Also

$$(3.60) \quad J'(t; \gamma, t_0) = \left( \frac{2}{t_0} - \frac{\lambda^*(\gamma; t_0)}{\alpha} \right) e^{\alpha t} + \frac{\lambda^*(\gamma; t_0)}{\alpha}$$

from which, by the definition of  $\lambda^*$ , it follows that

$$(3.61) \quad \left(\frac{\alpha\gamma t_0^2}{2}\right) J'(t; \gamma, t_0) = (k_1 + k_2\gamma)(1 - e^{-\alpha t}) + \alpha\gamma t_0,$$

where

$$(3.62a) \quad k_1 = \|\mathbf{u}_1\|^2(1 + 4\alpha T) + \bar{\mu} + 4\alpha T h_\alpha(T),$$

$$(3.62b) \quad k_2 = 1 + \alpha T.$$

Thus, if we choose

$$(3.63) \quad t_0 = t_{0,\gamma} \equiv \frac{(k_1 + k_2\gamma)}{\alpha\gamma} (e^{\alpha T} - 1), \quad \gamma > 0,$$

then  $J'(t; \gamma, t_{0,\gamma}) > 0$  for all  $t, 0 \leq t \leq T$ , and each real  $\gamma > 0$ , and we can state the following result.

**THEOREM 3.3.** *Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) with  $\mathbf{u}_0 = \mathbf{0}$  and  $\alpha > 0$  and assume that  $\mathbf{M}(0)$  satisfies (3.2a), (3.2b). Then for any  $T > 0$  there exists  $\mathcal{Q} > 0$  such that  $\|\mathbf{u}\|^2$  satisfies (3.56) and, or each real  $\gamma > 0$ ,  $\|\mathbf{u}\|^2$  also satisfies*

$$(3.64) \quad \|\mathbf{u}(t)\|^2 + \gamma(t + t_{0,\gamma})^2 \geq \gamma t_{0,\gamma}^2 \exp[J(t; \gamma, t_{0,\gamma})], \quad 0 \leq t < T,$$

where  $t_{0,\gamma}$  is defined by (3.61a), (3.62b), and (3.63) and  $J(t; \gamma, t_{0,\gamma})$ , defined by (3.59) with  $t_0 = t_{0,\gamma}$ , is nonnegative and strictly monotonically increasing on  $[0, T)$ .

The results obtained in Cases I and II did not involve any hypotheses concerning the sign of the initial energy  $\mathcal{E}(0)$ ; as we assumed  $\mathbf{u}_0 = \mathbf{0}$  in both cases,  $\mathcal{E}(0) = \frac{1}{2}\|\mathbf{u}_1\|^2 > 0$  if  $\mathbf{u}_1 \neq \mathbf{0}$ . In the cases considered below we remove the restriction that  $\mathbf{u}_0 = \mathbf{0}$ .

*Case III.*  $\mathbf{u}_0 \neq \mathbf{0}$ ,  $\alpha < 0$ , and  $\beta(t) = 0, 0 \leq t < T$ . In this case (provided we use the fact that  $\alpha < 0$  to delete the term  $H(t; \gamma, t_0)$ ) inequality (3.3) reduces to

$$(3.65) \quad FF'' - F'^2 \geq -2F(\|\mathbf{u}_1\|^2 - \langle \mathbf{u}_0, \mathbf{L}\mathbf{u}_0 \rangle + \mu) - |\alpha|FF'$$

with  $\mu$  given by (3.17). We now assume that the initial data  $\mathbf{u}_0, \mathbf{u}_1$  satisfies

$$(3.66) \quad \|\mathbf{u}_1\|^2 - \langle \mathbf{u}_0, \mathbf{L}\mathbf{u}_0 \rangle < -\bar{\mu},$$

where  $\bar{\mu} = \omega N^2 T \sup_{[0,T)} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}$ . Taking  $\gamma = 0$  in (3.65) we obtain  $(F(t) = \|\mathbf{u}(t)\|^2)$ ,

$$(3.67) \quad F(t)F''(t) - [F'(t)]^2 \geq -|\alpha|F(t)F'(t), \quad 0 \leq t < T,$$

Jensen's inequality then yields the upper bound

$$(3.68a) \quad \|\mathbf{u}(t)\|^2 \leq \|\mathbf{u}_0\|^{2\delta} \|\mathbf{u}(T)\|^{2(1-\delta)}, \quad 0 \leq t < T.$$

We note that the hypothesis that  $\mathbf{u} \in \mathcal{N}$ , and (3.68), imply that there exists  $R > 0$  such that

$$(3.68b) \quad \|\mathbf{u}(t)\|^2 \leq R^{1-\delta} \|\mathbf{u}_0\|^{2\delta}, \quad 0 \leq t < T.$$

However, as (3.66) can not be valid for  $\|\mathbf{u}_0\|$  sufficiently small, (3.68b) represents only an upper bound on  $\|\mathbf{u}(t)\|$  in terms of  $\|\mathbf{u}_0\|$  and not a stability estimate. A better result is found by integrating (3.67) according to the "tangent property" of convex functions; in fact, directly from (3.32) with  $\lambda = 0$  and  $F(t; \gamma, t_0)$  replaced by  $F(t) \equiv \|\mathbf{u}(t)\|^2$  we obtain

$$(3.69) \quad \|\mathbf{u}(t)\|^2 \geq \|\mathbf{u}_0\|^2 \exp\left[\frac{2\langle \mathbf{u}_1, \mathbf{u}_0 \rangle}{|\alpha| \|\mathbf{u}_0\|^2} (1 - e^{-|\alpha|t})\right], \quad 0 \leq t < T.$$



From the estimate (3.69) it is obvious that if either  $\langle \mathbf{u}_0, \mathbf{u}_1 \rangle = 0$  or  $\mathbf{u}_1 = \mathbf{0}$  (and  $\langle \mathbf{u}_0, \mathbf{L}\mathbf{u}_0 \rangle > \bar{\mu}$ ) then  $\|\mathbf{u}(t)\|^2 \geq \|\mathbf{u}_0\|^2$  for all  $t \in [0, T)$ . On the other hand, if  $\langle \mathbf{u}_1, \mathbf{u}_0 \rangle > 0$ , then on  $[0, T)$ ,  $\|\mathbf{u}(t)\|^2$  is bounded below by a monotonically increasing exponential function of  $t$ . Finally if  $\langle \mathbf{u}_0, \mathbf{u}_1 \rangle < 0$  then  $\|\mathbf{u}(t)\|^2$  can not decay any faster than a monotonically decreasing exponential function of  $t$ . Our results are summarized as:

**THEOREM 3.4.** *Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) with  $\mathbf{u}_0 \neq \mathbf{0}$ ,  $\alpha < 0$ , and  $\beta(t) \equiv 0$  on  $[0, T)$ . Assume that  $M(0)$  satisfies (3.2a) and (3.2b). Then*

(A) *If the initial data satisfy (3.66),  $\|\mathbf{u}(t)\|$  is bounded above by  $\|\mathbf{u}_0\|$  according to (3.68b),  $0 \leq t < T$ .*

(B) *If the initial data satisfy (3.66) then there exists  $K(\alpha)$  such that for all  $t$ ,  $0 \leq t < T$ ,*

$$(3.70) \quad \|\mathbf{u}(t)\|^2 \geq \|\mathbf{u}_0\|^2 \exp [K(\alpha)(1 - e^{-|\alpha|t})],$$

where for each real  $\alpha$ ,  $K(\alpha)$  is real-valued and

- (i)  $K(\alpha) = 0$  if either  $\mathbf{u}_1 = \mathbf{0}$  or  $\langle \mathbf{u}_0, \mathbf{u}_1 \rangle = 0$ ;
- (ii)  $K(\alpha) > 0$  if  $\langle \mathbf{u}_0, \mathbf{u}_1 \rangle > 0$ ;
- (iii)  $K(\alpha) < 0$  if  $\langle \mathbf{u}_0, \mathbf{u}_1 \rangle < 0$ ;

and

- (iv)  $|K(\alpha)| \rightarrow 0$  as  $|\alpha| \rightarrow \infty$ .

*Remark.* The case  $\mathbf{u}_0 \neq \mathbf{0}$ ,  $\alpha > 0$ , and  $\beta(t) \equiv 0$  can be treated in the same manner as Case III; in fact, from (3.50) (which was derived under the assumption that  $\mathbf{u}_0 = \mathbf{0}$  with  $\alpha > 0$ ) we can write down immediately the differential inequality

$$(3.71) \quad FF'' - F'^2 \geq -2F(\|\mathbf{u}_1\|^2 - \langle \mathbf{u}_0, \mathbf{L}\mathbf{u}_0 \rangle + \mu) + \alpha FF'$$

for the case where  $\mathbf{u}_0 \neq \mathbf{0}$ ,  $\alpha > 0$ , but  $\beta(t) \equiv 0$ ; in (3.71)  $\tilde{\mu}$  is defined by (3.51). Suppose we set  $\gamma = 0$ ; then if the initial data satisfy

$$(1 + 4\alpha T)\|\mathbf{u}_1\|^2 - \langle \mathbf{u}_0, \mathbf{L}\mathbf{u}_0 \rangle \leq -(\bar{\mu} + 4\alpha Th_\alpha(T)),$$

the above differential inequality reduces to

$$(3.72) \quad F(t)F''(t) - [F'(t)]^2 \geq \alpha F(t)F'(t), \quad 0 \leq t < T,$$

where  $F(t) = \|\mathbf{u}(t)\|^2$ . We leave the integration of (3.72) and the analysis of the resulting estimates on  $\|\mathbf{u}(t)\|^2$  to the reader and turn, instead, to consider a case where both  $\mathbf{u}_0 \neq \mathbf{0}$  and  $\beta(t) \neq 0$ .

*Case IV.*  $\mathbf{u}_0 \neq \mathbf{0}$ ,  $\beta(t) \neq 0$ ,  $\alpha < 0$  and  $\beta(0) > 0$ . In this case (3.3) is easily seen to imply that

$$(3.73) \quad \begin{aligned} FF'' - F'^2 &\geq -2F(2\mathcal{E}(0) + \mu) - |\alpha|FF' \\ &\quad + 2F\left(2 \int_0^t \dot{\beta}(\tau)\langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau - \beta(t)\langle \mathbf{u}, \mathbf{u}_0 \rangle\right) + 4F\beta(0)\|\mathbf{u}_0\|^2 \\ &= -2F(2\mathcal{E}(0) - 2\beta(0)\|\mathbf{u}_0\|^2 + \mu) - |\alpha|FF' \\ &\quad + 2F\left(2 \int_0^t \dot{\beta}(\tau)\langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau - \beta(t)\langle \mathbf{u}, \mathbf{u}_0 \rangle\right). \end{aligned}$$

In order to proceed further we must bound from below the third expression on the right-hand side of the differential inequality (3.73<sub>2</sub>); this is accomplished by the following lemma.

LEMMA. Suppose that  $\dot{\beta}(t)$  is bounded on  $[0, T)$  for each fixed  $T, 0 < T < \infty$ . Then there exists a constant  $C > 0$  such that

$$(3.74) \quad 2 \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau - \beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle \geq -C \|\mathbf{u}_0\|, \quad 0 \leq t < T.$$

*Proof.* We set  $\rho = \sup_{[0, T)} |\dot{\beta}(t)| < \infty$ . Then

$$(3.75) \quad \begin{aligned} \left| \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}(\tau), \mathbf{u}_0 \rangle d\tau \right| &= \left| \left\langle \int_0^t \dot{\beta}(\tau) \mathbf{u}(\tau) d\tau, \mathbf{u}_0 \right\rangle \right| \\ &\leq \left( \int_0^t |\dot{\beta}(\tau)| \|\mathbf{u}(\tau)\| d\tau \right) \|\mathbf{u}_0\| \\ &\leq \rho \left( \int_0^T \|\mathbf{u}(\tau)\| d\tau \right) \|\mathbf{u}_0\| \leq \rho \omega N T \|\mathbf{u}_0\| \end{aligned}$$

so

$$(3.76) \quad \int_0^t \dot{\beta}(\tau) \langle \mathbf{u}, \mathbf{u}_0 \rangle d\tau \geq -\rho \omega N T \|\mathbf{u}_0\|, \quad 0 \leq t < T.$$

Also

$$(3.77) \quad \begin{aligned} |\beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle| &\leq |\beta(t)| \cdot |\langle \mathbf{u}, \mathbf{u}_0 \rangle| \\ &\leq \omega N |\beta(t)| \cdot \|\mathbf{u}_0\| \\ &\leq \omega N \left| \int_0^t \dot{\beta}(\tau) d\tau + \beta(0) \right| \|\mathbf{u}_0\| \leq \omega N (\rho T + \beta(0)) \|\mathbf{u}_0\| \end{aligned}$$

so

$$(3.78) \quad -\beta(t) \langle \mathbf{u}, \mathbf{u}_0 \rangle \geq -\omega N (\rho T + \beta(0)) \|\mathbf{u}_0\|, \quad 0 \leq t < T.$$

Combining (3.76) and (3.78) we obtain (3.74) with

$$(3.79) \quad C = \omega N (3\rho T + \beta(0)) > 0.$$

We now return to (3.73<sub>2</sub>); in view of the last lemma this latter inequality implies that

$$(3.80) \quad FF'' - F'^2 \geq -2F(\|\mathbf{u}_1\|^2 + \sum(\mathbf{u}_0) + \mu) - |\alpha| FF',$$

where  $\sum: H_+ \rightarrow R^+$  is defined by

$$(3.81) \quad \sum(\mathbf{w}) = 2\beta(0) \|\mathbf{w}\| \left( \frac{C}{2\beta(0)} - \|\mathbf{w}\| \right) - \langle \mathbf{w}, \mathbf{Lw} \rangle, \quad \mathbf{w} \in H_+.$$

If we set  $\gamma = 0$  then (3.80) reduces to

$$(3.82) \quad F(t)F''(t) - [F'(t)]^2 \geq -2F(t)(\|\mathbf{u}_1\|^2 + \sum(\mathbf{u}_0) + \bar{\mu}) - |\alpha| F(t)$$

with  $F(t) = \|\mathbf{u}(t)\|^2$  and  $\bar{\mu} = \omega N^2 \sup_{[0, T)} \|\mathbf{M}(t)\|_{\mathcal{L}(H_+, H_-)}$  and we have the following result.

THEOREM 3.5. Let  $\mathbf{u} \in \mathcal{N}$  be any solution of (2.17), (2.18) where  $\mathbf{u}_0 \neq 0, \beta(t) \neq 0, \alpha < 0$ , and  $\beta(0) > 0$ . Assume that  $\mathbf{M}(0)$  satisfies (3.2a), (3.2b) and that  $\dot{\beta}(t)$  is bounded for  $0 \leq t < T$ . Then if the initial data satisfy

$$(3.83) \quad \|\mathbf{u}_1\|^2 + \sum(\mathbf{u}_0) \leq -\bar{\mu},$$

where  $\sum$  is defined by (3.81),  $\|\mathbf{u}(t)\|$  satisfies the estimates (3.68) and (3.69). In particular, if  $\mathbf{u}_1 = 0$  and  $\sum(\mathbf{u}_0) \leq -\bar{\mu}$  then  $\|\mathbf{u}(t)\|^2 \geq \|\mathbf{u}_0\|^2$  for all  $t, 0 \leq t < T$ .

*Remark.* We leave for the reader the consideration of the other cases possible when  $\mathbf{u}_0 \neq \mathbf{0}$  and  $\beta(t) \neq 0$ , e.g.,  $\alpha < 0$  and  $\beta(0) \leq 0$ ; the stability and growth estimates which apply in these situations may easily be derived by suitably modifying the last lemma and making use of the basic differential inequalities derived for the previous cases.

**4. Applications to bounds for electric displacement fields.** In order to apply the results of the previous section to solutions of the initial-boundary value problem (2.1), (2.10a), (2.10b) (associated with the constitutive relations (1.16a), (1.16b)) we must delineate the form assumed by the basic hypothesis (3.2a), (3.2b). In other words, for the operator  $\mathbf{M}(t)$ , which is defined by (2.14b), we wish to examine the implications of the requirement that

$$(4.1) \quad -\langle \mathbf{v}, \mathbf{M}(0)\mathbf{v} \rangle_{L_2} \geq \kappa \|\mathbf{v}\|_{H_0^1}^2, \quad \mathbf{v} \in H_0^1(\Omega),$$

with  $\kappa \geq \omega T \sup_{[0,T]} \|\mathbf{M}_t\|_{\mathcal{L}(H_0^1, H^{-1})}$ . From (2.14b) and (2.11) we easily compute

$$(4.2) \quad \begin{aligned} \langle \mathbf{v}, \mathbf{M}(0)\mathbf{v} \rangle_{L_2} &= - \int_{\Omega} (\mathbf{M}(0)\mathbf{v})_i v_i \, d\mathbf{x} \\ &= -b_0 \ddot{\Psi}(0) \int_{\Omega} \delta_{ij} v_i v_j \, d\mathbf{x} + \frac{b_0}{a_0} \Phi(0) \int_{\Omega} \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} v_i \, d\mathbf{x} \\ &= -b_0 \ddot{\Psi}(0) \|\mathbf{v}\|_{L_2}^2 + \frac{b_0}{a_0} \Phi(0) \int_{\Omega} \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} v_i \, d\mathbf{x} \end{aligned}$$

for any  $\mathbf{v} \in H_0^1$ . But if  $\mathbf{v} \in H_0^1$  then

$$(4.3) \quad \begin{aligned} \int_{\Omega} \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} v_i \, d\mathbf{x} &= \int_{\Omega} \delta_{jl} v_k \frac{\partial^2 v_k}{\partial x_j \partial x_l} \, d\mathbf{x} \\ &= - \int_{\Omega} \delta_{jl} \frac{\partial v_k}{\partial x_j} \frac{\partial v_k}{\partial x_l} \, d\mathbf{x} = -\|\mathbf{v}\|_{H_0^1}^2, \end{aligned}$$

where we have used integration by parts together with the fact that  $\mathbf{v}$  vanishes on  $\partial\Omega^4$ . Thus

$$(4.4) \quad \begin{aligned} -\langle \mathbf{v}, \mathbf{M}(0)\mathbf{v} \rangle_{L_2} &= -b_0 \ddot{\Psi}(0) \|\mathbf{v}\|_{L_2}^2 - \frac{b_0}{a_0} \Phi(0) \|\mathbf{v}\|_{H_0^1}^2 \\ &\geq -b_0 \left( \omega^2 |\ddot{\Psi}(0)| + \frac{1}{a_0} \Phi(0) \right) \|\mathbf{v}\|_{H_0^1}^2. \end{aligned}$$

Therefore, (4.1a) will be satisfied if

$$(4.5) \quad -b_0 \left( \omega^2 |\ddot{\Psi}(0)| + \frac{1}{a_0} \Phi(0) \right) \geq \kappa$$

with  $\kappa \geq \omega T \sup_{[0,T]} \|\mathbf{M}_t\|_{\mathcal{L}(H_0^1, H^{-1})}$ . For the sake of convenience we now set  $Y(t) = \ddot{\Psi}(t)$ . From (2.14b) again we have,

$$(4.6) \quad (\mathbf{M}_t \mathbf{v})_i = b_0 \left[ \dot{Y}(t) \delta_{ij} v_j - \frac{\dot{\Phi}(t)}{a_0} \delta_{ik} \delta_{jl} \frac{\partial^2 v_k}{\partial x_j \partial x_l} \right], \quad \mathbf{v} \in H_0^1,$$

<sup>4</sup> This follows from the definition of  $H_0^1$  and a standard trace theorem.

so

$$\begin{aligned}
 \langle \mathbf{v}, \mathbf{M}_t \mathbf{v} \rangle_{L_2} &= \left| \int_{\Omega} [\mathbf{M}_t v]_i v_i \, dx \right| \\
 &= \left| b_0 \dot{Y}(t) \|\mathbf{v}\|_{L_2}^2 - \frac{b_0}{a_0} \dot{\Phi}(t) \int_{\Omega} \delta_{jk} v_k \frac{\partial^2 v_k}{\partial x_j \partial x_j} \, dx \right| \\
 (4.7) \quad &= b_0 |\dot{Y}(t)| \|\mathbf{v}\|_{L_2}^2 + \frac{1}{a_0} \dot{\Phi}(t) \|\mathbf{v}\|_{H_0^1}^2 \\
 &\leq b_0 \left( \omega^2 |\dot{Y}(t)| + \frac{1}{a_0} |\dot{\Phi}(t)| \right) \|\mathbf{v}\|_{H_0^1}^2.
 \end{aligned}$$

It now follows that for each  $t, 0 \leq t < T$ ,

$$(4.8) \quad \|\mathbf{M}_t\|_{\mathcal{L}(H_0^2, H^{-1})} = \sup_{v \in H_0^1} \frac{|\langle \mathbf{v}, \mathbf{M}_t \mathbf{v} \rangle|}{\|\mathbf{v}\|_{H_0^1}^2} \leq b_0 \left( \omega^2 |\dot{Y}(t)| + \frac{1}{a_0} |\dot{\Phi}(t)| \right).$$

Thus, (4.1b) will be satisfied if

$$(4.9) \quad \kappa \geq \omega T b_0 \left( \omega^2 \sup_{[0, T]} |\dot{Y}(t)| + \frac{1}{a_0} \sup_{[0, T]} |\dot{\Phi}(t)| \right).$$

Combining (4.5) and (4.9) we find that a condition which insures the validity of (4.1) is

$$(4.10) \quad -\left( \omega^2 |Y(0)| + \frac{1}{a_0} \Phi(0) \right) \geq \omega T \left( \omega^2 \sup_{[0, T]} |\dot{Y}(t)| + \frac{1}{a_0} \sup_{[0, T]} |\dot{\Phi}(t)| \right).$$

It is clear, from (4.10), that this inequality can be satisfied only if  $\Phi(0) < 0$  with  $|\Phi(0)| > a_0 \omega^2 |Y(0)|$ . It is worthwhile, at this point, to recall the following result which has been proven in [6]:

LEMMA. Let  $\phi(t) \in C^1[0, T]$  and assume that the series defining  $\Phi(t)$  as well as the derived series, which is obtained by term by term differentiation, are uniformly convergent on every interval  $[0, T - \varepsilon], 0 < \varepsilon < T$ . If  $\sup_{[0, T]} |\phi(t)| < a_0/T$  then

$$(4.11) \quad (i) \quad \sup_{[0, T]} |\dot{\Phi}(t)| \leq \mathcal{F}(T);$$

$$(4.12) \quad (ii) \quad \sup_{[0, T]} |\dot{\Phi}(t)| \leq \frac{\mathcal{F}(T)}{T} \left\{ 1 + T \frac{\sup_{[0, T]} |\dot{\phi}(t)|}{\sup_{[0, T]} |\phi(t)|} \right\};$$

where

$$(4.13) \quad \mathcal{F}(T) = \sup_{[0, T]} |\phi(t)| \left( a_0 - T \sup_{[0, T]} |\phi(t)| \right).$$

Remark. Similar results hold for  $\sup_{[0, T]} |\Psi(t)|$  and  $\sup_{[0, T]} |\dot{\Psi}(t)|$ , of course, under analogous assumptions on  $\psi(t)$  and the series defining  $\Psi(t)$ , e.g., we require that  $\sup_{[0, T]} |\psi(t)| < b_0/T$ ; the constant  $\mathcal{F}(T)$  appearing in (4.11), (4.12) would, in this case, be replaced by

$$(4.14) \quad \mathcal{G}(T) = \sup_{[0, T]} |\psi(t)| \left( b_0 - T \sup_{[0, T]} |\psi(t)| \right).$$

In recalling the above lemma we have been motivated by a desire to replace the sufficient condition represented by (4.10) by a condition which involves only the basic memory functions  $\phi(t), \psi(t)$  specified in the constitutive relations (1.16a), (1.16b). To

this end we note that the equations defining  $\Phi(t)$  in terms of  $\phi(t)$  and  $\Psi(t)$  in terms of  $\psi(t)$  imply, respectively, that

$$(4.15a) \quad \Phi(t) + \frac{1}{a_0} \phi(t) = -\frac{1}{a_0} \int_0^t \phi(t-\tau)\Phi(\tau) \, d\tau,$$

$$(4.15b) \quad \Psi(t) + \frac{1}{b_0} \psi(t) = -\frac{1}{b_0} \int_0^t \psi(t-\tau)\Psi(\tau) \, d\tau.$$

From (4.15a) and (4.15b) we immediately obtain

$$(4.16) \quad \Phi(0) = -\frac{1}{a_0} \phi(0), \quad \Psi(0) = -\frac{1}{b_0} \psi(0)$$

and thus (4.10) can only be satisfied if  $\phi(0) > 0$ . Directly from (4.15b) we now compute that

$$(4.17a) \quad \dot{\Psi}(t) + \frac{1}{b_0} \dot{\psi}(t) = -\frac{1}{b_0} \psi(0)\Psi(t) - \frac{1}{b_0} \int_0^t \psi_t(t-\tau)\Psi(\tau) \, d\tau,$$

$$(4.17b) \quad \ddot{\Psi}(t) + \frac{1}{b_0} \ddot{\psi}(t) = -\frac{1}{b_0} \psi(0)\dot{\Psi}(t) - \frac{1}{b_0} \dot{\psi}(0)\Psi(t) - \frac{1}{b_0} \int_0^t \psi_{tt}(t-\tau)\Psi(\tau) \, d\tau.$$

Therefore,

$$(4.18) \quad \ddot{\Psi}(0) \equiv Y(0) = -\frac{1}{b_0} (\ddot{\psi}(0) + \psi(0)\dot{\Psi}(0) + \dot{\psi}(0)\Psi(0)).$$

However, from (4.16) and (4.17a),

$$(4.19) \quad \dot{\Psi}(0) = -\frac{1}{b_0} \dot{\psi}(0) - \frac{1}{b_0} \psi(0)\Psi(0) = -\frac{1}{b_0} \dot{\psi}(0) + \frac{1}{b_0^2} \psi^2(0).$$

Combining (4.16<sub>2</sub>) and (4.19<sub>2</sub>) with (4.18) we have, finally,

$$(4.20) \quad Y(0) = -\frac{1}{b_0} \left( \frac{1}{b_0^2} \psi^3(0) - \frac{2}{b_0} \psi(0)\dot{\psi}(0) + \ddot{\psi}(0) \right).$$

The left-hand side of (4.10) now assumes the form

$$(4.21) \quad \frac{1}{a_0^2} \phi(0) - \frac{\omega^2}{b_0} \left| \frac{1}{b_0^2} \psi^3(0) - \frac{2}{b_0} \psi(0)\dot{\psi}(0) + \ddot{\psi}(0) \right|.$$

We now turn our attention to the right-hand side of (4.10). Directly from (4.17b) we obtain

$$(4.22) \quad \dot{Y}(t) = -\frac{1}{b_0} \left( \psi^{(3)}(t) + \psi(0)Y(t) + \dot{\psi}(0)\dot{\Psi}(t) + \ddot{\psi}(0)\Psi(t) + \int_0^t \psi_{tt}(t-\tau)\Psi(\tau) \, d\tau \right).$$

Also,

$$(4.23) \quad \sup_{[0,T]} |Y(t)| \leq \frac{1}{b_0} \left[ \sup_{[0,T]} |\ddot{\psi}(t)| + |\psi(0)| \sup_{[0,T]} |\dot{\psi}(t)| + \left( |\dot{\psi}(0)| + T \sup_{[0,T]} |\ddot{\psi}(t)| \right) \sup_{[0,T]} |\Psi(t)| \right]$$

while, by (4.22),

$$(4.24) \quad \sup_{[0,T]} |\dot{Y}(t)| \leq \frac{1}{b_0} \left[ \sup_{[0,T]} |\psi^{(3)}(t)| + \psi(0) \sup_{[0,T]} |Y(t)| \right. \\ \left. + \dot{\psi}(0) \sup_{[0,T]} |\dot{\Psi}(t)| + \left( |\ddot{\psi}(0)| + T \sup_{[0,T]} |\psi^{(3)}(t)| \right) \sup_{[0,T]} |\Psi(t)| \right].$$

If we substitute for  $\sup_{[0,T]} |Y(t)|$  in (4.24) from (4.23) we obtain an estimate of the form

$$(4.25) \quad \sup_{[0,T]} |\dot{Y}(t)| \leq \mathcal{A} \sup_{[0,T]} |\Psi(t)| + \mathcal{B} \sup_{[0,T]} |\dot{\Psi}(t)| + \mathcal{C},$$

where, the constants  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  are given by

$$\mathcal{A} = \frac{1}{b_0} \left[ T \sup_{[0,T]} |\psi^{(3)}(t)| + |\ddot{\psi}(0)| + \frac{|\dot{\psi}(0)|}{b_0} \left( |\dot{\psi}(0)| + T \sup_{[0,T]} |\psi^{(2)}(t)| \right) \right], \\ \mathcal{B} = \frac{1}{b_0} \left[ |\dot{\psi}(0)| + \frac{\psi^2(0)}{b_0} \right], \\ \mathcal{C} = \frac{1}{b_0} \left[ \sup_{[0,T]} |\psi^{(3)}(t)| + \frac{1}{b_0} |\psi(0)| \sup_{[0,T]} |\psi^{(2)}(t)| \right].$$

As a result of the estimate (4.25), the right-hand side of the inequality (4.10) is bounded above by the expression

$$(4.26) \quad \omega^3 T \left( \mathcal{A} \sup_{[0,T]} |\Psi(t)| + \mathcal{B} \sup_{[0,T]} |\dot{\Psi}(t)| + \mathcal{C} \right) + \frac{\omega T}{a_0} \sup_{[0,T]} |\dot{\Phi}(t)|,$$

which, in view of the preceding lemma, is itself bounded above by

$$(4.27) \quad \mathcal{D} \equiv \omega^3 T \left[ \mathcal{A} \mathcal{G}(T) + \frac{\mathcal{B} \mathcal{G}(T)}{T} \left( 1 + T \frac{\sup_{[0,T]} |\dot{\psi}(t)|}{\sup_{[0,T]} |\psi(t)|} \right) + \mathcal{C} \right] \\ + \frac{\omega \mathcal{F}(T)}{a_0} \left( 1 + T \frac{\sup_{[0,T]} |\dot{\phi}(t)|}{\sup_{[0,T]} |\phi(t)|} \right)$$

provided  $\sup_{[0,T]} |\phi(t)| < a_0/T$  and  $\sup_{[0,T]} |\psi(t)| < b_0/T$ .

From (4.27), the definitions of the constants  $\mathcal{A}, \mathcal{B}, \mathcal{C}$ , (4.13), and (4.14), it is clear that

$$(4.28) \quad \mathcal{D} = \mathcal{D} \left( \omega, T, a_0, b_0, |\psi^{(i)}(0)|, \sup_{[0,T]} |\phi^{(i)}(t)|, \sup_{[0,T]} |\psi^{(k)}(t)| \right)$$

with  $i = 0, 1, 2, j = 0, 1$ , and  $k = 0, 1, 2, 3$ . Thus,  $\mathcal{D}$  is computable once  $\Omega, T > 0$ , and the constitutive relations (1.16a), (1.16b) are specified. Thus (4.1) will be satisfied provided

$$(4.29) \quad \frac{1}{a_0^2} \phi(0) - \frac{\omega^2}{b_0} \left| \frac{1}{b_0^2} \psi^3(0) - \frac{2}{b_0} \psi(0) \dot{\psi}(0) + \ddot{\psi}(0) \right| \geq \mathcal{D}.$$

We offer below an example of the kind of considerations which are involved in verifying that (4.29) is satisfied.

*Example.* In the constitutive equations (1.16a), (1.16b) we take

$$(4.30) \quad \phi(t) = e^{-Kt}, \quad \psi(t) = e^{-t},$$

where  $K > 0$  is arbitrary; for the sake of convenience we set  $T = 1$ . The region  $\Omega \subseteq \mathcal{R}^3$

(and hence the embedding constant  $\omega$ ) are left arbitrary at this point as are the constants  $a_0, b_0$ . From (4.30) we have

$$(4.31) \quad \phi(0) = \sup_{[0,1]} |\phi(t)| = 1, \quad \sup_{[0,1]} |\dot{\phi}(t)| = K$$

and

$$(4.32a) \quad \sup_{[0,1]} |\psi^{(k)}(t)| = 1, \quad k = 0, 1, 2, 3,$$

$$(4.32b) \quad \psi(0) = \dot{\psi}(0) = 1, \quad \dot{\psi}(0) = -1.$$

Therefore, the constants  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  in (4.25) are given by

$$(4.33) \quad \mathcal{A} = \frac{2}{b_0} \left(1 + \frac{1}{b_0}\right), \quad \mathcal{B} = \mathcal{C} = \frac{1}{b_0} \left(1 + \frac{1}{b_0}\right).$$

Also, if  $a_0 > 1, b_0 > 1$ , then from (4.13) and (4.14)

$$(4.34) \quad \mathcal{F}(1) = \frac{1}{a_0 - 1}, \quad \mathcal{G}(1) = \frac{1}{b_0 - 1}.$$

Combining our results it follows that (4.29) will be satisfied if  $a_0, b_0$ , and  $\omega$  are chosen so as to satisfy

$$(4.35) \quad \frac{1}{a_0^2} - \frac{\omega(1+K)}{a_0(a_0-1)} > \omega^3 \frac{(b_0-1)}{b_0^2} \cdot \frac{b_0+3}{b_0-1} + \frac{\omega^2}{b_0} \left(\frac{1}{b_0^2} + \frac{2}{b_0} + 1\right).$$

As  $b_0$  must be restricted to satisfy  $b_0 > 1$ , the right-hand side of (4.35), which we denote as  $\sigma(b_0, \omega)$ , is clearly positive. Thus, in order for (4.35) to be satisfied for an arbitrary  $a_0 > 1$ ,  $\omega$  must satisfy

$$(4.36) \quad \omega = \omega_K < \frac{1}{1+K} \left(1 - \frac{1}{a_0}\right) < \frac{1}{1+K}.$$

If we now choose  $\Omega$  so that (4.36) is satisfied and define

$$\tilde{\sigma}(a_0, \omega_K) = \frac{1}{a_0^2} - \frac{\omega_K(1+K)}{a_0(a_0-1)}$$

then (4.35) becomes

$$(4.37) \quad \tilde{\sigma}(a_0, \omega_K) > \sigma(b_0, \omega_K).$$

But

$$(4.38) \quad \lim_{b_0 \rightarrow +\infty} \sigma(b_0, \omega) = 0 \quad (\text{for any } \omega > 0)$$

and thus it is clear that for an arbitrary  $a_0 > 1$  and  $\omega = \omega_K$  defined by (4.36), the inequality (4.35) will be satisfied if  $b_0$  is chosen sufficiently large. We summarize our results in the following lemma.

**LEMMA.** *Consider the holohedral isotropic dielectric material which is defined by the constitutive relations*

$$(4.39a) \quad \mathbf{D}(\mathbf{x}, t) = a_0 \mathbf{E}(\mathbf{x}, t) + \int_0^t e^{-K(t-\tau)} \mathbf{E}(\mathbf{x}, \tau) d\tau,$$

$$(4.39b) \quad \mathbf{H}(\mathbf{x}, t) = b_0 \mathbf{B}(\mathbf{x}, t) + \int_0^t e^{-(t-\tau)} \mathbf{B}(\mathbf{x}, \tau) d\tau,$$

where  $K > 0$  and  $a_0 > 1$  are arbitrary and  $(\mathbf{x}, t) \in \Omega \times [0, 1)$  with  $\Omega \subseteq \mathcal{R}^3$  chosen so that the bedding constant  $\omega$ , defined by the inclusion map of  $H_0^1$  into  $L_2$ , satisfies (4.36). If  $\mathbf{D}(\mathbf{x}, t) = \mathbf{0}$ ,  $(x, t) \in \partial\Omega \times [0, 1)$ , then there exists a constant  $\Gamma > 1$  such that the operator  $\mathbf{M}(t)$ , defined by (2.14b), satisfies the basic hypotheses (4.1) whenever  $b_0 \geq \Gamma$ .

**5. Relation to previous estimates for holohedral isotropic dielectrics.** In the present paper we have considered a special case of nonconducting holohedral isotropic dielectric response under the assumption of zero past history, i.e.,  $\mathbf{E}(\tau) = \mathbf{0}$ ,  $\mathbf{B}(\tau) = \mathbf{0}$ ,  $-\infty < \tau < 0$ ; our constitutive relations were, therefore, of the form (1.16a), (1.16b); using a logarithmic convexity argument we then derived growth estimates for the time evolution of the components of the electric displacement field in a dielectric which conforms to these constitute hypotheses. In a recent work [11] we have derived different estimates for a closely related problem. Namely, we consider in [11] a holohedral isotropic material dielectric of the type (1.15a), (1.15b) with  $a_\nu = 0$ ,  $b_\nu = 0$ ,  $\nu > 1$  but with past history of the form

$$(5.1) \quad \begin{aligned} \mathbf{E}(\mathbf{x}, t) &= \begin{cases} \mathbf{0}, & -\infty < t < -t_h, \\ \mathbf{E}_h(\mathbf{x}, t), & -t_h \leq t < 0, \end{cases} \\ \mathbf{B}(\mathbf{x}, t) &= \begin{cases} \mathbf{0}, & -\infty < t < -t_h, \\ \mathbf{B}_h(\mathbf{x}, t), & -t_h \leq t < 0, \end{cases} \end{aligned}$$

where  $t_h > 0$  is a given positive constant and  $\mathbf{E}_h, \mathbf{B}_h$  satisfy appropriate smoothness assumptions on  $\Omega \times (-t_h, 0)$ . The constitutive hypotheses in [11] then take the form

$$(5.2) \quad \begin{aligned} \mathbf{D}(\mathbf{x}, t) &= a_0 \mathbf{E}(\mathbf{x}, t) + \int_{-t_h}^t \phi(t - \tau) \mathbf{E}(\mathbf{x}, \tau) d\tau, \\ \mathbf{H}(\mathbf{x}, t) &= b_0 \mathbf{B}(\mathbf{x}, t) + \int_{-t_h}^t \psi(t - \tau) \mathbf{B}(\mathbf{x}, \tau) d\tau \end{aligned}$$

on  $\Omega \times (-t_h, T)$ , and, in place of the evolution equations (2.1) considered in the present work, we obtain, under the additional assumption that  $\mathbf{D}_h(\mathbf{x}, -t_h) = \mathbf{0}$ , uniformly on  $\Omega$ , the evolution equations

$$(5.3) \quad \begin{aligned} \frac{\partial^2 D_i}{\partial t^2} + \Psi(0) \frac{\partial D_i}{\partial t} + \dot{\Psi}(0) \left[ D_i - \hat{c}_0 \delta_{ik} \delta_{jl} \frac{\partial^2 D_k}{\partial x_j \partial x_l} \right] \\ + \int_{-t_h}^t \left( \dot{\Psi}(t - \tau) D_i(\tau) - \frac{b_0}{a_0} \Phi(t - \tau) \delta_{ik} \delta_{jl} \frac{\partial^2 D_k(\tau)}{\partial x_j \partial x_l} \right) d\tau = 0 \end{aligned}$$

for  $i = 1, 2, 3$  with  $\hat{c}_0 \equiv b_0/a_0 \dot{\Psi}(0)$ . The same Hilbert space formalism used in the present work when leads in [11] to consideration of abstract initial-history value problems of the form (2.17), (2.18) but with  $\beta(t) \equiv 0$  and with the integral operator defined on  $[-t_h, T)$  instead of  $[0, T)$ . The basic differences, however, between [11] and the present work are as follows: In [11] we consider initial-history value problems corresponding to varying initial displacement fields and varying past histories, i.e.,

$$(5.4) \quad \begin{aligned} \mathbf{u}_t^\alpha + \Gamma \mathbf{u}_t^\alpha - \mathbf{N} \mathbf{u}^\alpha + \int_{-t_h}^t \mathbf{K}(t - \tau) \mathbf{u}^\alpha(\tau) d\tau = \mathbf{0}, \quad 0 \leq t < T, \\ \mathbf{u}^\alpha(0) = \alpha \mathbf{u}_0, \quad \mathbf{u}_t^\alpha(0) = \mathbf{v}_0, \quad \alpha > 0, \quad \mathbf{u}_0, \mathbf{v}_0 \in H_+, \\ \mathbf{u}^\alpha(\tau) = \mathbf{U}(\tau), \quad -t_h \leq \tau < 0, \end{aligned}$$



and

$$\begin{aligned}
 & \mathbf{u}_t^\beta + \Gamma \mathbf{u}_t^\beta - \mathbf{N} \mathbf{u}^\beta + \int_{-t_h}^t \mathbf{K}(t-\tau) \mathbf{u}^\beta(\tau) d\tau = \mathbf{0}, \quad 0 \leq t < T, \\
 (5.5) \quad & \mathbf{u}^\beta(0) = \mathbf{u}_0, \quad \mathbf{u}_t^\beta(0) = \mathbf{v}_0, \quad \beta > 0, \\
 & \mathbf{u}^\beta(\tau) = g(\beta) \mathbf{U}(\tau), \quad -t_h < \tau < 0,
 \end{aligned}$$

where  $g$  is monotonically increasing on  $[0, \infty)$ . The basic aim of the work in [11] is not to derive growth estimates for the time evolution of  $\|\mathbf{u}(t)\|$  but rather to derive lower bounds for  $\sup_{(-t_h, T)} \|\mathbf{u}^\alpha\|_+ + (\sup_{(-t_h, T)} \|\mathbf{u}^\beta\|_+)$  in terms of  $\alpha(\beta)$  and the data of the problem: the conditions (3.2a), (3.2b) in the present work are weakened in [11] to simply

$$(5.6) \quad -\langle \mathbf{v}, \mathbf{K}(0) \mathbf{v} \rangle \geq 0, \quad \forall \mathbf{v} \in H_+$$

and the a priori condition that  $\mathbf{u} \in \mathcal{N}$  (a class of bounded perturbations of the kind prescribed in § 1) is dropped in [11] as logarithmic convexity is not employed to derive the desired estimates. Additional assumptions are made, however, in [11] relative to the data and the integral operator; namely,

$$\begin{aligned}
 & \int_0^\infty \|\mathbf{K}(\tau)\|_{\mathcal{L}_s(H_+, H_-)} d\tau < \infty, \quad \int_0^\infty \|\mathbf{K}_\tau(\tau)\|_{\mathcal{L}_s(H_+, H_-)} d\tau < \infty, \\
 (5.7) \quad & \int_{-t_h}^0 \|\mathbf{U}(\tau)\|_+ d\tau < \infty; \\
 & \langle \mathbf{u}_0, \mathbf{v}_0 \rangle > 0, \quad \langle \mathbf{u}_0, \mathbf{N} \mathbf{u}_0 \rangle > 0 \quad \text{and} \quad \left\langle \mathbf{u}_0 \int_{-t_h}^0 \mathbf{K}(-\tau) \mathbf{U}(\tau) d\tau \right\rangle < 0.
 \end{aligned}$$

For the initial-history value problem (5.4) we then have the following result in [11]: Let  $\mathbf{u}^\alpha$  be a strong solution to (5.4) with

$$(5.8) \quad \|\mathbf{u}_0\|^2 \geq \frac{2}{\Gamma} \langle \mathbf{u}_0, \mathbf{v}_0 \rangle, \quad T > \frac{1}{\Gamma} \ln \left( \frac{2 \langle \mathbf{u}_0, \mathbf{v}_0 \rangle}{2 \langle \mathbf{u}_0, \mathbf{v}_0 \rangle - \Gamma \|\mathbf{u}_0\|^2} \right).$$

Then for each  $\alpha > \|\mathbf{v}_0\| / \langle \mathbf{u}_0, \mathbf{N} \mathbf{u}_0 \rangle^{1/2}$ ,

$$(5.9) \quad \sup_{[-t_h, T]} \|\mathbf{u}^\alpha\|_+ \geq \left[ \frac{|\langle \mathbf{u}_0, \int_{-t_h}^0 \mathbf{K}(-\tau) \mathbf{U}(\tau) d\tau \rangle|}{\omega \Sigma_T} \right]^{1/2} \sqrt{\alpha},$$

where

$$(5.10) \quad \Sigma_T = \frac{1}{2} \|\mathbf{N}\|_{\mathcal{L}_s(H_+, H_-)} + \int_0^\infty \|\mathbf{K}(\tau)\|_{\mathcal{L}_s(H_+, H_-)} d\tau + T \int_0^\infty \|\mathbf{K}_\tau(\tau)\|_{\mathcal{L}_s(H_+, H_-)} d\tau.$$

A similar result follows for the problem (5.5), with varying past history, under analogous assumptions. The basic idea behind the proof of the estimate (5.9) is as follows: Assume that (5.9) is false for some parameter value  $\bar{\alpha} > \|\mathbf{v}_0\| / \langle \mathbf{u}_0, \mathbf{N} \mathbf{u}_0 \rangle^{1/2}$  and show that  $F_{\bar{\alpha}}(t) = \|\mathbf{u}^{\bar{\alpha}}(t)\|^2$  satisfies the differential inequality

$$(5.11) \quad F_{\bar{\alpha}} F_{\bar{\alpha}}'' - (\bar{\alpha} + 1) F_{\bar{\alpha}}'^2 \geq -\Gamma F_{\bar{\alpha}} F_{\bar{\alpha}}', \quad 0 \leq t < T,$$

which, in turn, implies that

$$(5.12) \quad F_{\bar{\alpha}}^{\bar{\alpha}}(t) \geq F_{\bar{\alpha}}^{\bar{\alpha}}(0) [1 - (1 - e^{-\Gamma t}) \bar{\alpha} F_{\bar{\alpha}}'(0) / \Gamma F_{\bar{\alpha}}(0)]^{-1}.$$

The bracketed expression in (5.12) vanishes at

$$(5.13) \quad t_\infty \equiv \frac{1}{\Gamma} \ln \left( \frac{2\langle \mathbf{u}_0, \mathbf{v}_0 \rangle}{2\langle \mathbf{u}_0, \mathbf{v}_0 \rangle - \Gamma \|\mathbf{u}_0\|^2} \right)$$

and  $t_\infty < T$  by virtue of the hypothesis (5.8). Thus  $\sup_{(-t_h, T)} \|\mathbf{u}^{\bar{\alpha}}\| = +\infty$  and via the embedding of  $H_+$  into  $H$  this implies that  $\sup_{(-t_h, T)} \|\mathbf{u}^{\bar{\alpha}}\| = +\infty$  contradicting the assumption that

$$\sup_{(-t_h, T)} \|\mathbf{u}^{\bar{\alpha}}\|_+ \left[ \frac{|\langle \mathbf{u}_0, \int_{-t_h}^0 \mathbf{K}(-\tau) \mathbf{U}(\tau) d\tau \rangle|}{\omega \Sigma_T} \right]^{1/2} \sqrt{\bar{\alpha}}$$

and, thus, establishing (5.9). Estimates of the type (5.9) can be very useful in terms of deriving estimates for physical parameters which enter the definition of the integral operator; in this vein we refer to a recent work [12] on Maxwell–Hopkinson dielectrics where estimates of the type (5.9) have been shown to lead to bounds for constitutive parameters appearing in the memory functions of such materials.

In a more recent work [13] initial-history boundary value problems associated with (5.3) have been reconsidered with a view toward deriving asymptotic lower bounds on the norms of the electric displacement vector when the operators in the equivalent initial-history value problem do not satisfy the requisite coerciveness conditions that imply asymptotic stability [14]. In fact, it is shown, in [13], that solutions  $\mathbf{u} \in \mathcal{N}^*$  of the present abstract initial-history value problem ( $\mathcal{N}^* = \{\mathbf{v} \in C([-t_h, \infty); H_0^1) \mid \sup_{(-t_h, \infty)} \|\mathbf{v}\|_{H_0^1} \leq N\}$  for some  $N > 0$ ) satisfy the differential inequality

$$(5.14) \quad FF'' - \left( \frac{\beta + 1}{2\beta + 1} \right) F'^2 \geq -\Gamma FF', \quad F = \|\mathbf{u}(t)\|_{\mathcal{E}_2}^2,$$

for any  $\beta > 0$ ,  $0 \leq t < \infty$ , provided  $\mathcal{E}(0) = \frac{1}{2} \|\mathbf{v}_0\|_{\mathcal{E}_2}^2 - \langle u_0, \mathbf{N}\mathbf{u}_0 \rangle_{\mathcal{E}_2} < 0$  with  $|\mathcal{E}(0)| > \frac{3}{2} \omega N^2 [\|\mathcal{K}\|_{\mathcal{E}_1[0, \infty)} + \|\hat{\mathcal{K}}\|_{\mathcal{E}_1[0, \infty)}]$  where we assume that (5.6) holds, and in addition, that

$$(5.15) \quad \begin{aligned} \mathcal{K}(t) &\equiv \|\mathbf{K}(t)\|_{\mathcal{E}_s(H_0^1, H^{-1})} \text{ satisfies } \mathcal{K}(\cdot) \in \mathcal{L}_1[0, \infty), \\ \hat{\mathcal{K}}(t) &\equiv \int \|\mathbf{K}_t\|_{\mathcal{E}_s(H_0^1, H^{-1})} d\tau \text{ satisfies } \hat{\mathcal{K}}(\cdot) \in \mathcal{L}_1[0, \infty) \text{ with } \hat{\mathcal{K}}(0) = 0. \end{aligned}$$

The differential inequality (5.14) then yields the estimate

$$(5.16) \quad \lim_{t \rightarrow +\infty} \|\mathbf{u}(t)\|_{L_2}^2 \geq \|\mathbf{u}_0\|_{L_2}^2 \exp \left( \frac{2\langle \mathbf{u}_0, \mathbf{v}_0 \rangle_{L_2}}{\Gamma \|\mathbf{u}_0\|_{L_2}^2} \right)$$

so that  $\lim_{\Gamma \rightarrow +\infty} \lim_{t \rightarrow +\infty} \|\mathbf{u}(t)\|_{L_2}^2 \geq \|\mathbf{u}_0\|_{L_2}^2$ . In fact, the sharper estimate

$$(5.17) \quad \|\mathbf{u}\|_{L_2}^2 \geq \|\mathbf{u}_0\|_{L_2}^2 \left[ 1 + \left( \frac{2(1-\lambda)\langle \mathbf{u}_0, \mathbf{v}_0 \rangle_{L_2}}{\Gamma \|\mathbf{u}_0\|_{L_2}^2} \right) (1 - e^{-\Gamma t}) \right]^{1/(1-\lambda)}$$

is shown to obtain in [14] for all  $t > 0$  and any  $\lambda$ ,  $\frac{1}{2} < \lambda < 1$ . Thus the  $L_2$  norm of  $\mathbf{u}$  is bounded from below as  $t \rightarrow +\infty$  even as the damping becomes arbitrarily large.

REFERENCES

[1] F. BLOOM, *Stability and growth estimates for Volterra integrodifferential equations in Hilbert space*, Bull. Amer. Math. Soc., (1976), pp. 603–606.  
 [2] ———, *On stability in linear viscoelasticity*, Mech. Research Comm., (1976), pp. 143–150.

- [3] ———, *Growth estimates for solutions to initial-boundary value problems in viscoelasticity*, J. Math. Anal. Appl., 59 (1977), pp. 469–487.
- [4] ———, *Continuous data dependence for an abstract Volterra integrodifferential equation in Hilbert space with applications to viscoelasticity*, Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat., (4), 1 (1977), pp. 179–207.
- [5] R. J. KNOPS AND L. E. PAYNE, *Growth estimates for solution of evolutionary equations in Hilbert space with applications in elastodynamics*, Arch. Rational Mech. Anal., 41 (1971), pp. 363–398.
- [6] F. BLOOM, *Stability and growth estimates for electric fields in nonconducting material dielectrics*, J. Math. Anal. Appl., 67 (1979), pp. 296–322.
- [7] J. C. MAXWELL, *A Treatise on Electricity and Magnetism*, 1873 (reprinted by) Dover Press, New York.
- [8] J. HOPKINSON, *The residual charge of the Leyden jar*, Philos. Trans. Roy. Soc. London Ser. A., 167 (1877), pp. 599–626.
- [9] V. VOLTERRA, *Theory of Functionals*, Dover, New York, 1928.
- [10] R. A. TOUPIN AND R. S. RIVLIN, *Linear functional electromagnetic constitutive relations and plane waves in a hemihedral isotropic material*, Arch. Rational Mech. Anal., 6 (1960), pp. 188–197.
- [11] F. BLOOM, *Concavity arguments and growth estimates for damped linear integrodifferential equations with applications to a class of holohedral isotropic dielectrics*, Z. Angew. Math. Phys., 29 (1978), pp. 644–663.
- [12] ———, *Growth estimates for electric displacement fields and bounds for constitutive constants in the Maxwell–Hopkinson theory of dielectrics*, Internat. J. Eng. Sci., 17 (1979), pp. 1–15.
- [13] ———, *Asymptotic bounds for solutions to a system of damped integrodifferential equations of electromagnetic theory*, J. Math. Anal. Applic., to appear.
- [14] C. M. DAFERMOS, *An abstract Volterra equation with applications to linear viscoelasticity*, J. Differential Equations, (1970), pp. 554–569.

## CONTINUOUS DEPENDENCE AND INSTABILITY IN LINEAR THERMOELASTICITY\*

N. S. WILKES†

**Abstract.** We consider the linear theory of thermoelasticity. By means of a modification of the method of logarithmic convexity, certain results are proved concerning the continuous dependence and instability of the system in the case where the elasticity tensor is not positive definite.

**1. Introduction.** This paper is concerned with the theory of small disturbances superposed upon an initially stressed equilibrium state of a thermoelastic solid, that is a linear theory of thermoelasticity. Such a theory was considered first by Green [4] who formulated the equations of motion, and these can be written in the following form, cf. Knops and Wilkes [7],

$$(1.1) \quad (d_{ijkl}u_{k,l})_{,j} + (f_{ij}\theta)_{,j} = \rho\ddot{u}_i,$$

$$(1.2) \quad \dot{\theta} - cf_{ij}\dot{u}_{i,j} = (a_{ij}\theta_{,j})_{,i}$$

where  $u_i$  is the displacement from the equilibrium state and  $\theta$  is the temperature disturbance from the constant equilibrium temperature.

The physical quantities in equations (1.1), (1.2) are the elasticities  $d_{ijkl}$ , the conductivity tensor  $a_{ij}$ , the stress temperature tensor  $f_{ij}$ , the density  $\rho$  and the prescribed constant  $c$ . The elasticities and the conductivity tensor are assumed to satisfy the symmetry conditions  $d_{ijkl} = d_{klij}$  and  $a_{ij} = a_{ji}$ . Here we shall be concerned with various aspects of stability of these equations.

We consider equations (1.1) and (1.2) in an open, bounded, connected region  $\mathcal{B}$  of  $\mathbb{R}^3$ , which is assumed to have a regular boundary  $\partial\mathcal{B}$  and we postulate homogeneous boundary conditions of the form

$$(1.3) \quad u_i = 0 \quad \text{on } \sigma$$

$$(1.4) \quad n_j d_{ijkl}u_{k,l} + n_j f_{ij}\theta = 0 \quad \text{on } \partial\mathcal{B} - \sigma$$

$$(1.5) \quad \theta = 0 \quad \text{on } \Sigma$$

$$(1.6) \quad n_j a_{ij}\theta_{,j} = 0 \quad \text{on } \partial\mathcal{B} - \Sigma$$

where  $\mathbf{n}$  is the outward unit normal to  $\partial\mathcal{B}$ , and where either  $\partial\mathcal{B} - \sigma$  or  $\partial\mathcal{B} - \Sigma$  is assumed to be empty. We further assume that  $\mathbf{u}$ ,  $\dot{\mathbf{u}}$  and  $\theta$  are prescribed in  $\mathcal{B}$  at time  $t = 0$ .

In the study of the stability of the system, a fundamental result is the energy conservation law. It is easy to show that the quantity  $J(t)$  defined below is conserved, i.e.

$$(1.7) \quad \begin{aligned} J(t) &\equiv \int_{\mathcal{B}} \theta^2 dV + 2 \int_0^t \int_{\mathcal{B}} a_{ij}\theta_{,j}(\tau)\theta_{,i}(\tau) dV d\tau + c \int_{\mathcal{B}} \rho\dot{u}_i\dot{u}_i dV + c \int_{\mathcal{B}} d_{ijkl}u_{i,j}u_{k,l} dV \\ &= J(0). \end{aligned}$$

Using this result, together with the assumed positivity of  $c$  and  $a_{ij}$ , Knops and Wilkes [7]

\* Received by the editors March 28, 1979.

† Department of Mathematics, Heriot-Watt University, Riccarton, Currie, Edinburgh. Currently at Engineering Sciences Division, A.E.R.E. Harwell, Oxfordshire, OX11 0RA, England.

prove a simple stability theorem in the case where  $d_{ijkl}$  is positive definite in the sense that

$$(1.8) \quad \int_{\mathcal{B}} d_{ijkl} \xi_{ij} \xi_{kl} dV \cong d_0 \int_{\mathcal{B}} \xi_{ij} \xi_{ij} dV$$

for all  $\xi_{ij}$  and for some  $d_0 > 0$ . They show that if  $J(0)$  is initially small, then  $\int_{\mathcal{B}} \rho u_i u_i dV$  and  $\int_{\mathcal{B}} \theta^2 dV$  remain small for all time. In this case, the asymptotic stability of the system has also been demonstrated by Dafermos [2], [3] and Slemrod and Infante [10]. Results on stability have also been obtained by Brun [1] and Levine [8].

In this paper, we shall consider the case in which  $d_{ijkl}$  is not positive definite in the sense of (1.8). This case has been considered previously by Knops and Wilkes [7] and also by Knops and Payne [5]. Knops and Wilkes showed that in the case where  $d_{ijkl}$  was actually negative definite in the sense that

$$(1.9) \quad \int_{\mathcal{B}} d_{ijkl} \xi_{ij} \xi_{kl} dV \leq -d \int_{\mathcal{B}} \xi_{ij} \xi_{ij} dV$$

for all  $\xi_{ij}$  and some  $d > 0$ ; and when  $f_{ij}$  satisfied a bound of the form

$$(1.10) \quad f_{ij} f_{ij} \leq M_1^2$$

with  $d > cM_1^2$ , the solution had quadratic growth, provided the initial value of  $J(0)$  was negative. Knops and Payne, while not requiring and negative-definiteness of  $d_{ijkl}$ , also needed a bound on the derivative of  $f_{ij}$  of the form

$$(1.11) \quad f_{ij,i} f_{ik,k} \leq M_2^2$$

in order to prove a theorem on continuous dependence. We shall describe their result more fully later.

Here it is our intention to prove, without requiring  $d_{ijkl}$  to be negative-definite, that for certain prescribed initial conditions the solution to the system (1.1), (1.2), has exponential growth. Further continuous dependence results will also be proved which complement those of Knops and Payne.

The method of proof will be a modification of the logarithmic convexity technique, which has been used, for example, in the linear theory of nonthermal elasticity by Knops and Payne [6]. (As a general reference on the logarithmic convexity technique, see also the expository article of Payne [9].)

**2. The logarithmic convexity inequality.** From this point on, we shall assume that the conductivity tensor  $a_{ij}$  is positive definite, in the sense that

$$(2.1) \quad a_{ij} \xi_i \xi_j \geq a_0 \xi_i \xi_i$$

for some  $a_0 > 0$ . We shall also assume that  $c > 0$ , which is analogous to the restriction that the specific heat be positive. We consider the measure  $F(t; b, t_0)$  defined on the solutions to the system (1.1), (1.2) by

$$(2.2) \quad F(t; b, t_0) = \int_{\mathcal{B}} \rho u_i u_i dV + \frac{1}{c} \int_0^t \int_{\mathcal{B}} a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \times \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV + b(t + t_0)^2$$

for positive  $b$  and  $t_0$ , where  $h$  is to be determined later from the initial data. We shall

show that this measure satisfies an inequality similar to a logarithmic convexity inequality.

We first of all proceed to compute the first and second time derivatives of  $F(t; b, t_0)$ . Thus

$$(2.3) \quad \begin{aligned} \dot{F}(t; b, t_0) = & 2 \int_{\mathcal{B}} \rho u_i \dot{u}_i dV + \frac{1}{c} \int_{\mathcal{B}} a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) \\ & \cdot \left( \int_0^t \theta_{,i}(s) ds + h_{,i} \right) dV + 2b(t + t_0), \end{aligned}$$

$$(2.4) \quad \ddot{F}(t; b, t_0) = 2 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV + 2 \int_{\mathcal{B}} \rho u_i \ddot{u}_i dV + \frac{2}{c} \int_{\mathcal{B}} \theta_{,i} a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) dV + 2b.$$

Substituting for  $\rho \ddot{u}_i$  from (1.1) and integrating by parts we can rewrite this as

$$(2.5) \quad \begin{aligned} \ddot{F}(t; b, t_0) = & 2 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV - 2 \int_{\mathcal{B}} d_{ijkl} u_{i,j} u_{k,l} dV - 2 \int_{\mathcal{B}} u_{i,j} f_{ij} \theta dV \\ & - \frac{2}{c} \int_{\mathcal{B}} \theta \left( a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) \right)_{,i} dV + 2b. \end{aligned}$$

We can also substitute from the energy balance law (1.7) to obtain

$$(2.6) \quad \begin{aligned} \ddot{F}(t; b, t_0) = & 4 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV - \frac{2}{c} J(0) + \frac{2}{c} \int_{\mathcal{B}} \theta^2 dV + \frac{4}{c} \int_0^t \int_{\mathcal{B}} a_{ij} \theta_{,j}(\tau) \theta_{,i}(\tau) dV d\tau \\ & - 2 \int_{\mathcal{B}} u_{i,j} f_{ij} \theta dV - \frac{2}{c} \int_{\mathcal{B}} \theta \left( a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) \right)_{,i} dV + 2b. \end{aligned}$$

Integrating equation (1.2) with respect to time, we obtain that

$$(2.7) \quad \theta - cf_{ij}u_{i,j} = \theta(0) - cf_{ij}u_{i,j}(0) + \left( \int_0^t a_{ij} \theta_{,j}(s) ds \right)_{,i}.$$

Hence if we define  $h$  to be a solution of

$$(2.8) \quad (a_{ij}h_{,j})_{,i} = \theta(0) - cf_{ij}u_{i,j}(0)$$

subject to the same boundary conditions as the temperature, that is

$$(2.9) \quad h = 0 \quad \text{on } \Sigma,$$

$$(2.10) \quad n_i a_{ij} h_{,j} = 0 \quad \text{on } \partial\mathcal{B} - \Sigma$$

then  $h$  exists provided  $\Sigma$  has positive measure with respect to  $\partial\mathcal{B}$ , and

$$(2.11) \quad \ddot{F}(t; b, t_0) = 4 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV + \frac{4}{c} \int_{\mathcal{B}} \int_0^t a_{ij} \theta_{,j}(\tau) \theta_{,i}(\tau) d\tau dV - \frac{2}{c} J(0) + 2b.$$

We note that

$$(2.12) \quad \begin{aligned} & \frac{1}{c} \int_{\mathcal{B}} a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^t \theta_{,i}(s) ds + h_{,i} \right) dV - \frac{1}{c} \int_{\mathcal{B}} a_{ij} h_{,j} h_{,i} dV \\ & = \frac{2}{c} \int_{\mathcal{B}} \int_0^t \theta_{,i}(\tau) a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) d\tau dV. \end{aligned}$$

It then follows by Schwarz's inequality that

$$\begin{aligned}
 (2.13) \quad & F(t; b, t_0)\ddot{F}(t; b, t_0) - \left( \dot{F}(t; b, t_0) - \frac{1}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV \right)^2 \\
 & \cong - \left( \frac{2}{c} J(0) + 2b \right) F(t; b, t_0).
 \end{aligned}$$

We note that in the special case in which the initial conditions satisfy

$$(2.14) \quad \theta(0) - cf_{ij}u_{i,j}(0) \equiv 0$$

so that  $h \equiv 0$ , then the inequality (2.13) reduces to the form of logarithmic convexity found in classical elasticity. (See Knops and Payne.)

In the next two sections, we will use inequality (2.13) to obtain results concerning the continuous dependence and instability of solutions to the system (1.1) and (1.2).

**3. Continuous dependence.** Let  $F(t)$  be defined by

$$(3.1) \quad F(t) = F(t; 0, 0).$$

$F(t)$  then satisfies the inequality

$$(3.2) \quad F(t)\ddot{F}(t) - \left( \dot{F}(t) - \frac{1}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV \right)^2 \cong - \frac{2}{c} J(0)F(t).$$

We will consider continuous dependence only on a finite time interval  $[0, T]$ . We will consider two cases: first we shall discuss continuous dependence for solutions with  $J(0) \leq 0$  and then we shall discuss solutions with  $J(0) > 0$ .

Firstly, when  $J(0) \leq 0$ , define  $H_1(t)$  by

$$(3.3) \quad H_1(t) = \log \left\{ F(t) + \frac{T-t}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV \right\}.$$

It follows from (3.2) and (2.11) that

$$(3.4) \quad \ddot{H}_1(t) \cong 0.$$

Hence, by Jensen's inequality

$$(3.5) \quad F(t) + \frac{T-t}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV \cong \left\{ F(0) + \frac{T}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV \right\}^{1-t/T} F(T)^{t/T}.$$

Thus solutions within a class for which  $F(T)$  is bounded are Hölder continuously dependent upon the initial data on compact subintervals of  $[0, T]$ .

Secondly, when  $J(0) > 0$ , define  $H_2(t)$  by

$$(3.6) \quad H_2(t) = \log \left\{ F(t) + \frac{T-t}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV + \frac{1}{c} J(0) \right\} + t^2.$$

Then, again from (3.2) and (2.11), it follows that

$$(3.7) \quad \ddot{H}_2(t) \cong 0$$

and by Jensen's inequality

$$\begin{aligned}
 (3.8) \quad & F(t) + \frac{T-t}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV + \frac{1}{c} J(0) \\
 & \cong e^{t(T-t)} \left\{ F(0) + \frac{T}{c} \int_{\mathcal{B}} a_{ij}h_j h_{,i} dV + \frac{1}{c} J(0) \right\}^{1-t/T} \left\{ F(T) + \frac{1}{c} J(0) \right\}^{t/T}.
 \end{aligned}$$

Thus, again, solutions within a class for which  $F(T)$  is bounded are Hölder continuously dependent upon the initial data on compact subintervals of  $[0, T)$ .

We note, however, that the continuous dependence described by (3.8) is weaker than that of (3.5) as  $J(0)$  must also be small.

In their investigation into the continuous dependence upon the initial data of the system (1.1), (1.2), Knops and Payne [5] further assumed that the tensor  $f_{ij}$  and its derivative satisfied bounds of the form (1.10) and (1.11). By considering the logarithmic convexity of a function  $F^*(t)$  defined by

$$(3.9) \quad F^*(t) = \int_0^t \int_{\mathcal{B}} \rho u_i(\tau) u_i(\tau) dV d\tau + (T-t) \int_{\mathcal{B}} \rho u_i(0) u_i(0) dV + \gamma$$

for some nonnegative constant  $\gamma$ , they deduced that for solutions within a class defined by the inequality

$$(3.10) \quad \int_0^T \int_{\mathcal{B}} \rho u_i(\tau) u_i(\tau) dV d\tau \leq N^2$$

the following inequality was satisfied:

$$(3.11) \quad \int_0^t \int_{\mathcal{B}} \rho u_i(\tau) u_i(\tau) dV d\tau \leq K_1 N^{2\delta} \left\{ K_2 \int_{\mathcal{B}} \rho u_i(0) u_i(0) dV + K_3 \int_{\mathcal{B}} \rho \dot{u}_i(0) \dot{u}_i(0) dV + K_4 \left| \int_{\mathcal{B}} d_{ijk} u_{i,j}(0) u_{k,i}(0) dV \right| + K_5 \int_{\mathcal{B}} \theta(0)^2 dV \right\}^{1-\delta}$$

for computable constants  $K_i$  and where  $\delta$  is given by

$$(3.12) \quad \delta = \frac{1 - \exp(-K_0 t)}{1 - \exp(-K_0 T)}.$$

It can be seen that in some ways this result is stronger than the present result in that the temperature does not need to satisfy any bound at time  $t = T$  and also the quantity  $\int_{\mathcal{B}} a_{ij} h_{,j} h_{,i} dV$  need not be small. We include the results (3.5) and (3.8) to complement the results of Knops and Payne, and for the sake of completeness of the analysis.

**4. Instability.** We study instability directly from inequality (2.13). Suppose  $J(0) < 0$  and let the arbitrary constant  $b$  satisfy

$$(4.1) \quad b = -\frac{J(0)}{c}.$$

It follows that

$$(4.2) \quad F(t; b, t_0) \ddot{F}(t; b, t_0) - \left( \dot{F}(t; b, t_0) - \frac{1}{c} \int_{\mathcal{B}} a_{ij} h_{,j} h_{,i} dV \right)^2 \geq 0$$

and by writing

$$(4.3) \quad L = \frac{2}{c} \int_{\mathcal{B}} a_{ij} h_{,j} h_{,i} dV,$$

$$(4.4) \quad F(t; b, t_0) \ddot{F}(t; b, t_0) - \dot{F}(t; b, t_0)^2 \geq -L \dot{F}(t; b, t_0).$$

As  $F(t; b, t_0)$  never vanishes this is equivalent to

$$(4.5) \quad \frac{d}{dt} \left\{ \frac{\dot{F}(t; b, t_0)}{F(t; b, t_0)} \right\} \geq -\frac{L \dot{F}(t; b, t_0)}{F(t; b, t_0)^2},$$



and this inequality can be immediately integrated to give

$$(4.6) \quad \frac{\dot{F}(t; b, t_0) - L}{F(t; b, t_0)} \geq \frac{\dot{F}(0; b, t_0) - L}{F(0; b, t_0)}.$$

By choosing the arbitrary positive constant  $t_0$  to be large enough, we can satisfy

$$(4.7) \quad \dot{F}(0; b, t_0) > L,$$

and then (4.6) can itself be integrated to give

$$(4.8) \quad F(t; b, t_0) \geq \frac{F(0; b, t_0)\dot{F}(0; b, t_0)}{\dot{F}(0; b, t_0) - L} \exp\left\{\frac{\dot{F}(0; b, t_0) - L}{F(0; b, t_0)}t\right\} - \frac{F(0; b, t_0)L}{\dot{F}(0; b, t_0) - L}.$$

We have thus established that  $F(t) = F(t; 0, 0)$  grows exponentially for large time, providing  $J(0)$  can be chosen to be negative.

Now  $F(t)$  is given by

$$(4.9) \quad F(t) = \int_{\mathcal{B}} \rho u_i u_i dV + \frac{1}{c} \int_{\mathcal{B}} \int_0^t a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV$$

and it is of interest to examine the growth of the two components of  $F(t)$ . We first consider two methods for investigating the growth of the norm  $\int_{\mathcal{B}} \rho u_i u_i dV$  of the displacements.

Firstly, let us consider the function  $G(t)$  defined by

$$(4.10) \quad G(t) = \int_{\mathcal{B}} \rho u_i u_i dV - \frac{1}{c} \int_{\mathcal{B}} \int_0^t a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV + \alpha t.$$

for some positive constant  $\alpha$ .

As before we can compute the first and second derivatives of  $G$  as follows:

$$(4.11) \quad \dot{G}(t) = 2 \int_{\mathcal{B}} \rho u_i \dot{u}_i dV - \frac{1}{c} \int_{\mathcal{B}} a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^t \theta_{,i}(s) ds + h_{,i} \right) dV + \alpha,$$

$$(4.12) \quad \ddot{G}(t) = 2 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV + 2 \int_{\mathcal{B}} \rho u_i \ddot{u}_i dV - \frac{2}{c} \int_{\mathcal{B}} \theta_{,i} a_{ij} \left( \int_0^t \theta_{,j}(s) ds + h_{,j} \right) dV.$$

Substituting from the equations of motion (1.1), (1.2) and also for  $h$  from (2.8), we obtain that

$$(4.13) \quad \ddot{G}(t) = 2 \int_{\mathcal{B}} \rho \dot{u}_i \dot{u}_i dV - 2 \int_{\mathcal{B}} d_{ijkl} u_{i,j} u_{k,l} dV - 4 \int_{\mathcal{B}} f_{ij} u_{i,j} \theta dV + \frac{2}{c} \int_{\mathcal{B}} \theta^2 dV.$$

Hence, in the special case in which the elasticities are negative definite, in the sense of (1.9), and the tensor  $f_{ij}$  is subject to the restrictions (1.10), (1.11), it follows that

$$(4.14) \quad \ddot{G}(t) \geq 0.$$

On integrating twice, we find

$$(4.15) \quad G(t) \geq G(0) + \dot{G}(0)t$$

and by choosing  $\alpha$  appropriately, we may choose  $\dot{G}(0) > 0$ . It is then clear, by adding  $F(t)$  and  $G(t)$ , that in this case  $\int_{\mathcal{B}} \rho u_i u_i dV$  exhibits exponential growth for large time. We note that this is precisely the case in which Knops and Wilkes [5] were able to show quadratic growth.

Secondly, let us additionally assume that  $f_{ij}$  is subject to bounds of the form (1.10) and (1.11). From the equations of motion (1.2) and also (2.8), it is easy to derive the following relation:

$$\begin{aligned}
 (4.16) \quad & \int_{\mathcal{B}} \left( \int_0^t \theta(s) ds + h \right)^2 dV + 2 \int_{\mathcal{B}} \int_0^t a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \\
 & \cdot \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV \\
 & = 2c \int_{\mathcal{B}} \int_0^t f_{ij} u_{i,j} \left( \int_0^\tau \theta(s) ds + h \right) dV + \int_{\mathcal{B}} h^2 dV \\
 & = -2c \int_{\mathcal{B}} \int_0^t u_i \left[ f_{ij} \left( \int_0^\tau \theta(s) ds + h \right) \right]_{,j} dV + \int_{\mathcal{B}} h^2 dV.
 \end{aligned}$$

By using Schwarz inequality and Poincaré inequality (provided  $\Sigma$  is nonempty), we can then obtain an inequality of the form

$$\begin{aligned}
 (4.17) \quad & \int_{\mathcal{B}} \left( \int_0^t \theta(s) ds + h \right)^2 dV + \int_{\mathcal{B}} \int_0^t a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV \\
 & \leq \int_{\mathcal{B}} h^2 dV + K \int_{\mathcal{B}} \int_0^t \rho u_i u_i d\tau dV
 \end{aligned}$$

for some positive computable constant  $K$ . (A similar inequality was used by Knops and Payne [5, equation (3.9b)] in their investigation into continuous dependence).

It follows from (4.17) that in this case

$$(4.18) \quad F(t) \leq \int_{\mathcal{B}} \rho u_i u_i dV + K \int_{\mathcal{B}} \int_0^t \rho u_i(\tau) u_i(\tau) d\tau dV + \int_{\mathcal{B}} h^2 dV,$$

and hence it is easy to show that  $\int_{\mathcal{B}} \int_0^t \rho u_i(\tau) u_i(\tau) d\tau dV$  must grow exponentially for large time.

From this it can further be shown that there exists a sequence  $\{t_n\} \rightarrow \infty$  for which  $t_n \int_{\mathcal{B}} \rho u_i(t_n) u_i(t_n) dV$  is exponentially large and we thus have instability with respect to the weighted  $L_2$ -norm of the displacements.

We next proceed to investigate the growth of the norm of the temperature, that is

$$\int_{\mathcal{B}} \int_0^t a_{ij} \left( \int_0^\tau \theta_{,j}(s) ds + h_{,j} \right) \left( \int_0^\tau \theta_{,i}(s) ds + h_{,i} \right) d\tau dV.$$

Clearly, in general, it will not be possible to show that it has exponential growth, because in the special case in which  $f_{ij} \equiv 0$ , equation (1.2) reduces to the classical heat equation which is known to have bounded solutions.

When  $f_{ij}$  is not identically zero, it remains an open question whether the temperature norm grows exponentially or not. However in certain one dimensional examples, equations (1.1) and (1.2) can be solved explicitly and the temperature norm does indeed grow exponentially.

#### REFERENCES

- [1] L. BRUN, *Méthodes énergétiques dan les systèmes évolutifs linéaires*, J. de Mécanique, 8 (1969).
- [2] C. M. DAFERMOS, *On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity*, Arch. Rational. Mech. Anal., 29 (1968), pp. 241–271.
- [3] ———, *Contraction semigroups and trend to equilibrium in continuum mechanics*, Proc. I.U.T.A.M./I.M.U. conference on applications of functional analysis to mechanics, 1975.

- [4] A. E. GREEN, *Thermoelastic stresses in initially stressed bodies*, Proc. Roy. Soc., A226 (1962), pp. 1–19.
- [5] R. J. KNOPS AND L. E. PAYNE, *On uniqueness and continuous dependence in dynamical problems of linear thermoelasticity*, Internat. J. Solids Structures, 6 (1970), pp. 1173–1184.
- [6] ———, *Growth estimates for solutions of evolutionary equations in Hilbert space with applications in elastodynamics*, Arch Rational. Mech. Anal., 41 (1971), pp. 363–398.
- [7] R. J. KNOPS AND E. W. WILKES, *Theory of elastic stability*, Handbuch der Physik, C. Truesdell, ed., Vol. VIa/3. Springer-Verlag, Berlin, 1973.
- [8] H. A. LEVINE, *On a theorem of KNOPS and PAYNE in dynamical linear thermoelasticity*, Arch. Rational. Mech. Anal., 38 (1970).
- [9] L. E. PAYNE, *Logarithmic convexity and related techniques applied to problems in continuum mechanics*.
- [10] M. SLEMROD AND E. F. INFANTE, *An Invariance Principle for Dynamical Systems on Banach Space: Application to the General Problem of Thermoelastic Stability*. IUTAM Symposium on Instability of Continuous Systems—Herrenalb., Springer-Verlag, New York, 1969.

## ASYMPTOTIC EXPANSIONS OF INTEGRALS WITH OSCILLATORY KERNELS AND LOGARITHMIC SINGULARITIES\*

JUDITH A. ARMSTRONG† AND NORMAN BLEISTEIN†

**Abstract.** This paper is a follow-up to an earlier paper by Bleistein which derived asymptotic expansions of integral transforms of functions with logarithmic singularities. That result dealt with exponentially decaying kernels. In this paper the results are expanded to include the case of general oscillatory kernels—e. g., Fourier or Hankel transforms. The results also include kernels such as Airy functions of negative argument and oscillatory Weber functions.

**1. Introduction.** We shall develop the asymptotic expansion of a class of integrals of the form

$$(1.1) \quad I(\lambda) = \int_0^T h(\lambda t)f(t) dt, \quad \lambda \rightarrow \infty.$$

For this class we shall assume that  $h(t)$  is an “oscillatory” kernel; that is,

$$(1.2) \quad h(t) \sim \exp\{i\omega^v\} \sum_{m=0}^{\infty} \sum_{n=0}^{N(m)} \alpha_{mn} t^{-r_m} (\log t)^n, \quad t \rightarrow \infty.$$

Here,  $\text{Re } r_m \uparrow \infty$  and  $N(m)$  is finite for each  $m$ ,  $\omega$  and  $v$  real,  $\omega \neq 0$ . We assume that  $h$  and  $f$  are infinitely differentiable on  $(0, T)$ . Furthermore,  $f(t)$  is assumed to vanish “ $C^\infty$  smoothly” at  $T < 1$ .

Thus the integral  $I(\lambda)$  is one which might arise from a more general integral by applying the appropriate van der Corput (1948) “neutralizer” to isolate the critical point at the origin. The class of integrals is further distinguished by the nature of  $f(t)$  near the origin, namely

$$(1.3) \quad f(t) \sim \sum_{m=0}^{\infty} \sum_{n=0}^{N(m)} c_{mn} t^{\alpha_m} (\log t)^{\beta_{mn}}, \quad t \rightarrow 0^+.$$

Here,  $\text{Re } \alpha_m \uparrow \infty$ ,  $N(m)$  is finite for each  $m$ , and the  $\beta_{mn}$ ’s are any complex numbers. Furthermore, we assume that the asymptotic expansion of any derivative of  $f$  is obtained by differentiating (1.3).

This work is a continuation of an earlier paper by one of the authors, Bleistein (1977), in which  $h(t)$  was instead an “exponential” kernel— $i\omega$  in (1.2) replaced by a negative real number. Unfortunately, the method of proof of that paper does not suffice here. The relevant literature for both classes of integrals is cited in that earlier paper and will not be repeated here. We do remark, however, that, in comparison to the earlier literature, the distinguishing feature in both classes of integrals is that the coefficients  $\beta_{mn}$  may be something other than nonnegative integers.

In the interim between these two papers, Wong (1977) and Wong and Lin (1978) have derived results for the Fourier transform and the Hankel transform. These are special cases of the results given here. In these two papers the nature of the particular kernel is exploited and factors of the form  $t^{\alpha_m}$  occurring in  $f(t)$  are viewed as multipliers of the kernel in order to derive the asymptotic result. The method of proof here is more general and allows statement of the asymptotic result for arbitrary oscillatory kernels.

\* Received by the editors July 20, 1977, and in final revised form May 29, 1979.

† Department of Mathematics, University of Denver, Denver, Colorado 80208. This research was supported in part by the Office of Naval Research under Contract # N00014-76-C.0039.

The reader unfamiliar with asymptotic techniques will find these results useful when dealing with such common kernels as Weber functions or Airy functions of negative argument.

To carry out the analysis below, we shall further assume that  $h(t)$  is locally integrable on  $(0, \infty)$  and

$$(1.4) \quad h(t) = O(t^{-a}), \quad t \rightarrow 0^+, \quad a < \operatorname{Re} r_0, \quad \operatorname{Re} \alpha_0 - a > -1.$$

**2. Technique of integration.** We shall calculate the asymptotic expansion of  $I(\lambda)$  by the Mellin transform technique. (See Bleistein and Handelsman (1975), Chaps. 4–7.) To do so we define

$$(2.1) \quad M[h(t); z] = \int_0^\infty t^{z-1} h(t) dt, \quad z = x + iy,$$

$$(2.2) \quad M[f(t); 1-z] = \int_0^\infty t^{-z} f(t) dt, \quad z = x + iy$$

and use the Mellin–Parseval theorem to write

$$(2.3) \quad I(\lambda) = \frac{1}{2\pi i} \cdot \int_{c-i\infty}^{c+i\infty} \lambda^{-z} M[h; z] M[f; 1-z] dz.$$

We shall now quote results about this integral. They are proven in the above cited references and in the papers by Handelsman and Lew listed in the references.

(i)  $M[h; z]$  exists and is analytic for  $a < x < \operatorname{Re} r_0$ .

(ii)  $M[h; z]$  may be analytically continued as a holomorphic function to the right half plane  $\operatorname{Re} r_0 < x$ , however,

$$(2.4) \quad M[h; z] = O(|y|^{(x-\operatorname{Re} r_0)v-\frac{1}{2}}), \quad |y| \rightarrow \infty, \quad x \text{ fixed;}^1$$

that is, its rate of growth on vertical lines increase with  $x$ .

(iii)  $M[f; 1-z]$  is analytic for  $x < \operatorname{Re} \alpha_0 + 1$

(iv) For the Bromwich contour in (2.3)

$$(2.5) \quad a < c < \operatorname{Re} \alpha_0 + 1.$$

The asymptotic expansion of  $I(\lambda)$  is generated by replacing the Bromwich contour by a sum of loop integrals around singularities of the analytic continuation of  $M[f; 1-z]$  plus a vertical contour further to the right. The asymptotic expansion arises from the loop integrals while the integral on the vertical contour is explicitly of lower order in  $\lambda$  than the original integral. To allow for the deformation of contour, we must impose conditions on  $f(t)$  which will insure sufficient decay of its Mellin transform, thereby compensating for the growth of  $M[h; z]$ .

We now state Theorem 1 which concerns  $M[f; 1-z]$ .

**THEOREM 1.** *Suppose  $f(t)$  locally integrable on  $(0, 1)$  with an expansion*

$$f(t) \sim \sum_{m=0}^{\infty} \sum_{n=0}^{N(m)} c_{mn} t^{\alpha_m} (\log t)^{\beta_{mn}}, \quad t \rightarrow 0,$$

where  $\operatorname{Re} \alpha_m \uparrow \infty$  and  $N(m)$  is finite for each  $m$ .

<sup>1</sup> In Bleistein and Handelsman (1975), the result with “greatest integer less than”  $(x-r_0)/v$  is proven in Chapter 4 and (2.4) is outlined in the exercises in Chapter 7 as a straightforward application of the method of steepest descents.

Then

- (i)  $M[f; 1 - z]$  is analytic for  $x < \text{Re } \alpha_0 + 1$ .
- (ii) The analytic continuation of  $M[f; 1 - z]$  to the right takes the form

$$(2.6) \quad M[f; 1 - z] = \sum_{\text{Re } (\alpha_m - \alpha_0) < k} \left\{ \sum_{n=0}^{N(m)} \frac{c_{mn} e^{i\pi\beta_{mn}} \Gamma(\beta_{mn} + 1)}{(\alpha_m + 1 - z)^{\beta_{mn} + 1}} + \sum_{\beta_{mn} = -l} \frac{c_{mn} (\alpha_m + 1 - z)^{l-1}}{(l-1)!} \log(z - \alpha_m - 1) \right\} + M_k(z).$$

Here, in  $\Sigma^*$ , we exclude the terms with  $\beta_{mn}$  a negative integer, while, in  $\Sigma'$ , we include only terms with  $\beta_{mn}$  a negative integer. The function  $M_k(z)$  is analytic for  $x < \text{Re } \alpha_0 + k + 1$  and the result is correct for any  $k$ .

- (iii)  $M[f; 1 - z] = O(y^{-k})$ , any  $k$ , as  $y \rightarrow \infty$ .

The form of (2.6) exhibits the singularities of  $M[f; 1 - z]$  in  $x < \text{Re } \alpha_0 + k + 1$  but suggests growth of  $M$  on vertical lines. In fact  $M_k(z)$  compensates for this growth to make (iii) true but  $M_k(z)$  has no singularities in  $x < \text{Re } \alpha_0 + k + 1$ .

Results (i) and (ii) were proved in Bleistein (1977). The proof of (iii) is given in the Appendix.

From Theorem 1, we see negative integer powers of  $\beta_{mn}$ , lead to logarithmic branch points, nonnegative integer powers lead to poles, and all other  $\beta_{mn}$  lead to algebraic branch points.

The principal part in the expansion of  $M[f; 1 - z]$  about such singularities takes the following form:

Case 1:  $\beta_{mn} = l \geq 0$ .

$$t^{\alpha_m} (\log t)^{\beta_{mn}} \rightarrow -\frac{1}{(z - \alpha_m - 1)} l + 1.$$

Case 2:  $\beta_{mn} = l < 0$ .

$$t^{\alpha_m} (\log t)^{\beta_{mn}} \rightarrow \frac{(z - \alpha_m - 1)^{l-1}}{(l-1)!} \log(z - \alpha_m - 1).$$

Case 3:  $\beta_{mn}$  not an integer.

$$t^{\alpha_m} (\log t)^{\alpha_{mn}} \rightarrow \frac{e^{i\pi\beta_{mn}} \Gamma(\beta_{mn} + 1)}{(\alpha_m + 1 - z)^{\beta_{mn} + 1}}.$$

**3. Main result.** We can now state the main result about the asymptotic expansion of  $I(\lambda)$ .

**THEOREM 2.** Let  $I(\lambda)$ , given by (1.1) be an absolutely convergent integral with  $f$  and  $h$  locally integrable on  $(0, \infty)$  and  $h$  satisfying (1.2) and (1.4) with  $\text{Re } \alpha_0 - a > -1$ , then with  $f$  neutralized about 0 and  $f \equiv 0$  for  $t \geq T$ ,  $I(\lambda)$  has the expansion

$$(3.1) \quad I(\lambda) = \sum_{\text{Re } (\alpha_m - \alpha_0) < k} \sum_{n=0}^{N(m)} c_{mn} \Gamma(\beta_{mn} + 1) J(\alpha_m, \beta_{mn}, \lambda) + \sum_{\text{Re } (\alpha_m - \alpha_0) < k} \sum_{n=0}^{N(m)} \frac{c_{mn}}{(l-1)!} K(\alpha_m, l, \lambda) + O(\lambda^{-\alpha_0 - 1 - k + \epsilon}), \quad \text{any } \epsilon > 0.$$

Here, for each choice of  $n$ ,  $\Sigma^*$  indicates those  $mn$ 's for which  $\beta_{mn}$  is not a negative integer, while  $\Sigma'$  includes exactly those  $mn$ 's for which  $\beta_{mn} = -l$ , a negative integer. Also, the functions  $J$  and  $K$  are defined respectively by (3.2) and (3.6).

As we shall see, the functions  $J$  and  $K$  are related to an ordered asymptotic sequence with increasing  $\operatorname{Re} \alpha_m$ . Their definition is fairly complicated but their asymptotic expansions are more straightforward.

Recall  $I(\lambda)$  is given by (2.3) repeated here:

$$I(\lambda) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \lambda^{-z} M[h; z] M[f; 1-z] dz.$$

Also recall the earlier comment on  $M[h; z]$  stated after (2.3).

We need to consider  $I(\lambda)$  with  $M[f; 1-z]$  taking the form (2.6) with the estimate (iii) below that equation.

Thus, we may deform the Bromwich contour to the right an arbitrary but finite distance, so long as we include loop contours around the singularities of (2.6).

To study the contributions from the singularities in  $\Sigma^*$ , we define

$$(3.2) \quad J(\alpha, \beta, \lambda) = \frac{e^{i\pi\beta}}{2\pi i} \oint_{\epsilon} (\alpha + 1 - z)^{-\beta-1} \lambda^{-z} M[h; z] dz,$$

where the contour is taken around the singularity at  $z = \alpha + 1$ . Here we envision retaining only the relevant principal part of  $M[f; 1-z]$  with respect to a given singular point, or each such point in  $\Sigma^*$  in (2.6). We must look at two separate cases.

*Case 1:*  $\beta = l$ , a nonnegative integer. In this case,  $J(\alpha, l, \lambda)$  is given as a residue of the integrand at  $\alpha + 1$ . So

$$(3.3) \quad J(\alpha, l, \lambda) = \frac{1}{l!} \left( \frac{d}{dz} \right)^l \{ \lambda^{-z} M[h; z] \}_{z=\alpha+1}.$$

*Case 2:*  $\beta \neq l$ . The result is

$$(3.4) \quad J(\alpha, \beta, \lambda) \sim \sum_{j=0}^{\infty} \frac{C_j e^{i\pi\beta} (\log \lambda)^{\beta-j}}{\lambda^{\alpha+1} \Gamma(1+\beta-j)}$$

with

$$(3.5) \quad C_j = \frac{M^{(j)}[h; \alpha+1]}{j!}.$$

This is derived by Watson's lemma for loop integrals (see Bleistein and Handelsman, (1975), p. 162) with large parameter  $\log \lambda$ .

We must also deal with integrands arising from the second sum in (3.1). We define

$$(3.6) \quad K(\alpha, l, \lambda) = \frac{1}{2\pi i} \int_{\epsilon} \lambda^{-z} (\alpha + 1 - z)^{l-1} \log(z - \alpha - 1) M[h; z] dz.$$

The result is

$$(3.7) \quad K(\alpha, l, \lambda) \sim \lambda^{-\alpha-1} \sum_{j=0}^{\infty} C_j \binom{l+j}{j} (\log \lambda)^{-l-j},$$

with  $C_j$  determined by

$$(3.8) \quad z^{l-1} M[h; z + \alpha + 1] = \sum_{j=0}^{\infty} \frac{C_j}{j!} \cdot z^{j+l-1}.$$

This is derived in the same manner as (3.5). For explicit details on the above contour integrals we refer the reader to the earlier paper by the second author.

Now, by referring back to (3.1), let us comment on the nature of the asymptotic sequence. As the contour is “moved” to the right, the singularities  $\alpha_m + 1$  are encountered beginning with  $\alpha_0 + 1$ , continuing with increasing  $m$ . Consider the possibilities for the contributions from  $\alpha_0 + 1$ . If  $\beta_{on}$  is a nonnegative integer for all  $n$ , then we obtain a finite expansion in powers of  $\log \lambda$  for each  $\beta_{on}$  from (3.3). We would then proceed to  $\alpha_1 + 1$ . However, if  $\beta_{on}$  is other than a nonnegative integer, for some  $n$ , we obtain from (3.4) an infinite expansion in powers of  $\log \lambda$  at  $\alpha_0 + 1$ , and that  $\beta_{on}$ . In this case, it makes no sense to proceed to  $\alpha_1 + 1$ , since it is already of lower algebraic order in  $\lambda$  and hence, asymptotically zero with respect to the sequence

$$\{\lambda^{-\alpha_0-1}(\log \lambda)^{-\beta_{on}-j}\}.$$

If  $\beta_{mn}$  are all nonnegative integers, we have the case

$$(3.9) \quad I(\lambda) \sim \sum_{\text{Re}(\alpha_m - \alpha_0) < k} \sum_{n=0}^{N(m)} c_{mn} \left(-\frac{d}{dz}\right)^n \{\lambda^{-z} M[h; z]\}_{z=\alpha_m+1}.$$

Only in this very special case does one obtain contributions from each singularity as the contour moves to the right. If one  $\beta_{mn}$  is not a nonnegative integer, the loop integral around  $\alpha_m + 1$  has an infinite expansion in powers of  $\log \lambda$  of the form  $J$  or  $K$ .

We shall close this section with examples. We consider the integral

$$(3.10) \quad I(\lambda) = \int_0^1 h(t) |\ln t|^{3/2} t dt.$$

Here  $f(t)$  is a single term of the form (1.3) with

$$(3.11) \quad \alpha_0 = 1, \quad \beta_{00} = \frac{3}{2}, \quad c_{00} = e^{3\pi i/2},$$

the last being chosen so that

$$(3.12) \quad c_{00}(\ln t)^{3/2} = |\ln t|^{3/2}$$

is real and positive for  $0 < t < 1$ .

Our asymptotic expansion will be of the form (3.1) with only terms of the form  $J(\alpha_0, \beta_{00}, \lambda)$  since  $\beta_{00} = \frac{3}{2}$ . We have

$$(3.13) \quad I(\lambda) \sim c_{00} \Gamma(\beta_{00} + 1) J(\alpha_0, \beta_{00}, \lambda).$$

A two-term expansion of  $I$  will be

$$(3.14) \quad I(\lambda) \sim -i\Gamma\left(\frac{5}{2}\right) \left[ C_0 \frac{-i(\log \lambda)^{3/2}}{\lambda^2 \Gamma\left(\frac{5}{2}\right)} + C_1 \frac{-i(\log \lambda)^{1/2}}{\lambda^2 \Gamma\left(\frac{3}{2}\right)} \right]$$

with

$$(3.15) \quad C_j = \frac{M^{(j)}[h; \alpha_0 + 1]}{j!}.$$

In (3.14) a two-term expansion is used in order to see the imaginary part. Note also that the relative error is  $O\{(\log \lambda)^{-2}\}$ . For

$$(3.16) \quad h(t) = e^{it}, \quad \alpha_0 = 1,$$

we obtain

$$(3.17) \quad C_0 = -1, \quad C_1 = -(1 - \gamma) - i(\pi/2)$$

where  $\gamma = .57721$ , the Euler–Mascheroni constant (see Erdélyi (1954)). This agrees



with the result which would be obtained by Wong and Lin (1978). For

$$(3.18) \quad h(t) = J_\nu(t), \quad \alpha_0 = 1,$$

we obtain<sup>2</sup>

$$(3.19) \quad C_0 = \frac{2\Gamma\left(\frac{2+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \quad C_1 = \frac{1}{2} \frac{d}{dz} \left[ \frac{2^{z-1}\Gamma\left(\frac{z+\nu}{2}\right)}{\left(\frac{\nu-z+2}{2}\right)} \right]_{z=2}$$

which would agree with Wong. Using (3.14), we can also derive the result for

$$(3.20) \quad h(t) = \text{Ai}(-t),$$

where<sup>3</sup>

$$(3.21) \quad C_0 = \frac{3^{1/6}}{2\pi} \Gamma\left(\frac{2}{3}\right),$$

$$C_1 = \frac{1}{2} \frac{d}{dz} \left[ \frac{3((2z/3) - (7/6))}{\pi} \Gamma\left(\frac{z}{3}\right) \Gamma\left(\frac{z+1}{3}\right) \sin\left(\frac{\pi z}{3} + \frac{\pi}{6}\right) \right]_{z=2}$$

and for

$$(3.22) \quad h(t) = D_\nu(e^{i(\pi/4)t}),$$

$$(3.23) \quad C_0 = \frac{-i\sqrt{\pi}2^{(3+\nu)/2}}{\Gamma\left(\frac{3-\nu}{2}\right)} F\left(\frac{3}{2}; \frac{1-\nu}{2}; \frac{3-\nu}{2}; -1\right)$$

and

$$(3.24) \quad C_1 = \frac{-i}{2} \frac{d}{dz} \left[ \frac{\sqrt{\pi}\Gamma(z)2^{(z+1-\nu)/2}}{\Gamma\left(\frac{z+1-\nu}{2}\right)} F\left(\frac{z+1}{2}; \frac{1+\nu}{2}; \frac{z+1-\mu}{2}; -1\right) \right]_{z=2}$$

We return to the case with  $h(t)$  given by (3.16). Inserting  $C_0$  and  $C_1$  into (3.14) gives the result

$$(3.25) \quad I(\lambda) \sim -\frac{(\log \lambda)^{3/2}}{\lambda^2} + \frac{3}{2}(.42279 + 1.57079i) \frac{(\log \lambda)^{1/2}}{\lambda^2}.$$

A numerical integration of (3.10) was carried out using Simpson's rule. In Table 1 we tabulate  $I(\lambda)$  for  $\lambda = 10, 50, 100$  and compare it with the results for the real and imaginary parts of  $I(\lambda)$  obtained by numerical integration. We tabulate  $\log \lambda$  as well, because it is the "large" parameter in the asymptotic result. Note that the results are surprisingly good for  $\log \lambda$  as small as 3.912. We also include  $(\log \lambda)^{-2}$  to give an indication of percentage error to be expected from a two term expansion with leading order  $(\log \lambda)^{3/2}$  and error term  $O\{(\log \lambda)^{-1/2}\}$ .

<sup>2</sup> See Erdélyi (1954).

<sup>3</sup> See Appendix in Bleistein and Handelsman (1975).

TABLE 1

$\lambda$	10	50	100
$\log \lambda$	2.3025	3.9120	4.6051
$(\log \lambda)^{-2}$	.189	.065	.047
ASYMPTOTIC RESULT			
Real Part	-.0252	-.002593	-.00085216
Imaginary Part	.03575	.00186	.00050562
VIA SIMPSON'S RULE			
Real Part	-.01647	-.002497	-.000825
Imaginary Part	.03199	.00183	.000496
RELATIVE ERRORS			
Real Part	34%	3.7%	3.1%
Imaginary Part	10.5%	1.6%	1.9%

**Appendix.** Here, a proof of Theorem 1 (iii), as provided by the referee, will be presented.

Define

$$(A.1) \quad S_k(t) = \sum_{\text{Re}(\alpha_m - \alpha_0) < k} \sum_{n=0}^{N(m)} c_{mn} t^{\alpha_m} (\log t)^{\beta_{mn}} (1 - t^k)^L$$

for  $0 < t < 1$ , and  $S_k(t) \equiv 0$  for  $t \geq 1$ . Set

$$f_k(t) = f(t) - S_k(t).$$

Then, or any sufficiently small  $\varepsilon > 0$ , we have

$$(A.2) \quad f_k(t) = O(t^{\alpha_0 + k - \varepsilon}) \quad \text{as } t \rightarrow 0^+.$$

In (A.1) we choose  $L$  to be so large that

$$(A.3) \quad \left. \frac{d^j}{dt^j} [(\log t)^{\beta_{mn}} (1 - t^k)^L] \right|_{t=1} = 0$$

for all  $\beta_{mn}$  in  $S_k(t)$  and  $j = 0, 1, \dots, k - 1$ . The factor  $(1 - t^k)$  is introduced to assure that  $f(t)$  and  $S_k(t)$  have the same asymptotic expansion to order  $\text{Re } \alpha_0 + k - \varepsilon$ ; see (A.2). From (A.2) we have

$$(A.4) \quad M[f_k; 1 - z] = O(|y|^{-k})$$

for all  $x < 1 + \text{Re } \alpha_0 + k - \varepsilon$ . This follows from Lemma 2 in [Handelsman and Bleistein (1973)]. Here we have used the assumption on the differentiability of the expansion (1.3).

Now we consider  $M[S_k; 1 - z]$ . A typical term in the sum takes the form

$$(A.5) \quad I(z) = \int_0^1 t^{\alpha - z} (\log t)^\beta (1 - t^k)^L dt.$$

Integrating by parts  $k$  times gives

$$(A.6) \quad I(z) = \left[ \prod_{n=0}^k (\alpha - z + n) \right]^{-1} \int_0^1 t^{\alpha-z+k} \left[ \frac{d}{dt} \right]^k \{(\log t)^\beta (1-t^k)^L\} dt.$$

All the boundary terms vanish in view of (A.3).

The representation in (A.6) is an analytic continuation of  $I(z)$  from  $x < \operatorname{Re} \alpha + 1$  to  $x < \operatorname{Re} \alpha + k + 1$ . From (A.6) it follows that, for  $x < \operatorname{Re} \alpha + k + 1$  and  $|y| \rightarrow \infty$ ,

$$(A.7) \quad I(z) = O(|y|)^{-k}.$$

Since there are only a finite number of terms of the form  $I(z)$  in  $M[S_k; 1-z]$ , we also have

$$(A.8) \quad M[S_k; 1-z] = O(|y|)^{-k},$$

for  $x < \operatorname{Re} \alpha_0 + k + 1$  and  $|y| \rightarrow \infty$ . Coupling the results (A.4) and (A.8), we have

$$M[f; 1-z] = M[S_k; 1-z] + M[f_k; 1-z] = O(|y|)^{-k}$$

for  $x < \operatorname{Re} \alpha_0 + k + 1$  and  $|y| \rightarrow \infty$ .

Note that the “ $k$ ” in the proof is not the same as given in (2.6), and that it can be taken sufficient large to compensate the growth of  $M[h; z]$ .

#### REFERENCES

- N. BLEISTEIN AND R. A. HANDELSMAN (1975), *Asymptotic Expansions of Integrals*, Holt, Rinehart, and Winston, New York.
- N. BLEISTEIN (1977), *Asymptotic expansions of integral transforms of functions with logarithmic singularities*, this Journal, 8, pp. 655–672.
- J. G. VANDER CORPUT (1948), *On the method of critical points*, Proc. Anst. Akad. Weten., 51, pp. 405–433.
- A. ERDÉLYI (1954), *Table of Integral Transforms*, McGraw-Hill, New York.
- R. A. HANDELSMAN AND N. BLEISTEIN (1973), *Asymptotic expansions of integral transforms with oscillatory kernels: A generalization of the method of stationary phase*, this Journal, 4, pp. 519–535.
- R. A. HANDELSMAN AND J. S. LEW (1970), *Asymptotic expansion of Laplace transforms near the origin*, this Journal, 1, pp. 118–130.
- (1971), *Asymptotic expansions of a class of integral transforms with algebraically dominated kernels*, J. Math. Anal. Appl., 35, pp. 405–433.
- R. WONG (1977), *Asymptotic expansions of Hankel transforms of functions with logarithmic singularities*, J. Comput. Appl. Math. 3, pp. 271–286.
- R. WONG AND J. F. LIN (1978), *Asymptotic expansions of Fourier transforms of functions with logarithmic singularities*, J. Math. Anal. Appl., 64, pp. 173–180.

## SYNTHESIS OF POLYNOMIC SYSTEMS\*

WILLIAM A. PORTER†

**Abstract.** With  $H$  a Hilbert space and  $\{(x_i, y_i): i = 1, \dots, m\} \subset H \times H$  a basic problem is to determine the existence and uniqueness of causal functions,  $f$ , on  $H$  satisfying  $y_i = fx_i, i = 1, \dots, m$ . The present paper considers classes of polynomic functions which minimize an operator norm. The results include explicit necessary and sufficient conditions and an explicit synthesis procedure for realizing the resultant polynomic functions.

**1. Introduction.** The identification and/or representation of a ‘black-box’ phenomena from external measurements is a mathematical problem of contemporary interest. Several lines of development have a rich associated literature. Automatic control theorists, for example, have been pursuing the identification problem for linear dynamic systems (see the survey [1]).

In the nonlinear setting the representation of a black-box by polynomic or multilinear models has had a recent resurgence of interest (see the survey [2]). Such an approach is essential where the input-output behavior is nonlinear even for small signals. A recent article by Palm and Poggio [3] has underscored the importance of polynomic modeling, i.e., Volterra–Wiener expansions, in the biological system domain.

The problem treated here can be construed as an identification problem. We start with a collection of observed input-output pairs  $\{(u_i, y_i): i = 1, \dots, m\}$ . From these a polynomic map,  $\Phi$ , is constructed such that

$$(1) \quad y_i = \Phi(u_i), \quad i = 1, \dots, m.$$

If the pairs are derived from experimental observation then the map  $\Phi$  is obviously a representation of the black-box, valid over existing data. However, because the procedure is constructive and is valid independent of the source of the pairs, it is also a synthesis tool for polynomic maps.

For perspective and to sharpen this introduction we turn first to a review of relevant existing results.

**2. Recent results.** In reviewing the results of [4], [5], [6] it is convenient to focus specifically on the Hilbert space  $L_2(\nu)$  over the finite or infinite interval,  $\nu$ , and equipped with the usual inner product. We shall need also the orthoprojector family  $\{P^t: t \in \nu\}$  given by

$$(2) \quad (P^t x)(\beta) = \begin{cases} x(\beta), & \beta \leq t, \\ 0, & \beta > t. \end{cases}$$

A function  $f: L_2(\nu) \rightarrow L_2(\nu)$  is *causal* provided  $P^t f = P^t f P^t$ , all  $t \in \nu$ .

The set  $\{(u_i, y_i): i = 1, \dots, m\}$  is said to be *linearly well posed* if it satisfies the condition

$$(3) \quad \sum_{i=1}^m \alpha_i (P^t u_i) = 0 \Rightarrow \sum_{i=1}^m \alpha_i (P^t y_i) = 0, \quad \text{all } t \in \nu.$$

In [4] it is shown that linearly well-posed sets admit a causal linear map,  $\Phi$ , satisfying (1).

---

\* Received by the editors March 20, 1978, and in final revised form June 25, 1979. This work was supported in part by the Air Force Office for Scientific Research under Grant 78-3500.

† Department of Electrical Engineering, Louisiana State University, Baton Rouge, Louisiana 70803.

In addition one such map is explicitly constructed. In [5] the linear solution is shown to have a state realization embodied in a family of  $m$  linear differential equations.

When the set  $\{(u_i, y_i): i = 1, \dots, m\}$  is not linearly well-posed then [4] provides a causal polynomic map satisfying the input-output constraints. This polynomic map is of order  $m - 1$  (we shall clarify ‘order’ later) and is by no means unique.

A related development, [6], considers the polynomic approximation of continuous functions on  $L_2(\nu)$ . In particular if  $K \subset L_2(\nu)$  is an arbitrary compact set and if,  $f$ , is a continuous causal function on  $L_2(\nu)$  then [6] shows the existence of a causal polynomic map,  $\hat{f}$ , on  $L_2(\nu)$  such that

$$\sup_{u \in K} \|f(u) - \hat{f}(u)\| < \varepsilon$$

holds for arbitrary  $\varepsilon > 0$ . Moreover  $\hat{f}$  has a state variable realization which is linear in state behavior and polynomic in its state to output map.

In the present paper we supplement the results cited above in the following fashion. First a norm is specified on the class of polynomic operators. This norm is minimized with respect to the constraints of (1). As in the earlier studies an explicit test for existence of minimal  $\Phi$  and subsequent explicit synthesis procedures result.

**3. Polynomic functions.**<sup>1</sup> In this study we consider first operators,  $\Phi$ , on  $L_2(\nu)$  of the form

$$\begin{aligned} (\Phi u)(t) = & \phi_0(t) + \int_{\nu} \phi_1(t, \alpha_1) + \iint_{\nu} \phi_2(t, \alpha_1, \alpha_2)u(\alpha_1)u(\alpha_2) + \dots \\ (4) \quad & + \int_{\nu} \dots \int_{\nu} \phi_n(t, \alpha_1 \dots \alpha_n)u(\alpha_1) \dots u(\alpha_n), \quad t \in \nu, \end{aligned}$$

where the differentials  $d\alpha_1, \dots, d\alpha_n$  have been suppressed for simplicity. Without loss of generality the kernels  $\phi_j(t, \dots)$  are assumed symmetric in the  $\alpha_1$  variables. The map  $P_j$  computed by

$$(P_j u)(\cdot) = \int \dots \int_{\nu} \phi_j(\cdot, \alpha_1 \dots \alpha_j)u(\alpha_1) \dots u(\alpha_j)$$

is said to be a  $j$  power map. The map  $\phi$  of (4) is said to be polynomic of order  $n$ . We note that  $\Phi$  is not causal and will make suitable modifications later.

We shall assume that each kernel satisfies a Hilbert-Schmidt type condition namely that

$$(5) \quad \|P_j\|^2 = \int_{\nu} \int_{\nu} |\phi_j(t, \alpha_1 \dots \alpha_j)|^2 dt d\alpha_1 \dots d\alpha_j = M_j^2 < \infty.$$

and compute a norm on  $\Phi$  by

$$(6) \quad \|\Phi\|^2 = \sum_{j=0}^n \|P_n\|^2.$$

In our study we shall keep  $n$  finite. The limiting case as  $n \rightarrow \infty$  requires a modified  $\|\Phi\|$ . We note that such modifications have been extensively studied by Dwyer [7], and others under the heading of Fock spaces.

<sup>1</sup> During review, the recent work of L. Zyla and R. J. P. DeFigueredo were brought to the author’s attention [8]. In [8] interpolating spline theory is exploited in a Fock space setting to study noncausal analytic interpolations. Having available the present manuscript, reference [8] has been extended to include causality properties [9].

To solve the constrained minimization problems the Lagrange multiplier method will be utilized. In short the functional

$$(7) \quad J(\Phi) = \|\Phi\|^2 + \sum_{i=1}^m \langle \lambda_i, y_i - \Phi(u_i) \rangle$$

where  $\langle \cdot, \cdot \rangle$  denote the  $L_2(\nu)$  inner product, is formed. The kernels  $\phi_i$  used to form  $\Phi$  are varied independently. The first variation condition,  $\delta J = 0$ , is examined. Since a clear pattern will emerge from these steps it is sufficient to consider the case  $n = 2$  in detail.

For  $n = 2$  we have

$$\|\Phi\|^2 = \int |\phi_0(t)|^2 + \iint |\phi_1(t, \alpha)|^2 + \iiint |\phi_2(t, \alpha, \beta)|^2$$

and

$$\begin{aligned} \langle \lambda_i, y_i - \Phi(u_i) \rangle = & \int \lambda_i(t) \left[ y_i(t) - \phi_0(t) - \int \phi_1(t, \alpha) u_i(\alpha) + \cdots \right. \\ & \left. + \iiint \phi_2(t, \alpha, \beta) u_i(\alpha) u_i(\beta) \right] dt \end{aligned}$$

where  $d\alpha_i$  have been suppressed and all integrals are over  $\nu$ . Using elementary manipulations it follows that

$$(8) \quad \begin{aligned} \delta \|\Phi\|^2 = 2 \int \left\{ \phi_0(t) \delta \phi_0(t) + \int \phi_1(t, \alpha) \delta \phi_1(t, \alpha) \right. \\ \left. + \iiint \phi_2(t, \alpha, \beta) \delta \phi_2(t, \alpha, \beta) \right\} dt. \end{aligned}$$

By similar computation we have

$$(9) \quad \begin{aligned} \delta \langle \lambda_i, y_i - \Phi(u_i) \rangle = - \int \lambda_i(t) \left[ \delta \phi_0(t) + \int u_i(\alpha) \delta \phi_1(t, \alpha) \right. \\ \left. + \iiint u_i(\alpha) u_i(\beta) \delta \phi_2(t, \alpha, \beta) \right] dt. \end{aligned}$$

In view of (7), (8) the condition  $\delta J = 0$  then yields the equations

$$(10) \quad \begin{aligned} 2\phi_0(t) - \sum_{i=1}^m \lambda_i(t) &= 0, \\ 2\phi_1(t, \alpha) - \sum_{i=1}^m \lambda_i(t) u_i(\alpha) &= 0, \\ 2\phi_2(t, \alpha, \beta) - \sum_{i=1}^m \lambda_i(t) u_i(\alpha) u_i(\beta) &= 0. \end{aligned}$$

The original constraint set remains,

$$(11) \quad y_i(t) = \phi_0(t) + \int \phi_1(t, \alpha) u_i(\alpha) + \iiint \phi_2(t, \alpha, \beta) u_i(\alpha) u_i(\beta), \quad i = 1, \cdots, m.$$

Substituting (9) into (10) in the obvious fashion produces

$$(12) \quad \begin{aligned} 2y_i(t) &= \sum_{j=1}^m \lambda_j(t) + \sum_{j=1}^m \int u_j(\alpha) u_i(\alpha) \lambda_j(t) \\ &+ \sum_{j=1}^m \iint u_j(\alpha) u_j(\beta) u_i(\alpha) u_i(\beta) \lambda_j(t), \quad i = 1, \dots, m. \end{aligned}$$

To simplify notation we introduce the definition

$$(13) \quad \mu_{ij}(m, 2) = 1 + \langle u_i, u_j \rangle + \langle u_i, u_j \rangle^2$$

and form the  $m \times m$  symmetric matrix

$$(14) \quad V(m, 2) = [\mu_{ij}(m, 2)].$$

Letting  $y(t) = \text{col}(y_1(t), \dots, y_m(t))$  and  $\lambda(t) = \text{col}(\lambda_1(t), \dots, \lambda_m(t))$  the vector form of (11) is apparently

$$(15) \quad 2y(t) = V(m, 2)\lambda(t).$$

Assuming invertibility for the moment we have an explicit solution for  $\lambda(t)$ .

Returning now to (9) let us define the  $m$ -tuplets  $\pi_j$  by

$$(16) \quad \begin{aligned} \pi_0 &= \text{row}(1, 1, \dots, 1), \\ \pi_1(\alpha) &= \text{row}(u_1(\alpha), u_2(\alpha), \dots, u_m(\alpha)), \\ \pi_2(\alpha, \beta) &= \text{row}(u_1(\alpha)u_1(\beta), \dots, u_m(\alpha)u_m(\beta)). \end{aligned}$$

In view of (9) and (14) we have

$$(17) \quad \begin{aligned} \phi_0(t) &= \pi_0 V(m, 2)^{-1}y(t), \\ \phi_1(t, \alpha) &= \pi_1(\alpha) V(m, 2)^{-1}y(t), \\ \phi_2(t, \alpha, \beta) &= \pi_2(\alpha, \beta) V(m, 2)^{-1}y(t), \end{aligned}$$

as the explicit construction of  $\Phi$ .

Several observations are available which refine and sharpen our result. With regard to the functions  $\mu_{ij}(m, 2)$  of (12) we note that one term  $1 = \langle u_i, u_j \rangle^0$  is directly attributable to the assumption of a  $\phi_0 \neq 0$  in (4); similarly the terms  $\langle u_i, u_j \rangle$  and  $\langle u_i, u_j \rangle^2$  follow from the assumptions  $\phi_1 \neq 0$  and  $\phi_2 \neq 0$  respectively.

To generalize then we consider the class of all power maps  $P_j$  satisfying (5) for finite  $j$ . Let  $\mathbb{N}$  be any finite subset of the integers. Let  $\phi$  be any map of the form

$$\Phi_{\mathbb{N}} = \sum_{i \in \mathbb{N}} P_i.$$

It is easily shown that the variational method utilized above needs only trivial adjustments. One adjustment is that the definition of  $\mu_{ij}(m, 2)$  is replaced by

$$\mu_{ij}(m, \mathbb{N}) = \sum_{k \in \mathbb{N}} \langle u_i, u_j \rangle^k.$$

The matrix  $V(m, \mathbb{N})$  replaces  $V(m, 2)$  with the result that (14) remains valid. The tuplets  $\pi_j(\alpha_1, \dots, \alpha_j)$ ,  $j > 2$  are defined as the obvious extension of the pattern evidenced in (15). The kernels  $\phi_j = \pi_j V(m, \mathbb{N})^{-1}y$  for  $j \in \mathbb{N}$  are well defined and meaningful in the context of synthesizing the requisite  $\Phi_{\mathbb{N}}$ .

*Example 1.* For the linear case we take  $\mathbb{N} = \{1\}$ . It is apparent that  $V(m, \{1\})$  is the Grammian matrix of the set  $\{u_1, \dots, u_m\}$ . Hence  $V$  is invertible if and only if this set is

linearly independent. The linear map constructed, namely

$$P_1 u = y(t)^* V(m, \{1\})^{-1} \int \pi_1^*(\alpha) u(\alpha) \phi \alpha,$$

is easily seen to have the requisite input-output properties; moreover null space  $(P_1) = \text{span} \{u_1, \dots, u_m\}^\perp$ . We noted earlier that (5) was a Hilbert-Schmidt assumption and hence it is not surprising that our solution has, in the above sense, a maximum null space.

*Example 2.* The affine solution is also easily reviewed and for this we take  $\mathfrak{N} = \{0, 1\}$ . The matrix  $V(m, \{0, 1\})$  is the Grammian of  $\{u_1, \dots, u_m\}$  with 1 added to each entry. For example with  $m = 2$  it is easily verified that

$$\det V(m, \{0, 1\}) = \|u_1 - u_2\|^2 + \|u_1\|^2 \|u_2\|^2 - |\langle u_1, u_2 \rangle|^2.$$

Clearly then the condition  $u_1 \neq u_2$ , rather than linear independence, suffices for a solution. In general an affine solution will exist provided  $\text{rank} \{u_1, \dots, u_m\} \geq m - 1$ .

To summarize our results we present the following theorem.

**THEOREM 1.** *A minimal map  $\Phi_{\mathfrak{N}}$  for  $J(\cdot)$  exists if and only if  $y(t) \in \text{Range } V(m, \mathfrak{N})$  a.e.  $t \in \nu$ . If  $V(m, \mathfrak{N})$  is nonsingular the solution is unique and given by*

$$\Phi_{\mathfrak{N}} \sim \sum_{j \in \mathfrak{N}} \pi_j(\cdot, \cdot, \cdot) V(m, \mathfrak{N})^{-1} y(\cdot).$$

*Proof.* It suffices to note that if  $y(t) \in \text{Range } V(m, \mathfrak{N})$  a.e. then any left inverse of  $V(m, \mathfrak{N})$  solves (14) for  $\lambda$ . Using (9) this suffices to construct the solution kernels.

**4. Causality.** In the preceding sections the maps in question were not causal. Many applications, however, require a synthesis procedure which guarantees a causal solution. The causality requirement, which was incorporated in [4], [5], [6], can be added to the present development without great difficulty.

The simplest change in the development of § 3 is to add the assumption

$$\phi_j(t, \alpha_1, \dots, \alpha_j) = 0, \quad \text{any } \alpha_k > t, \quad j = 1, \dots, n.$$

We shall also assume that variations  $\delta\phi_j$  are taken only over kernels with the same property.

The variational method of § 3 proceeds to (7), (8) without difficulty. Invoking the causality requirement we have

$$\begin{aligned} \delta J(\Phi) = \int_{\nu} dt & \left\{ \left( 2\phi_0(t) - \sum_i \lambda_i(t) \right) \delta\phi_0(t) + \int_0^t [2\phi_1(t, \alpha) - \sum \lambda_j(t) u_j(\alpha)] \delta\phi_1(t, \alpha) d\alpha \right. \\ & \left. + \int_0^t \int [2\phi_2(t, \alpha, \beta) - \sum \lambda_j(t) u_j(\alpha) u_j(\beta)] \delta\phi_2(t, \alpha, \beta) d\alpha d\beta + \dots \right\}. \end{aligned}$$

Using obvious arguments the condition  $\delta J(\Phi) = 0$  now yields the modified equation set

$$\begin{aligned} (10') \quad & 2\phi_0(t) - \sum_{i=1}^m \lambda_i(t) = 0, \\ & 2\phi_1(t, \alpha) - \sum_{i=1}^m \lambda_i(t) (P^t u_i)(\alpha) = 0, \\ & 2\phi_2(t, \alpha, \beta) - \sum_{i=1}^m \lambda_i(t) (P^t u_i)(\alpha) (P^t u_i)(\beta) = 0, \end{aligned}$$

where  $P^t$  is defined as in (2).



The original development proceeds as before with the substitutions  $P^t u_i$  for  $u_i, i = 1, \dots, m$ . To summarize we have the modified definition

$$(18) \quad \mu_{ij}(m, \mathfrak{N}, t) = \sum_{k \in \mathfrak{N}} \langle u_i, P^k u_j \rangle^k, \quad i, j = i, \dots, m,$$

which yields the  $m \times m$  matrix

$$(19) \quad V(m, \mathfrak{N}, t) = [\mu_{ij}(m, \mathfrak{N}, t)].$$

The  $m$ -tuples  $\pi_j$  take the modified form

$$(20) \quad \begin{aligned} \pi_0 &= \text{row}(1, 1, \dots, 1), \\ \pi_1(t, \alpha) &= \text{row}((P^t u_1)(\alpha), \dots, (P^t u_m)(\alpha)), \\ \pi_2(t, \alpha, \beta) &= \text{row}((P^t u_1)(\alpha)(P^t u_1)(\beta), \dots, (P^t u_m)(\alpha)(P^t u_m)(\beta)) \end{aligned}$$

and finally when  $V$  is invertible

$$(21) \quad \begin{aligned} \phi_0(t) &= \pi_0 V(m, \mathfrak{N}, t)^{-1} y(t), \\ \phi_1(t, \alpha) &= \pi_1(t, \alpha) V(m, \mathfrak{N}, t)^{-1} y(t), \\ \phi_2(t, \alpha, \beta) &= \pi_2(t, \alpha, \beta) V(m, \mathfrak{N}, t)^{-1} y(t), \\ &\vdots \end{aligned}$$

We note that the kernels  $\phi_j$  inherit the causal property  $\phi_j(t, \alpha_1, \dots, \alpha_j) = 0$  any  $\alpha_i > t$  from the  $\pi_i$ .

Our development leads to the following modification of Theorem 1.

**THEOREM 2.** *A minimal causal map  $\Phi_{\mathfrak{N}}$  exists for  $J(\cdot)$  if and only if  $y(t) \in \text{Range } V(m, \mathfrak{N}, t)$  for a.e.  $t \in \nu$ . When  $V(m, \mathfrak{N}, t)$  is nonsingular,*

$$\Phi_{\mathfrak{N}} \sim \sum_{j \in \mathfrak{N}} \pi_j(t, \cdot, \cdot, \cdot) V(m, \mathfrak{N}, t)^{-1} y(t).$$

The causal solution developed above raises at least two interesting questions not inherent in the noncausal case. The first issue stems from the fact that  $\langle u_i P^t u_j \rangle \rightarrow 0$  as  $t \rightarrow 0$ . This implies

$$\lim_{t \rightarrow 0} V(m, \mathfrak{N}, t) = 0.$$

In some cases the singular behavior is called for by the problem formulation. For instance if  $y_i(0) \neq 0$  and  $u_i$  continuous and finite as  $t \rightarrow 0$  then satisfying  $\Phi(u_i) = y_i$  requires a singular behavior at  $t = 0$ .

In the case where the  $(u_i, y_i)$  pairs are measurements from a physical low pass system it is to be expected that  $y_i(0) = 0$  even if  $u_i(0) \neq 0$ . In some such cases the singular behavior is not necessarily present in the kernels. Other aspects of the singularity question are considered in some detail in [4] and we will not dwell further on this here.

**5. The rank of  $V(m, \mathfrak{N}, t)$ .** In Theorem 2 we see that the range of the matrix  $V(m, \mathfrak{N}, t)$  as  $t$  varies is of obvious interest. To explore this situation note that the dimension of the parameterized manifold span  $\{P^t u_1, \dots, P^t u_m\}$  is nondecreasing for increasing  $t$ . Thus the Grammian matrix of this set has nondecreasing rank with increasing  $t$ . This Grammian matrix is  $V(m, \{1\}, t)$ , as indicated in Example 1. The results of § 4 in fact include the results of [4] as the case  $\mathfrak{N} = \{1\}$ .

Consider now the discrete scale of Hilbert spaces

$$H(\mathfrak{N}) = \bigotimes_{j \in \mathfrak{N}} L_2(\nu)^{(j)}.$$

For example with  $\mathfrak{N} = \{0, 1, 3, 4\}$ ,  $H(\mathfrak{N})$  consists of tuples

$$\underline{x} = (x_0, x_1(t), x_3(t, \alpha, \beta), x_4(t, \alpha, \beta, \gamma)),$$

where each entry is square integrable with respect to the appropriate product measure. We equip  $H(\mathfrak{N})$  with the natural inner product namely

$$\langle \underline{x}, \underline{y} \rangle_{\mathfrak{N}} = \sum_{j \in \mathfrak{N}} \langle x_j, y_j \rangle_j,$$

where  $\langle \cdot, \cdot \rangle_j$  is the inner product on  $L_2(\nu)^j$ .

Now for each  $x \in L_2(\nu)$  we associate the tuple  $\hat{x} \in H(\mathfrak{N})$  by

$$\hat{x} \sim (1, x(\alpha_1), x(\alpha_1)x(\alpha_2), x(\alpha_1)x(\alpha_2)x(\alpha_3), \dots),$$

where we leave out components of  $\hat{x}$  not indexed by  $\mathfrak{N}$ . The tuple  $P^t[\hat{x}]$  is taken to be  $[P^t \hat{x}]$ . It is then a matter of direct inspection to verify

PROPOSITION 1.  $V(m, \mathfrak{N}, t)$  is the Grammian matrix of  $\{P^t \hat{u}_1, \dots, P^t \hat{u}_m\}$ .

Several other results now come into view. First we note that a set  $\{\underline{x}, \underline{y}, \dots, \underline{z}\} \subset H(\mathfrak{N})$  is linearly dependent if and only if each component  $\{x_j, y_j, \dots, z_j\} \subset L_2(\nu)^j$  is linearly dependent (and with at least one set of common scalars). Thus  $\text{rank } V(m, \{1\}, t) \leq \text{rank } V(m, \mathfrak{N}, t)$  whenever  $1 \in \mathfrak{N}$ . For example  $\{1, t, t+1\}$  has linear span dimension 2 whereas the tuples (here  $\mathfrak{N} = \{0, 1\}$ ),  $\{(1, 1), (1, t), (1, t+1)\}$  are linearly independent in  $R \otimes L_2$ . More generally the following is true.

PROPOSITION 2.

- (a) If  $\mathfrak{N}' \subset \mathfrak{N}$  then  $\text{rank } V(m, \mathfrak{N}', t) \leq \text{rank } V(m, \mathfrak{N}, t)$ .
- (b) If  $t' \leq t$  then  $\text{rank } V(m, \mathfrak{N}, t') \leq \text{rank } V(m, \mathfrak{N}, t)$ .

Part (b) follows from the nesting property  $\text{range}(P^t) \subseteq \text{range}(P^\beta)$  all  $t \leq \beta$ , of the projection family.

Following [4] we shall say that a set  $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \subset L_2(\nu) \times L_2(\nu)$  is well posed over  $\mathfrak{N}$  if and only if the set  $\{(\hat{x}_1, y_1), (\hat{x}_2, y_2), \dots, (\hat{x}_m, y_m)\}$  is linearly well posed (see (3)) in  $H(\mathfrak{N}) \times L_2(\nu)$ . An easy modification of the results of [4] leads to

PROPOSITION 3.  $y(t) \in \text{range } V(m, \mathfrak{N}, t)$  for a.e.  $t \in \nu$  if and only if

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \text{ is well posed over } \mathfrak{N}.$$

We have indicated earlier that it is relatively difficult for a set  $\{\hat{x}_1, \dots, \hat{x}_m\}$  to be linearly dependent. As a further example consider scalars  $k_1, k_2, \dots, k_m$  and the functions  $x_i = k_i x_0, i = 1, \dots, m$ . When  $\mathfrak{N} = \{0, 1, \dots, l\}$  where  $l \geq m - 1$  it can be shown easily that the set  $\{\hat{x}_1, \dots, \hat{x}_m\}$  is linearly dependent if and only if the scalars  $k_i$  are not all distinct.

**6. Summary.** The earlier results [4] on polynomial synthesis provided a causal polynomial map,  $\Phi$  of order  $m - 1$  satisfying  $\phi(x_i) = y_i, i = 1, \dots, m$ . The present study provides necessary and sufficient conditions for synthesis of entire classes of causal polynomial maps  $\{\Phi(\mathfrak{N}): \mathfrak{N} \subset I\}$  satisfying the same input-output constraints. Moreover, the present synthesis procedure results in maps which minimize an operator norm.

REFERENCES

[1] R. K. MEHRA, *Optimal input signals for parameter estimation in systems—Survey and new results*, IEEE Trans. Automatic Control, AC-19 (1974), pp.753-768.  
 [2] W. A. PORTER, *An overview of polynomial system theory*, IEEE Proceeding Special Issue on System Theory, January 1976, pp. 18-23.

- [3] G. PALM AND T. POGGIO, *The Volterra representation and the Wiener expansion: Validity and pitfalls*, SIAM J. Appl. Math., 33 (1977), pp. 195–216.
- [4] W. A. PORTER, *Data interpolation, causality structure and system identification*, Information and Control, 29 (1975), pp. 217–233.
- [5] ———, *Causal realization from input–output pairs*, J. Control Optimization, 15 (1977), pp. 120–128.
- [6] ———, *Approximation by Bernstein systems*, Math. System Theory, 11 (1977), pp. 259–274.
- [7] T. A. DWYER III, *Holomorphic representations of tempered distributions and weighted Fock spaces*, Analyse Fonctionnell et Applications, Actualité's, Sci. et Industr., 1367, Herman, Paris, 1975.
- [8] L. V. ZYLA, *A Theory of nonlinear system approximation and identifications based on Volterra expansions*, Ph.D. thesis, Rice University, Houston, TX, May 1977.
- [9] R. J. P. DE FIGUEIREDO AND L. ZYLA, *Nonlinear system identification based on a Fock space framework*, E.E. Tech. Rep. 79, February 1979 (revised April 1979), Rice University, Houston, TX.

## A QUALITATIVE STUDY OF THE STEADY-STATE SOLUTIONS FOR A CONTINUOUS FLOW STIRRED TANK CHEMICAL REACTOR\*

M. GOLUBITSKY† AND B. L. KEYFITZ‡

**Abstract.** An approach to the bifurcation of steady-state equilibria using singularity theory is applied to the problem of multiple equilibria in a continuous flow stirred tank chemical reactor where the flow rate is the bifurcation parameter. Under the assumption of a single first-order exothermic chemical reaction, all the qualitatively different bifurcation diagrams which occur locally are found. They form the universal unfolding of the singular bifurcation problem  $x^3 + \lambda^2 = 0$ .

**Introduction.** It is well-known to chemical engineers that a complex reacting system can exhibit multiple equilibria which may differ dramatically from each other as to the extent of the reaction, the equilibrium temperature, and other phenomena. Analysis of this sort of problem is complicated by the fact that the equations are highly nonlinear, and contain many parameters, or control variables, which affect the configuration of the equilibria. This paper is an attempt to bring a new method to bear on such problems by the application of singularity theory to a chemical reactor problem. Singularity theory is a nonlinear theory which provides a framework for a qualitative analysis of many-parameter problems via the notions of contact equivalence, in terms of which "qualitatively similar" behavior can be precisely defined, and a universal unfolding, by means of which essential parameters can be identified. When a particular universal unfolding can be found for a complex problem, it may then be regarded as a perturbation of a simpler problem with the parameters varied about a particular choice. We feel that this technique, of building up a complete description of the solution from the behavior near this particular choice, or "organizing center" of the problem, may be widely applicable in those chemical engineering and combustion problems where a diversity of multiple steady-state phenomena makes any global analysis very difficult. The possibility of providing such a description was suggested by some work of Uppal, Ray and Poore [6], [7], on a continuous flow stirred tank reactor model in which an analysis of the steady-state behavior is a prerequisite for an understanding of the dynamic behavior of the model. Uppal, Ray and Poore were unable to prove that their analysis was complete, but provided some partial results supplemented by numerical experiments. Using singularity theory, we have been able to show that they did indeed identify all the qualitatively different types of equilibrium behavior of the system, and that the same classification also applies to a generalized system in which the standard temperature dependence of the reaction is replaced by a function with similar properties. To be precise, Uppal, Ray and Poore consider a single-step chemical reaction with Arrhenius-type kinetics, that is a reaction rate term of the form  $\exp(-E/RT)$ . For a class of reaction rate terms which includes a  $C^3$ -open neighborhood of the Arrhenius terms, we show that the structure of solutions is the same. In § 1, we describe the model used by Uppal, Ray and Poore and its generalization.

For physical reasons it is often convenient to analyze the steady states of a system by examining the dependence of these states on a distinguished parameter which is

---

\* Received by the editors October 10, 1978.

† Department of Mathematics, Arizona State University, Tempe, AZ 85281. The work of this author was supported in part by the National Science Foundation under Grant MCS 77-03655 and by the Institute for Advanced Study.

‡ Department of Mathematics, Arizona State University, Tempe, AZ 85281. The work of this author was supported in part by the National Science Foundation under Grant MCS 77-04164.

varied “quasi-statically”—i.e., the system is supposed always to remain in equilibrium. Of particular interest are the parameter values where the number of equilibria changes (bifurcation of equilibria)—hence the term “bifurcation parameter” which will be used to describe this variable throughout the paper. Although the approach to the reactor and similar systems as bifurcation problems is natural, classical bifurcation theory (for example [3]) has generally not considered such problems because there is no “trivial solution” about which to look for bifurcation points. Instead, we have the familiar *S*-curves of combustion theory. The recent approach of Golubitsky and Schaeffer [4] to bifurcation problems via singularity theory extends and specializes the theorems and techniques of singularity theory to steady-state bifurcation problems, and it is this theory that we apply to the reactor problem. Specifically the theorems of singularity theory are adapted to include the bifurcation parameter indicated above as a distinguished control variable. A brief description of the theory and an analysis of the singularities that appear in this problem are given in § 2. The “organizing center” for the problem turns out to be a singularity we have named the *winged cusp*: it corresponds to a particular, physically reasonable, choice of control variables. This singularity is of codimension three: that is, three independent controls must be varied in the neighborhood of the organizing center to yield all the qualitatively different types of bifurcation diagrams. These perturbed bifurcation diagrams are also listed in § 2. In § 3 we verify that the winged cusp singularity is present in this problem, and that the physical parameters do indeed provide a complete set of perturbations (or “unfolding parameters”) not only near the organizing center but everywhere in control space.

We are grateful to Rutherford Aris for pointing out this problem to us, and would like to thank David Schaeffer for many helpful conversations. Articles by Ray [5] and Aris [1], where an attempt was made to adapt the catastrophe theory cusp, by the addition of a wing, to explain the results of Uppal, Ray and Poore, served as a guide for our intuition. Elementary catastrophe theory now seems an inappropriate theory for the analysis of this model, although the type of mathematics ultimately used is identical in spirit to that of elementary catastrophe theory. Needless to say, our name for the organizing center of this problem, the winged cusp, was motivated by the papers of Ray and Aris.

**1. A mathematical model for a continuous flow stirred tank chemical reactor.** In this section we derive an equation to describe the steady-state temperature and concentration for a first-order, single-step, exothermic, irreversible, volume-preserving chemical reaction which takes place in a continuously stirred tank with in- and out-flow, and heat loss to the surroundings. If a reactant,  $\mathcal{R}$ , is converted to a product,  $\mathcal{P}$ , in the reaction, the assumption that the tank is stirred permits the concentration of  $\mathcal{R}$ ,  $c$ , and the temperature inside the tank,  $T$ , to be described as functions of time,  $t'$ , alone, while the heat-loss rate is modeled by a term of the form  $-hS(T - T_0)$ , where  $T_0$  is the ambient temperature and  $h$  is a heat-transfer coefficient which depends on the thermal conductivity of the mixture and of the walls, and  $S$  is the heat-transfer area (surface area of the container). In the reaction,  $\mathcal{R}$  is converted to  $\mathcal{P}$  at a rate  $k(T)c$ , where  $k(T)$  is the temperature-dependent reaction rate. For chemical reactor problems, in which radiation is usually ignored,  $k(T)$  is assumed to have a temperature dependence of the Arrhenius form,

$$(1.1) \quad k(T) = Z e^{-(E/RT)},$$

where  $Z$  is a frequency factor and  $E$  is called the *activation energy* of the reaction. The constant  $R$  is the Boltzmann constant. The heat release of such a reaction is  $(-\Delta H)k(T)$ , where  $\Delta H$ , the heat of reaction, is negative for an exothermic reaction.

Finally, if reactant with concentration  $c_f$  and temperature  $T_f$  are fed into the tank at a flow rate  $F$ , and the mixture of reactant and product removed at the same rate, the equations governing the time-evolution of  $T$  and  $c$  are

$$(1.2) \quad \begin{aligned} V \frac{dc}{dt'} &= F(c_f - c) - Vk(T)c, \\ V\rho C_\rho \frac{dT}{dt'} &= \rho C_\rho F(T_f - T) + V(-\Delta H)k(T)c - hS(T - T_0), \end{aligned}$$

where  $V$  is the volume of the container and  $\rho$  and  $C_\rho$  are the density and specific heat of the mixture (assumed constant). This standard system is discussed in [6], [1].

The following scalings are also conventionally used to develop nondimensionalized equations. Concentration and temperature are scaled by feeder concentration and temperature so that

$$(1.3) \quad x = \frac{c_f - c}{c_f}$$

measures the extent of conversion of  $\mathcal{R}$  to  $\mathcal{P}$ , and

$$(1.4) \quad y = \frac{T - T_f}{T_f}$$

is the rise above entrance temperature. Note that  $y > -1$ . Time is conveniently scaled by the heat-transfer rate,

$$(1.5) \quad t = \frac{hS}{V\rho C_\rho} t'.$$

Then (1.2) is replaced by

$$(1.6) \quad \begin{aligned} \frac{dx}{dt} &= -\varepsilon x + D(1-x)A(y) = f_1(x, y), \\ \frac{dy}{dt} &= -(1+\varepsilon)y + BD(1-x)A(y) + \eta = f_2(x, y), \end{aligned}$$

where now the essential parameters appearing are

$$(1.7) \quad \varepsilon = \frac{F\rho C_\rho}{hS} = \frac{1}{\theta},$$

which can be identified as a flow-rate based on the time-scale (1.5) (its reciprocal,  $\theta$ , is called the *residence time*, and is often in the literature taken as the fundamental flow-rate parameter),

$$(1.8) \quad D = \frac{k(T_f)V\rho C_\rho}{hS},$$

a Damköhler number relating the chemical heat-gain rate at  $T_f$  to the heat-loss rate,

$$(1.9) \quad B = \frac{(-\Delta H)C_f}{\rho C_\rho T_f},$$

which is proportional to the exothermicity, and also measures the “adiabatic temperature rise” which would occur if the reaction proceeded to completion in the absence of heat-loss or flow in the reactor, and

$$(1.10) \quad \eta = \frac{T_0 - T_f}{T_f},$$

the ambient temperature scaled by (1.4).

The function

$$(1.11) \quad A(y) = \frac{k(T_f y + T_f)}{k(T_f)}$$

is the temperature-dependent reaction rate, scaled by the rate at  $T_f$ . For an Arrhenius temperature dependence,

$$(1.12) \quad A(y) = \exp\left(\frac{\gamma y}{1 + y}\right),$$

where

$$(1.13) \quad \gamma = \frac{E}{RT_f}$$

is a scaled activation energy. For a truly temperature-dependent reaction,  $\gamma$  cannot be too small, and, in fact, in many applications to ignition problems,  $y$  is further scaled by  $\bar{y} = \gamma y$  and the approximation  $\gamma = \infty$  is used. Alternatively,  $A(y)$  is often approximated by a low-degree polynomial for the range of  $y$  known to occur in some particular problem. These approximations are introduced to make computations simpler, and will, in general, change the qualitative properties of solutions of (1.6) outside the range in which they are valid. In the present paper, we will not insist that  $A(y)$  be an Arrhenius term, but we will, in § 3, impose on  $A(y)$  a set of conditions, satisfied by all Arrhenius terms with  $\gamma > 8/3$ , which will guarantee a certain qualitative behavior for steady-state solutions of (1.6).

The system (1.6) has the property that multiple steady states, that is solutions to  $f_1 = f_2 = 0$ , can exist for certain values of the parameters  $\varepsilon$ ,  $D$ ,  $B$  and  $\eta$ . In this paper, we shall classify these steady states by means of the bifurcation diagrams which occur when  $D$ ,  $B$  and  $\eta$  are regarded as fixed control parameters, and  $\varepsilon$  is varied quasi-statically as a bifurcation parameter. This was the approach of Uppal, Ray and Poore in [7]. While it is possible to regard any of the parameters as a bifurcation variable, in any experiment it is clear that  $\varepsilon$  can be varied independently by adjusting the flow rate, while it would be difficult to design an experiment in which changing a single physical variable changed only one other dimensionless variable.

Thus, in what follows, a “bifurcation diagram” is defined as the graph of the steady-state solutions of (1.6) versus  $\varepsilon$ . The description is simplified somewhat in this problem because  $x$  or  $y$  can be eliminated from the equations  $f_1 = f_2 = 0$  and the equilibrium is determined by a single state variable, temperature or concentration, alone. Since (1.6) is linear in  $x$ , it is convenient to eliminate  $x$  by

$$(1.14) \quad x = \frac{DA(y)}{\varepsilon + DA(y)} = \frac{\eta - (1 + \varepsilon)y + BDA(y)}{BDA(y)}.$$

Introducing the notation  $\delta = 1/D$  and  $\mathcal{A}(y) = 1/(A(y))$ , we find the equilibrium temperature satisfies  $G = 0$ , where

$$(1.15) \quad G(y, \varepsilon, B, \delta, \eta) = \eta - (1 + \varepsilon)y + \frac{B\varepsilon}{1 + \varepsilon\delta\mathcal{A}(y)}.$$

All the qualitative analysis of the bifurcation diagrams is based on an analysis of the  $C^\infty$  function  $G$ .

**2. The theory.** In this section we shall state the theorems of [4] specialized to one state variable and discuss in detail the “winged cusp” singularity which we claim is the organizing center for the bifurcation problem associated to the stirred tank reactor described in § 1.

Let  $\mathcal{E}_{x,\lambda}$  be the space of  $C^\infty$  germs of mappings from  $\mathbb{R}^2 \rightarrow \mathbb{R}$  at 0 depending on the variables  $x$  and  $\lambda$ . A *bifurcation problem* is the solution of

$$(2.1) \quad G(x, \lambda) = 0,$$

where  $G(0, 0) = 0$  for  $G$  in  $\mathcal{E}_{x,\lambda}$ . Two bifurcation problems  $G$  and  $H$  are *contact equivalent* if

$$(2.2) \quad G(x, \lambda) = T(x, \lambda)H(X(x, \lambda), \Lambda(\lambda)),$$

where  $T(0, 0) \neq 0$ ,  $(\partial X/\partial x)(0) > 0$ ,  $(\partial \Lambda/\partial \lambda)(0) > 0$ , and  $X(0) = \Lambda(0) = 0$ . We shall use contact equivalence as our formalization of the term “qualitatively similar” for bifurcation problems as discussed in the Introduction.

There are two problems about contact equivalence which need to be investigated in order to analyze the stirred tank reactor. Although these problems have similar statements their resolution requires different methods. First, when is a bifurcation problem  $G$  contact equivalent to a (simple) polynomial and if it is how does one find this normal form? Second, we ask this question for a  $k$ -parameter family of given bifurcation problems. As we shall see the theoretical answer to both questions is the same although the mathematical sophistication needed to prove the second is of a much higher order.

Let

$$(2.3) \quad \tilde{T}G = \left\langle G, \frac{\partial G}{\partial x} \right\rangle$$

be the ideal in  $\mathcal{E}_{x,\lambda}$  generated by  $G$  and  $\partial G/\partial x$ ; that is, all function germs of the form

$$a(x, \lambda)G(x, \lambda) + b(x, \lambda) \frac{\partial G}{\partial x}(x, \lambda),$$

where  $a, b \in \mathcal{E}_{x,\lambda}$ .

**DEFINITION 2.4.**  $G$  has *finite codimension* if there exists a finite dimensional vector space  $V \subset \mathcal{E}_{x,\lambda}$  such that  $\tilde{T}G \oplus V = \mathcal{E}_{x,\lambda}$ .

Theorem 2.8 of [4] states that if  $G$  has finite codimension then  $G$  is contact equivalent to a polynomial. More interesting is the question of how one finds this normal form. The main step is given by the following proposition whose proof is elementary, requiring only the standard existence theorem for ordinary differential equations, and is a special case of the discussion after Lemma 3.8 of [4].

**PROPOSITION 2.5.** *Let  $H = G + P$  and define  $G_t$  to be  $G + tP$ . Then  $H$  is contact equivalent to  $G$  if  $\tilde{T}G_t = \tilde{T}G$  for  $0 \leq t \leq 1$ .*

The following is useful for checking the hypothesis of Proposition 2.5. Let  $\mathcal{M} = \langle x, \lambda \rangle$  be the maximal ideal generated by  $x$  and  $\lambda$ .

**LEMMA 2.6** (Nakayama’s lemma). *Let  $\mathcal{I} = \langle p_1, \dots, p_k \rangle$  be the ideal in  $\mathcal{E}_{x,\lambda}$  generated by  $p_1, \dots, p_k$  and suppose that  $q_1, \dots, q_k$  are in  $\mathcal{M}\mathcal{I}$ . Then  $\mathcal{I} = \langle p_1 + q_1, \dots, p_k + q_k \rangle$ .*



Note.  $\mathcal{MS}$  denotes the product of the ideals  $\mathcal{M}$  and  $\mathcal{S}$  and is the ideal generated by the products of the generators of  $\mathcal{M}$  and  $\mathcal{S}$ .

Proof. See, for example, Lemma 3.10 of [4].

Before discussing the second problem we analyze two bifurcation problems which both occur in the stirred tank problem and serve as examples of the general theory.

PROPOSITION 2.7. Let  $H(x, \lambda)$  satisfy one of the following set of conditions:

$$(2.8a) \quad \bar{H} = \bar{H}_x = \bar{H}_\lambda = \bar{H}_{xx} = \bar{H}_{x\lambda} = 0 \quad \text{and} \quad \bar{H}_{xxx}\bar{H}_{\lambda\lambda} > 0,$$

$$(2.8b) \quad \bar{H} = \bar{H}_x = \bar{H}_\lambda = \det(\bar{d}^2\bar{H}) = 0 \quad \text{and} \quad \bar{H}_{xx}\bar{d}^3\bar{H}(v, v, v) > 0,$$

where the bar indicates evaluation at  $x = \lambda = 0$  and  $v \neq 0$  satisfies  $(\bar{d}^2\bar{H})(v) = 0$ . Then (i)  $\tilde{T}H$  is computed to be

$$(2.9a) \quad \langle \lambda^2, x^2 + 2(\bar{H}_{xx\lambda}/\bar{H}_{xxx})x\lambda \rangle$$

or

$$(2.9b) \quad \left\langle \lambda^3, x + \frac{\bar{H}_{x\lambda}}{\bar{H}_{xx}}\lambda + \left( \frac{\bar{H}_{x\lambda\lambda}}{2\bar{H}_{xx}} - \frac{\bar{H}_{x\lambda}\bar{H}_{xx\lambda}}{\bar{H}_{xx}^2} + \frac{\bar{H}_{x\lambda}^2\bar{H}_{xxx}}{2\bar{H}_{xx}^3} \right)\lambda^2 \right\rangle$$

and (ii)  $H$  is contact equivalent to

$$(2.10a) \quad x^3 + \lambda^2$$

or

$$(2.10b) \quad x^2 + \lambda^3$$

respectively.

Note. We call the bifurcation problem  $G(x, \lambda) = x^3 + \lambda^2$  a winged cusp.

Proof. The main part of the proof is the computation of  $\tilde{T}H$ . We show first how (ii) follows from this computation along with Proposition 2.5. The assumption (2.8a) implies

$$(2.11a) \quad H(x, \lambda) = a\lambda^2 + bx^3 + cx^2\lambda + dx\lambda^2 + e\lambda^3 + Q(x, \lambda),$$

where  $Q(x, \lambda)$  begins with terms of order four and  $ab > 0$ . Observe that by a change in coordinates of the form  $x = \tilde{x} + B\lambda$  we can assume that  $2c = \bar{H}_{xx\lambda} = 0$ . After this preliminary change of coordinates the computation of  $\tilde{T}H$  given by (2.9a) shows that  $\tilde{T}H = \langle \lambda^2, x^2 \rangle$ . Let  $P = dx\lambda^2 + e\lambda^3 + Q(x, \lambda)$  and apply Proposition 2.5 to see that  $H$  is contact equivalent to  $bx^3 + a\lambda^2$ . Since multiplication by  $-1$  and scaling are contact equivalences (2.10a) is proved. As case (b) of Proposition 2.7 is similar we just point out briefly that assumption (2.8b) implies

$$(2.11b) \quad H(x, \lambda) = ax^2 + bx\lambda + c\lambda^2 + dx^3 + ex^2\lambda + fx\lambda^2 + g\lambda^3 + Q(x, \lambda),$$

where  $Q$  is as above and  $a \neq 0$ . The computation of  $\tilde{T}H$  given in (2.9b) shows that if we can make preliminary changes of coordinates so that  $b = f = 0$  then (2.10b) will follow from Proposition 2.5. The assumption that  $\det(\bar{d}^2\bar{H}) = 0$  implies

$$(2.12) \quad ax^2 + bx\lambda + c\lambda^2 = a\left(x + \frac{b}{2a}\lambda\right)^2.$$

Letting  $\tilde{x} = x + (b/2a)\lambda$  puts  $H$  in the form (2.11b) with  $b = c = 0$ . A short calculation shows that letting  $\tilde{x} = \tilde{x} + B\lambda^2$  will now put  $H$  in the form (2.11b) with  $f = 0$  also.

To compute (2.9a) and (2.9b) we will make repeated use of Nakayama’s lemma along with the following simple observation. Let  $P, A, B, f$  be in  $\mathcal{E}_{x,\lambda}$ . Then

$$(2.13) \quad \langle A, P \rangle = \langle B, P \rangle \quad \text{if } A = B + fP.$$

First we compute (2.9a). From (2.11a) we see that

$$(2.14) \quad \tilde{T}H = \left\langle \lambda^2 + C, x^2 + \frac{2c}{3b} x\lambda + C' \right\rangle,$$

where  $C$  and  $C'$  begin with terms of order three. Observe that

$$\mathcal{M}^3 \subset \mathcal{M} \left\langle \lambda^2, x^2 + \frac{2c}{3b} x\lambda \right\rangle$$

so that Nakayama’s lemma implies

$$(2.15) \quad \tilde{T}H = \left\langle \lambda^2, x^2 + \frac{2c}{3b} x\lambda \right\rangle.$$

As  $c = \bar{H}_{xx\lambda}/2$  and  $b = \bar{H}_{xxx}/6$  (2.9a) is proved.

To compute (2.9b) observe that (2.11b), (2.12), and (2.13) imply

$$(2.16) \quad \tilde{T}H = \langle dx^3 + ex^2\lambda + fx\lambda^2 + g\lambda^3 + Q', 2ax + b\lambda + 3dx^2 + 2ex\lambda + f\lambda^2 + C \rangle,$$

where  $C = \text{cubic} + \dots$  and  $Q' = \text{quartic} + \dots$ . Note that  $x + (b/2a)\lambda \equiv \text{quadratic} + \dots \pmod{\tilde{T}H}$ ; thus  $(x + (b/2a)\lambda)^2 \equiv \text{quartic} + \dots \pmod{\tilde{T}H}$ . Hence the cubic terms in the first generator of  $\tilde{T}H$  in (2.16) are the same as the cubic terms of  $H$  as in (2.11b). Next observe that  $x \equiv -(b/2a)\lambda + \dots \pmod{\tilde{T}H}$ ; thus (2.16) implies

$$(2.17) \quad \tilde{T}H = \langle K\lambda^3 + Q''(\tilde{x}, \lambda), \tilde{x} + C'(\tilde{x}, \lambda) \rangle,$$

where  $K = (\overline{d^3H})(v, v, v) \neq 0$  and  $\tilde{x} = 2ax + b\lambda + 2ex\lambda + 3dx^2 + f\lambda^2$ . To see that  $K$  is as claimed one needs the following observation:

$$(2.18) \quad 6(d^3\bar{H})(v, v, v) = -\bar{H}_{xxx} \left( \frac{\bar{H}_{x\lambda}}{\bar{H}_{xx}} \right)^3 + 3\bar{H}_{xx\lambda} \left( \frac{\bar{H}_{x\lambda}}{\bar{H}_{xx}} \right)^2 - 3\bar{H}_{x\lambda\lambda} \left( \frac{\bar{H}_{x\lambda}}{\bar{H}_{xx}} \right) + \bar{H}_{\lambda\lambda\lambda},$$

which is obtained from the fact that  $v$  may be taken to be  $(-\bar{H}_{x\lambda}/\bar{H}_{xx}, 1)$ .

Since  $a \neq 0$ ,  $\tilde{x}$  is a legitimate change of coordinates. One may use Nakayama’s lemma in the  $\tilde{x}, \lambda$  coordinates to obtain

$$(2.19) \quad \tilde{T}H = \langle \lambda^3, \tilde{x} \rangle = \left\langle \lambda^3, x + \frac{b\lambda + f\lambda^2}{2a + 3dx + 2e\lambda} \right\rangle$$

so  $\mathcal{M}^3 \subset \tilde{T}H$ . Next compute

$$(2.20) \quad \frac{b\lambda + f\lambda^2}{2a + 3dx + 2e\lambda} \equiv \frac{b}{2a} \lambda + \left( \frac{f}{2a} - \frac{be}{2a^2} \right) \lambda^2 - \frac{3bd}{4a^2} x\lambda \pmod{\mathcal{M}^3}.$$

Therefore using (2.13) we have

$$(2.21) \quad \tilde{T}H = \left\langle \lambda^3, x + \frac{b}{2a} \lambda + \left( \frac{f}{2a} - \frac{be}{2a^2} + \frac{3b^2d}{8a^3} \right) \lambda^2 \right\rangle.$$

Using the fact that  $a = \bar{H}_{xx}/2$ ,  $b = \bar{H}_{x\lambda}$ ,  $d = \bar{H}_{xxx}/6$ ,  $e = \bar{H}_{xx\lambda}/2$ , and  $f = \bar{H}_{x\lambda\lambda}/2$  the proposition is proved.

We now turn to the second problem; polynomial normal forms for  $k$ -parameter families of bifurcation problems. This is formalized through the notion of unfoldings and solved through the notion of universal unfoldings.

DEFINITION 2.22. (i)  $F: (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^k, 0) \rightarrow \mathbb{R}$  is a  $k$ -parameter unfolding of  $G$  in  $\mathcal{E}_{x,\lambda}$  if  $F(x, \lambda, 0) = G(x, \lambda)$ .

(ii) Let  $H(x, \lambda, \beta)$  be an  $m$ -parameter unfolding of  $G$ . Then  $H$  factors through  $F$  if

$$(2.23) \quad H(x, \lambda, \beta) = F(X(x, \lambda, \beta), \Lambda(\lambda, \beta), \alpha(\beta)),$$

where all mappings are smooth and  $\alpha(0) = 0$ .

(iii) Two unfoldings  $H$  and  $F$  are equivalent if  $H$  factors through  $F$  and the map  $\alpha \rightarrow \beta(\alpha)$  in (2.23) is an invertible change in coordinates (so  $m = l$ ).

(iv)  $F$  is a universal unfolding<sup>1</sup> of  $G$  if every unfolding  $H$  factors through  $F$ .

Note (a). The number of parameters in  $H$  need not be the same as the number in  $F$ .

Note (b). Equation (2.23) means that for every  $\beta$ ,  $H(\cdot, \cdot, \beta)$  is contact equivalent to  $F(\cdot, \cdot, \alpha)$  for some  $\alpha$ . Thus, if  $H$  factors through  $F$  then every bifurcation problem included in the unfolding  $H$  is already included in the unfolding  $F$ , at least up to contact equivalence.

In what follows we shall show why it is relatively easy to put a universal unfolding into a polynomial normal form.

PROPOSITION 2.24. Let  $F$  and  $H$  be universal unfoldings of  $G$  depending on the same number of parameters. Then  $F$  and  $H$  are equivalent.

Proof. Proposition 2.5 of [4].

THEOREM 2.25. Let  $F(x, \lambda, \alpha)$  be an  $l$ -parameter unfolding of  $G(x, \lambda)$  and assume that  $G$  has finite codimension. Then  $F$  is a universal unfolding if

$$(2.26) \quad \mathcal{E}_{x,\lambda} = \tilde{T}G + \mathcal{E}_\lambda \left\{ \frac{\partial G}{\partial \lambda} \right\} + \mathbb{R} \left\{ \left. \frac{\partial F}{\partial \alpha_1} \right|_{\alpha=0}, \dots, \left. \frac{\partial F}{\partial \alpha_k} \right|_{\alpha=0} \right\}.$$

Proof. Theorem 2.4 of [4].

We see from (2.26) that  $G$  has a universal unfolding precisely when  $G$  has finite codimension. The following remarks should make this clear.

Note. Equation (2.26) may be restated as follows: for every germ  $p(x, \lambda)$  there exist function germs  $a(x, \lambda)$ ,  $b(x, \lambda)$ , and  $c(\lambda)$ —not  $c(x, \lambda)$ —and real numbers  $r_1, \dots, r_k$  such that

$$(2.27) \quad p(x, \lambda) = a(x, \lambda)G + g(x, \lambda) \frac{\partial G}{\partial x} + c(\lambda) \frac{\partial G}{\partial \lambda} + r_1 \frac{\partial F}{\partial \alpha_1}(x, \lambda, 0) + \dots + r_k \frac{\partial F}{\partial \alpha_k}(x, \lambda, 0).$$

This condition may look difficult to check but, in reality, it is not. Consider:

Example 2.28. Let  $G(x, \lambda) = x^3 + \lambda^2$ . Then  $F(x, \lambda, \alpha_1, \alpha_2, \alpha_3) = x^3 + (\alpha_2\lambda + \alpha_3)x + \alpha_1 + \lambda^2$  is a universal unfolding of  $G$ . Moreover  $\text{codim } G = 3$ .

DEFINITION 2.29. The codimension of  $G$  is the minimum number of parameters necessary for a universal unfolding of  $G$ .

Proof. From (2.9a),  $\tilde{T}G = \langle x^2, \lambda^2 \rangle$ . Hence (2.27) becomes

$$(2.30) \quad p(x, \lambda) = a(x, \lambda)x^2 + b(x, \lambda)\lambda^2 + c(\lambda)\lambda + r_1 + r_2\lambda x + r_3x.$$

It is easy to check that (2.30) holds for all  $p$  by Taylor's theorem.

<sup>1</sup> In the literature the term "universal" is reserved for the unfolding in (iv) with the minimum number of parameters, and "versal" for what we have defined. We shall not make this distinction.

All of the germs  $G$  that will be considered in this paper have the property that  $\lambda(\partial G/\partial \lambda)$  is contained in  $\tilde{T}G$ . As a result (2.26) may be reduced to a question of linear algebra.

**COROLLARY 2.31.** *Assume  $\lambda(\partial G/\partial \lambda)$  is in  $\tilde{T}G$ , and let  $q_1(x, \lambda), \dots, q_s(x, \lambda)$  be a basis for a complementary subspace to  $\tilde{T}G$  in  $\mathcal{E}_{x,\lambda}$ . Let  $F(x, \lambda, \alpha)$  be an  $l$ -parameter unfolding of  $G(x, \lambda)$ .*

Let

$$\frac{\partial F}{\partial \alpha_i}(x, \lambda, 0) = c_{i,1}q_1 + \dots + c_{i,s}q_s + t_i$$

and

$$\frac{\partial G}{\partial \lambda}(x, \lambda) = c_{l+1,1}q_1 + \dots + c_{l+1,s}q_s + t_{l+1},$$

where  $t_i$  is in  $\tilde{T}G$  for  $1 \leq i \leq l+1$ . Then  $F$  is a universal unfolding if  $\text{rank } C = s$  where  $C = (c_{ij})$  is the  $(l+1) \times s$  matrix described above.

*Note.* Example 2.28 is now a triviality as a complementary space to  $\langle x^2, \lambda^2 \rangle = \tilde{T}G$  is spanned by  $1, x, \lambda, x\lambda$ .

Examples of the application of Theorem 2.25 and its Corollary 2.31 in identifying universal unfoldings can be found in [4]. We provide the specific results for the new singularities—the winged cusp  $(x^3 + \lambda^2)$  and  $x^2 + \lambda^3$ —which arise in the present application in the next proposition.

**PROPOSITION 2.32.** *Let  $F(x, \lambda, \alpha_1, \alpha_2, \alpha_3)$  be an unfolding of  $G(x, \lambda)$ . Assume that (i)  $G$  satisfies (2.8a) and suppose that  $\text{rank } C = 4$  where  $C$  is the matrix*

$$\begin{pmatrix} F_{\alpha_1} & F_{\alpha_1 x} & F_{\alpha_1 \lambda} & F_{\alpha_1 x \lambda} - \frac{G_{xx\lambda}}{G_{xxx}} F_{\alpha_1 xx} \\ F_{\alpha_2} & F_{\alpha_2 x} & F_{\alpha_2 \lambda} & F_{\alpha_2 x \lambda} - \frac{G_{xx\lambda}}{G_{xxx}} F_{\alpha_2 xx} \\ F_{\alpha_3} & F_{\alpha_3 x} & F_{\alpha_3 \lambda} & F_{\alpha_3 x \lambda} - \frac{G_{xx\lambda}}{G_{xxx}} F_{\alpha_3 xx} \\ 0 & 0 & G_{\lambda\lambda} & G_{x\lambda\lambda} - \frac{G_{xx\lambda}^2}{G_{xxx}} \end{pmatrix}$$

evaluated at  $x = \lambda = 0$ , or

(ii)  $G$  satisfies (2.8b) and suppose that  $\text{rank } C = 2$  where  $C$  is the matrix

$$\begin{pmatrix} F_{\alpha_1} & F_{\alpha_1 \lambda} - \frac{G_{x\lambda}}{G_{xx}} F_{\alpha_1 x} \\ F_{\alpha_2} & F_{\alpha_2 \lambda} - \frac{G_{x\lambda}}{G_{xx}} F_{\alpha_2 x} \\ F_{\alpha_3} & F_{\alpha_3 \lambda} - \frac{G_{x\lambda}}{G_{xx}} F_{\alpha_3 x} \end{pmatrix}$$

evaluated at  $x = \lambda = 0$ . Then  $F$  is a universal unfolding of  $G$ .

*Note.* One may apply Proposition 2.24 to see that if  $F$  is a universal unfolding as in (i) then  $F$  is contact equivalent as a parameterized family to  $x^3 + (\alpha_2 + \alpha_3 \lambda)x + \alpha_1 + \lambda^2$ , thus solving our second problem for the winged cusp.

*Proof.* The heart of the proof has already been completed by the computation of  $\tilde{T}G$  in (2.9a) and (2.9b). Given a germ  $Q(x, \lambda)$  in  $\mathcal{E}_{x,\lambda}$  we may write—where  $t(x, \lambda) \in \tilde{T}G$ —

$$(2.33a) \quad Q(x, \lambda) = \bar{Q} + \bar{Q}_{xx}x + \bar{Q}_\lambda\lambda + \left( \bar{Q}_{x\lambda} - \frac{\bar{G}_{xx\lambda}}{\bar{G}_{xxx}} \bar{Q}_{xx} \right) x\lambda + t(x, \lambda)$$

or

$$(2.33b) \quad Q(x, \lambda) = \bar{Q} + (\bar{Q}_\lambda - A\bar{Q}_x)\lambda + K(Q)\lambda^2 + t(x, \lambda),$$

where  $A = \bar{G}_{x\lambda}/\bar{G}_{xx}$  and

$$(2.34) \quad K(Q) = \frac{1}{2}[\bar{Q}_{xx}A^2 - 2\bar{Q}_{x\lambda}A - 2\bar{Q}_xB + \bar{Q}_{\lambda\lambda}]$$

and

$$B = \frac{\bar{G}_{x\lambda\lambda}}{2\bar{G}_{xx}} - \frac{\bar{G}_{x\lambda}\bar{G}_{xx\lambda}}{\bar{G}_{xx}^2} + \frac{\bar{G}_{x\lambda}^2\bar{G}_{xxx}}{2\bar{G}_{xx}^3}.$$

In case (i) the proposition follows from Corollary 2.31 directly along with the observation that (2.8a) implies that  $\bar{G}_\lambda = \bar{G}_{\lambda x} = 0$ .

In case (ii) Corollary 2.31 implies that  $F$  is a universal unfolding of  $G$  if

$$\text{rank} \begin{pmatrix} F_{\alpha_1} & F_{\alpha_1\lambda} - AF_{\alpha_1x} & K(F_{\alpha_1}) \\ F_{\alpha_2} & F_{\alpha_2\lambda} - AF_{\alpha_2x} & K(F_{\alpha_2}) \\ F_{\alpha_3} & F_{\alpha_3\lambda} - AF_{\alpha_3x} & K(F_{\alpha_3}) \\ G_\lambda & G_{\lambda\lambda} - AG_{\lambda x} & K(G_\lambda) \end{pmatrix} = 3.$$

One computes—using (2.18)—that  $K(G_\lambda) = (\bar{d}^3\bar{G})(v, v, v) \neq 0$  by (2.8b). Also by (2.8b)  $\bar{G}_\lambda = 0$  and  $\bar{G}_{\lambda\lambda} - A\bar{G}_{\lambda x} = \det(\bar{d}^3\bar{G})/\bar{G}_{xx} = 0$ . So the proposition is proved.

We are now ready to discuss the problem of classifying—up to contact equivalence—the types of bifurcations which occur in the universal unfolding of a given problem. Suppose one has a bifurcation problem  $G(x, \lambda)$  and an  $l$ -parameter universal unfolding  $F(x, \lambda, \alpha)$ , how does one classify in a qualitative way the types of bifurcation diagrams  $F(\cdot, \cdot, \alpha) = 0$  for various  $\alpha$ ? A good start at the answer is given by the following theorem. First observe that if  $G$  has a universal unfolding then it is contact equivalent to a polynomial and if  $G$  is a polynomial then  $F$  may also be assumed to be a polynomial. (This is Corollary 2.9 of [4].) Next define

$$\begin{aligned} (\mathcal{B}) &= \{\alpha \in \mathbb{R}^l \mid \exists x, \lambda \text{ with } F = F_x = F_\lambda = 0 \text{ at } (x, \lambda, \alpha)\}, \\ (\mathcal{H}) &= \{\alpha \in \mathbb{R}^l \mid \exists x, \lambda \text{ with } F = F_x = F_{xx} = 0 \text{ at } (x, \lambda, \alpha)\}, \\ (\mathcal{DL}) &= \{\alpha \in \mathbb{R}^l \mid \exists (x_1, \lambda_1) \text{ and } (x_2, \lambda_2) \text{ with } F = F_x = 0 \\ &\quad \text{at both } (x_1, \lambda_1, \alpha) \text{ and } (x_2, \lambda_2, \alpha)\}. \end{aligned}$$

These are called the *bifurcation*, *hysteresis*, and *double limit* varieties, respectively.

**THEOREM 2.35.** *Let  $\Sigma = (\mathcal{B}) \cup (\mathcal{H}) \cup (\mathcal{DL}) \subset \mathbb{R}^l$ . (Note that  $\Sigma$  is a codimension one algebraic variety in  $\mathbb{R}^l$ .) Then there exist open neighborhoods  $\mathcal{U}$  of 0 in  $\mathbb{R} \times \mathbb{R}$  and  $\mathcal{V}$  of 0 in  $\mathbb{R}^l$  such that if  $\alpha_1$  and  $\alpha_2$  are in the same connected component of  $\mathcal{V} \sim \Sigma$  then  $F(\cdot, \cdot, \alpha_1)$  and  $F(\cdot, \cdot, \alpha_2)$  are contact equivalent on  $\mathcal{U}$ .*

*Proof.* This is Corollary 2.16 of [4].

Using this theorem we analyze the local nature of bifurcation diagrams near the winged cusp.

**PROPOSITION 2.36.** *Let  $F(x, \lambda, \alpha) = x^3 + (\alpha_2 + \alpha_3\lambda)x + \alpha_1 + \lambda^2$ . Then*

$$(\mathcal{H}) = \{\alpha_2^2 + \alpha_1\alpha_3^2 = 0; \alpha_2 \leq 0\}, \quad (\mathcal{DL}) = \emptyset$$

and  $(\mathcal{B})$  is parameterized by the equations

$$\alpha_1 = 2x^3 - \frac{\alpha_3^2 x^2}{4}, \quad \alpha_2 = -3x^2 + \frac{\alpha_3 x}{2}.$$

*Proof.* A short computation.

To visualize how the varieties  $(\mathcal{B})$  and  $(\mathcal{H})$  intertwine it is perhaps easiest to graph  $(\mathcal{B})$  and  $(\mathcal{H})$  for  $\alpha_3$  fixed. The results are given in Fig. 2.1. The numbered regions correspond to connected components of the complement of  $\Sigma$ . The lettered regions correspond to various branches of the variety  $\Sigma$ . The bifurcation diagrams are given in Fig. 2.2. (Note that the diagrams associated with  $\Sigma$  are obtained by continuity as one crosses  $\Sigma$ .) Also observe that  $(\mathcal{H})$  is just the ‘‘Whitney Umbrella’’ while  $(\mathcal{B})$  is a cylinder over a cusp curve. They are pictured in Fig. 2.3.

In the Introduction we stated that the winged cusp is an ‘‘organizing center’’ for bifurcation diagrams associated with the stirred tank reactor described in § 1. We are now in a position to make that statement more precise.









**PROPOSITION 2.37.** *Let  $G(x, \lambda)$  be defined on  $\Omega$  in  $\mathbb{R}^2$ . Assume that the following sets of equations are never satisfied in  $\Omega$*

- (i)  $G_{\lambda\lambda} = 0$ ;
- (ii)  $G = G_x = G_{xx} = G_{xxx} = 0$ ; and
- (iii)  $G = G_x = G_\lambda = \det(d^2G) = d^3G(v, v, v) = 0$ ;

where  $(d^2G)(v, v) = 0$ . Then at any point  $(x_0, \lambda_0)$  in  $\Omega$  for which  $G(x_0, \lambda_0) = 0$ , the local nature of the bifurcation diagram  $\{G = 0\}$  is described by one of the eight singularities in Table 2.1. Moreover each of these local situations occurs in the universal unfolding of the winged cusp.

*Proof.* A simple check shows that conditions (1)–(8) of Table 2.1 yield an exhaustive list for the possibilities for  $G$  satisfying (i)–(iii). The normal forms for the singularities (1)–(4) and (6)–(7) are given by Proposition 4.1 of [4]. Singularities (5) and (8) were given in Proposition 2.7.

TABLE 2.1

Defining conditions at $(x_0, \lambda_0)$	Normal form	Bifurcation diagram	Codimension
(1) $G = 0, G_x \neq 0$	$x$		0
(2) $G = G_x = 0, G_{xx} \cdot G_\lambda \neq 0$	$x^2 \pm \lambda$		0
(3) $G = G_x = G_\lambda = 0, G_{xx} \cdot \det d^2G \neq 0$	$x^2 - \lambda^2$		1
(4) $G = G_x = G_\lambda = 0, G_{xx} \cdot \det(d^2G) \neq 0$ index $d^2G = 1$	$x^2 + \lambda^2$		1
(5) $G = G_x = G_\lambda = \det(d^2G) = 0$ $G_{xx} \cdot (d^3G)(v, v, v) \neq 0$	$x^2 + \lambda^3$		2
(6) $G = G_x = G_{xx} = 0$ $G_{xxx} \cdot G_\lambda \neq 0$	$x^3 \pm \lambda$		1
(7) $G = G_x = G_{xx} = G_\lambda = 0$ $G_{xxx} \cdot G_{\lambda x} \neq 0$	$x^3 \pm \lambda x$		2
(8) $G = G_x = G_{xx} = G_\lambda = G_{\lambda x} = 0$ $G_{xxx} \cdot G_{\lambda\lambda} \neq 0$	$x^3 \pm \lambda^2$		3

We shall use the following specialized result in our analysis for the stirred tank reactor in the next section.

**PROPOSITION 2.38.** *Let  $F(x, \lambda, \alpha_1, \alpha_2, \alpha_3) = \alpha_1 + \tilde{F}(x, \lambda, \alpha_2, \alpha_3)$  be an unfolding of  $G(x, \lambda)$  as in Proposition 2.37. Then  $F$  is a universal unfolding of  $G$  if—in each of the eight cases listed in Table 2.1—the following conditions are satisfied.*

TABLE 2.2

Case	Condition
(1)–(4)	Always
(5)	$F_{\alpha_3 x} F_{\alpha_2 \lambda} - F_{\alpha_2 x} F_{\alpha_3 \lambda} \neq 0$
(6)	$\text{rank} \begin{pmatrix} F_{\alpha_2 x} & F_{\alpha_3 x} & G_{x\lambda} \end{pmatrix} = 1$
(7)	$\text{rank} \begin{pmatrix} F_{\alpha_2 x} & F_{\alpha_2 xx} - F_{\alpha_2 \lambda} \frac{G_{xxx}}{G_{x\lambda}} \\ F_{\alpha_3 x} & F_{\alpha_3 xx} - F_{\alpha_3 \lambda} \frac{G_{xxx}}{G_{x\lambda}} \\ G_{\lambda x} & G_{\lambda xx} - G_{\lambda \lambda} \frac{G_{xxx}}{G_{x\lambda}} \end{pmatrix} = 2$
(8)	$\text{rank} \begin{pmatrix} F_{\alpha_2 x} & F_{\alpha_2 \lambda} & F_{\alpha_2 x \lambda} - \frac{G_{xx\lambda}}{G_{xxx}} F_{\alpha_2 xx} \\ F_{\alpha_3 x} & F_{\alpha_3 \lambda} & F_{\alpha_3 x \lambda} - \frac{G_{xx\lambda}}{G_{xxx}} F_{\alpha_3 xx} \\ 0 & G_{\lambda \lambda} & G_{x\lambda \lambda} - \frac{G_{xx\lambda}^2}{G_{xxx}} \end{pmatrix} = 3$

*Proof.* Cases (5) and (8) are easy consequences of Proposition 2.32. The remaining cases are proved in a fashion similar to that proposition.

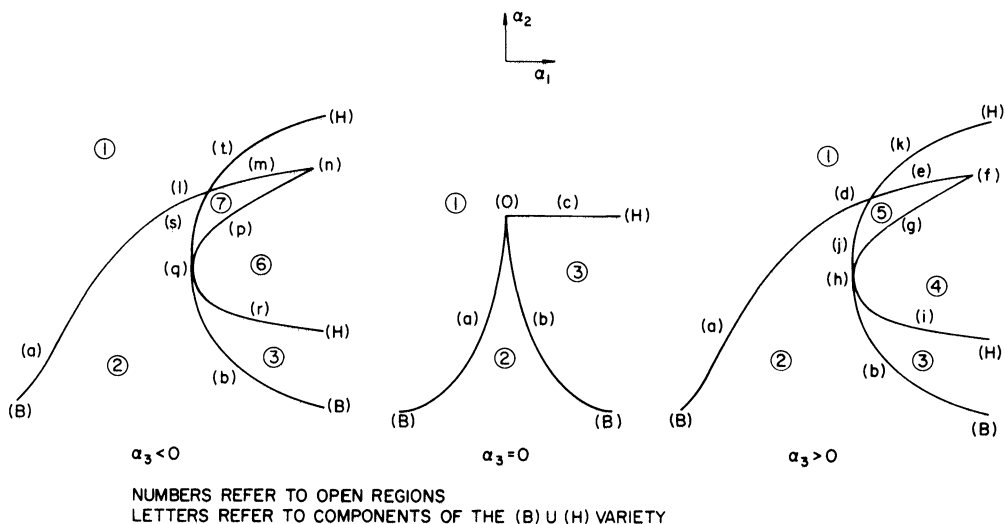
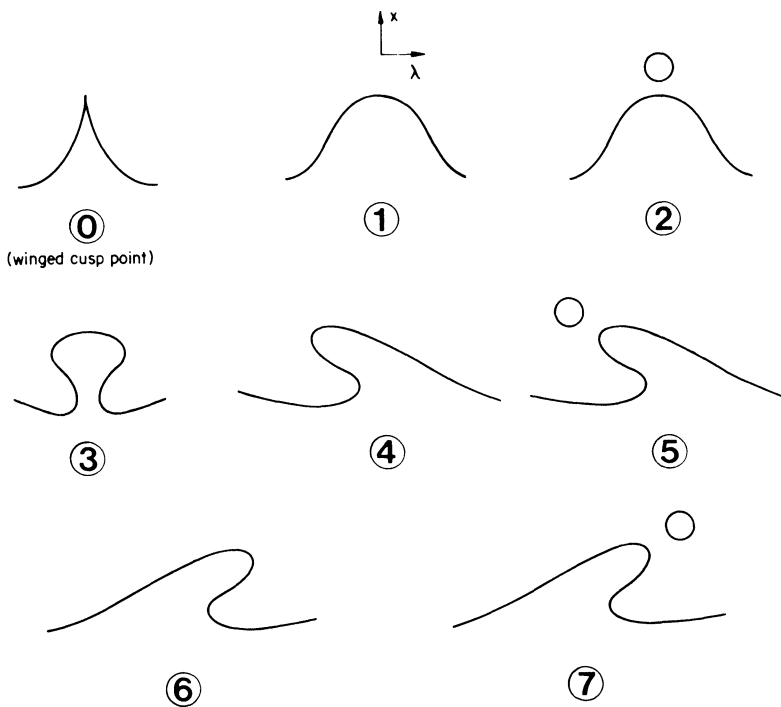


FIG. 2.1

**3. The local nature of the bifurcation diagrams.** In § 1 we showed that the steady state solutions to our model chemical reactor are described by the equation:

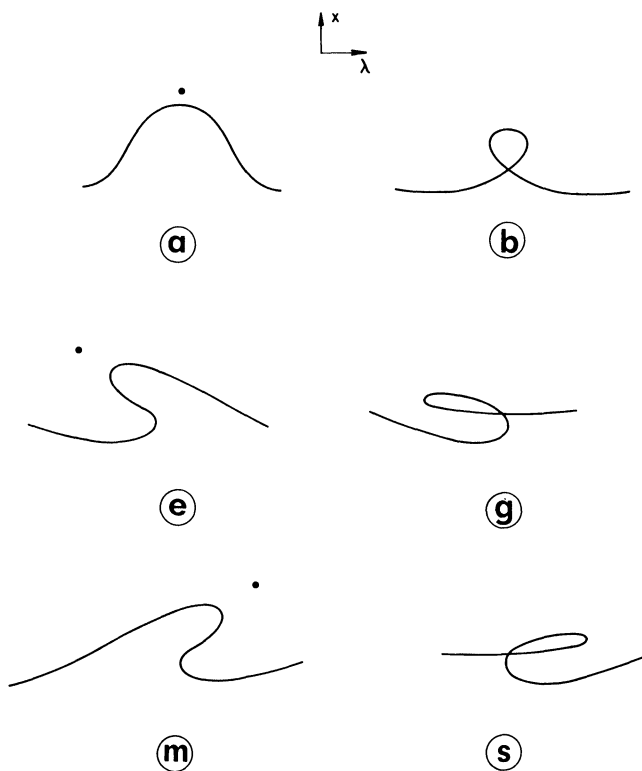
$$(3.1) \quad G(y, \varepsilon, B, \delta, \eta) = \eta - (1 + \varepsilon)y + \frac{B\varepsilon}{1 + \varepsilon\delta\mathcal{A}(y)} = 0,$$

where  $y$  is a nondimensionalized temperature,  $\varepsilon$  is a nondimensionalized flow rate,  $B$ ,  $\delta$  and  $\eta$  are parameters, and  $\mathcal{A}$  is a reaction rate term which is usually assumed to have the



THE OPEN REGIONS

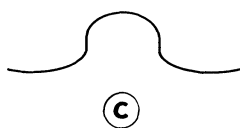
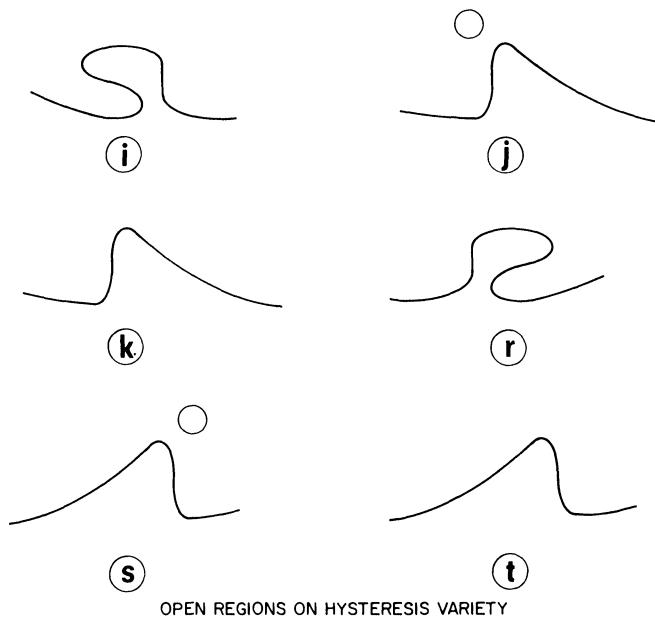
FIG. 2.2a



OPEN REGIONS ON BIFURCATION VARIETY

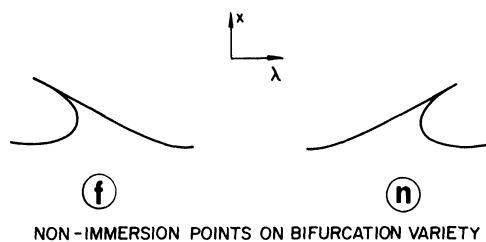
FIG. 2.2b



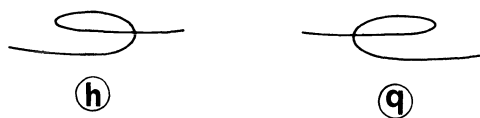


SELF-INTERSECTION OF HYSTERESIS VARIETY

FIG. 2.2c

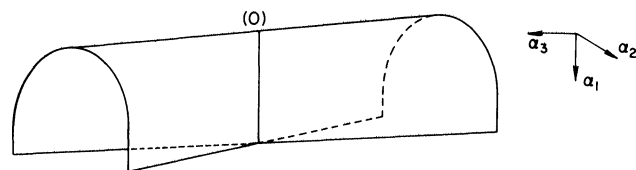


TRANSVERSE INTERSECTION OF BIFURCATION AND HYSTERESIS VARIETIES

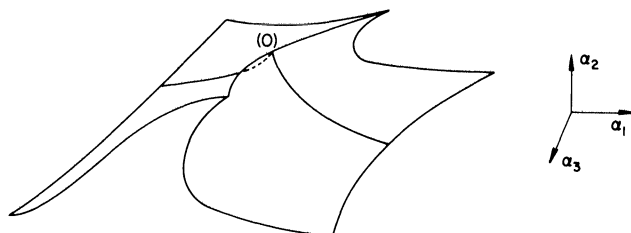


TANGENCY OF BIFURCATION AND HYSTERESIS VARIETIES

FIG. 2.2d



THE HYSTERESIS VARIETY



THE BIFURCATION VARIETY

FIG. 2.3

form

$$(3.2) \quad \mathcal{A}(y) = \exp\left(\frac{-\gamma y}{1+y}\right).$$

When  $\mathcal{A}$  has the form (3.2) we call  $\mathcal{A}$  an *Arrhenius term* with activation energy  $\gamma > 0$ .

The problem we address in this section is the global classification of the local bifurcation problems which appear in this model. We shall prove that for each member of a class of reaction terms which are both open and include the Arrhenius terms when  $\gamma > 8/3$ , there is a unique winged cusp point and that globally the only local bifurcation problems which occur are those found in the universal unfolding for the winged cusp. Moreover, the physically motivated parameters  $B$ ,  $\delta$ , and  $\eta$  turn out to be universal unfolding parameters; it is indeed a curious fact that these parameters—given physically—are the minimum number necessary to determine the qualitative classification. This fact suggests strongly that the winged cusp should be considered as the “organizing center” for this model.

The region in space which we consider is

$$(3.3) \quad \Omega = \{B > 0, \delta > 0, \eta > -1, y > -1, \varepsilon > 0\}.$$

The main assumptions about  $\mathcal{A}$  are

$$(3.4) \quad \begin{aligned} (A) & \quad \mathcal{A}(y) > 0, & y > -1, \\ (B) & \quad \mathcal{A}'(y) < 0, & y > -1, \\ (C) & \quad \mathcal{A}''(y) > 0, & y > -1, \\ (D) & \quad \{y, \mathcal{A}\} \equiv \frac{2\mathcal{A}'\mathcal{A}''' - 3(\mathcal{A}'')^2}{(\mathcal{A}')^2} < 0, & y > -1. \end{aligned}$$

The expression  $\{y, \mathcal{A}\}$  is called the *Schwarzian derivative* of  $\mathcal{A}$  and has been useful in

projective geometry. (See E. Cartan [2].) We mention one fact; namely  $\{y, g\} = 0$  if  $g$  is a fractional linear transformation; that is,  $g(y) = (ay + b)/(cy + d)$ .

A calculation shows that for (3.2) with  $y > -1$

- (a)  $\mathcal{A} > 0,$
- (b)  $\mathcal{A}'(y) = \frac{-\gamma}{(1+y)^2} \mathcal{A}(y) < 0,$
- (c)  $\mathcal{A}''(y) = \frac{\gamma + 2 + 2y}{(1+y)^4} \mathcal{A}(y) > 0,$
- (d)  $\{y, \mathcal{A}\} = \frac{-\gamma^2}{(1+y)^4} < 0.$

So the assumptions (A–D) are indeed satisfied for the usual Arrhenius terms.

*Remark 3.5.* Equation (d) shows that  $\{y, \mathcal{A}\} = -(\mathcal{A}'/\mathcal{A})^2$ . We claim that this differential equation along with the boundary conditions  $\mathcal{A}(0) = 1$  and  $\mathcal{A}(-1) = +\infty$  uniquely define the Arrhenius terms up to  $\gamma$ . For assume  $\mathcal{A} = e^g$  then a computation yields  $\{y, \mathcal{A}\} = -(g')^2 + \{y, g\}$ . As  $\mathcal{A}'/\mathcal{A} = g'$  we see that  $\{y, g\} = 0$ . As noted above this implies that  $g$  is fractional linear; the boundary conditions yield the claim.

Before stating our main results we need two lemmas.

**LEMMA 3.6.** *Let  $\mathcal{A}$  satisfy (B), (C), and (D). Then there exists a unique point  $y_0 > 0$  such that  $\nu = y\mathcal{A}'' + \mathcal{A}' = 0$ .*

*Proof.* Observe that solutions to  $\nu = 0$  are obtained as intersections of the two functions  $f(y) = -y$  and  $g(y) = y + 2\mathcal{A}'/\mathcal{A}''$ . Assumption (D) shows that  $g$  is monotone increasing while  $f$  is clearly monotone decreasing to  $-\infty$ . As  $f(0) = 0$  and  $g(0) = 2\mathcal{A}'(0)/\mathcal{A}''(0) < 0$  the result is proved.

*Remark 3.7.* For (3.2)  $y_0 = \sqrt{1 + (\gamma^2/4)} - (\gamma/2)$ .

**LEMMA 3.8.** *Let  $\mathcal{A}$  satisfy (A–D). Then there is at most one point  $y_z$  such that  $(\mathcal{A}^{-1})'' = 0$ .*

*Proof.* Let  $\mathcal{F} = 2(\mathcal{A}')^2 - \mathcal{A}\mathcal{A}''$ , then  $(\mathcal{A}^{-1})'' = \mathcal{F}/\mathcal{A}^3$  and we need only find points where  $\mathcal{F} = 0$ . Consider the following identity:

$$(3.9) \quad 2\mathcal{F}' = -\mathcal{A}'\mathcal{A}\{y, \mathcal{A}\} + 3\mathcal{A}''\mathcal{F}/\mathcal{A}'$$

and observe that if  $\mathcal{F} \cong 0$ , then  $\mathcal{F}' < 0$ . This proves the lemma.

*Note.* If  $\mathcal{F} > 0$  for all  $y$ , let  $y_z = +\infty$  and if  $\mathcal{F} < 0$  for all  $y$ , let  $y_z = -1$ .

*Remark.* For (3.2),  $y_z = (\gamma/2) - 1$ .

The following list of derivatives of (3.1) will be needed for subsequent computations.

**LEMMA 3.10.** *Let  $\Delta = 1 + \varepsilon\delta\mathcal{A}(y)$ . Then*

- (i)  $G = \eta - (1 + \varepsilon)y + \frac{B\varepsilon}{\Delta};$
- (ii)  $G_\varepsilon = -y + B/\Delta^2;$
- (iii)  $G_y = -(1 + \varepsilon) - B\varepsilon^2\delta\mathcal{A}'/\Delta^2;$
- (iv)  $G_{\varepsilon\varepsilon} = -2B\delta\mathcal{A}/\Delta^3;$
- (v)  $G_{y\varepsilon} = -1 - 2B\varepsilon\delta\mathcal{A}'/\Delta^3;$
- (vi)  $G_{yy} = B\varepsilon^2\delta Q/\Delta^3$  where  $Q = 2\varepsilon\delta(\mathcal{A}')^2 - \Delta\mathcal{A}'';$

and

$$(vii) \quad G_{yyy}|_{Q=0} = B\epsilon^2\delta(3\epsilon\delta\mathcal{A}'\mathcal{A}'' - \Delta\mathcal{A}''')/\Delta^3.$$

PROPOSITION 3.11. *There exists at most one point  $(y_0, \epsilon_0, B_0, \delta_0, \eta_0) \in \Omega$  where the bifurcation problem  $G(y, \epsilon, B_0, \delta_0, \eta_0) = 0$  is contact equivalent to the winged cusp on a neighborhood of  $(y_0, \epsilon_0)$ .*

*Moreover if the following assumptions are made on  $\mathcal{A}$  then such a point actually exists in  $\Omega$ .*

- (E)  $(\ln \mathcal{A})''(y_0) > 0,$
- (F)  $y_z > y_0,$  i.e.,  $\mathcal{F}(y_0) > 0$  (Lemma 3.8),
- (G)  $y_0^2 + y_0 + \mathcal{A}(y_0)/\mathcal{A}'(y_0) < 0.$

*In fact, these points may be computed as follows:*

(i)  $y_0$  as in Lemma 3.6,

$$(ii) \quad \epsilon_0 = -\frac{\mathcal{A} + 2y_0\mathcal{A}'}{\mathcal{A} + y_0\mathcal{A}'} \Big|_{y=y_0},$$

$$(iii) \quad \eta_0 = -y_0(1 + \epsilon_0),$$

$$(iv) \quad \delta_0 = \frac{-1}{(\mathcal{A} + 2y_0\mathcal{A}')\epsilon_0} \Big|_{y=y_0},$$

$$(v) \quad B_0 = y_0\Delta_0^2.$$

REMARK 3.12. Assumptions (E) and (G) are satisfied for all Arrhenius terms while (F) is satisfied when  $\gamma > 8/3$ . As one is really interested in  $\gamma$  large—say of the order of 10—this is a reasonable hypothesis.

*Proof.* Proposition 2.7 states that to prove this proposition one must show that there is a unique choice of  $\beta_0, \delta_0, \eta_0$  yielding a unique solution  $(y_0, \epsilon_0)$  to the equation (2.8a) with  $G, y, \epsilon$  replacing  $H, x, \lambda$ . Observe that the equations  $G_{yy} = G_{y\epsilon} = 0$  imply that

$$(3.13a) \quad w(2(\mathcal{A}')^2 - \mathcal{A}\mathcal{A}'') = \mathcal{A}'',$$

$$(3.13b) \quad w(\mathcal{A} + 2y\mathcal{A}') = -1,$$

where  $w = \epsilon\delta$ . (Here one substitutes  $B = y\Delta^2$  into  $G_{y\epsilon}$ .) Thus  $\mathcal{A}''(b) + (a)$  implies

$$(3.14) \quad 2w\mathcal{A}'(\mathcal{A}' + y\mathcal{A}'') = 0.$$

As  $w\mathcal{A}' \neq 0$  in  $\Omega$  we have that  $y_0$  is given by Lemma 3.6. Next solve (3.13a) for

$$(3.15) \quad w = \mathcal{A}''/(2(\mathcal{A}')^2 - \mathcal{A}\mathcal{A}'') = \mathcal{A}''/F.$$

Hence  $w_0 = \epsilon_0\delta_0 > 0$  by (C) and (F). Next substitute  $B = y\Delta^2$  into  $G_y = 0$  to obtain

$$(3.16) \quad \epsilon_0 = -1/(1 + y_0w_0\mathcal{A}'(y_0)) = -\frac{\mathcal{A} + 2y_0\mathcal{A}'}{\mathcal{A} + y_0\mathcal{A}'} \Big|_{y=y_0}.$$

The last equality is obtained by solving (3.13b) for  $w_0$ . Recall that at  $y_0, \mathcal{A}'' = -\mathcal{A}'/y_0$ . Thus (E) implies that  $\mathcal{A}(y_0) + y_0\mathcal{A}'(y_0) > 0$  and (F) (or (3.13b)) implies  $\mathcal{A}(y_0) + 2y_0\mathcal{A}'(y_0) < 0$ . So  $\epsilon_0 > 0$  and  $\delta_0 > 0$ . Now  $B_0 = y_0\Delta_0^2 > 0$  where  $\Delta_0 = 1 + w_0\mathcal{A}(y_0)$  and from  $G = 0$

$$(3.17) \quad \eta_0 = (1 + \epsilon_0)y_0 - B_0\epsilon_0/\Delta_0 = -(1 + \epsilon_0)y_0.$$

The last equality is obtained as follows: from  $G_{\varepsilon y} = 0$  derive  $\Delta = -2y\mathcal{A}'$  and from  $G_y = 0$  derive  $\Delta = 2(\varepsilon + 1)/\varepsilon$ . Now use (3.17) and (3.16) to obtain

$$(3.18) \quad \eta_0 = \frac{y^2 \mathcal{A}'}{\mathcal{A} + y\mathcal{A}'} \Big|_{y=y_0}.$$

So  $\eta_0 > -1$  is equivalent to assumption (G).

To complete the proof of the proposition one shows that  $G_{\varepsilon\varepsilon} < 0$  and  $G_{yyy} < 0$  at  $(y_0, \varepsilon_0, B_0, \delta_0, \eta_0)$ . In fact  $G_{\varepsilon\varepsilon} < 0$  on  $\Omega$  by (A) and Lemma 3.10 (iv) and  $G_{yyy} < 0$  on  $\Omega \cap \{G_{yy} = 0\}$ . To obtain this last fact, note that Lemma 3.10 (vi) and (vii) imply that sign  $(G_{yyy})$  on  $\{G_{yy} = 0\}$  is just sign  $(3\varepsilon\delta\mathcal{A}'\mathcal{A}'' - \Delta\mathcal{A}''')$ . Now

$$(3.19) \quad 3\varepsilon\delta\mathcal{A}'\mathcal{A}'' - \Delta\mathcal{A}''' = -\frac{\varepsilon\delta(\mathcal{A}')^3}{\mathcal{A}''} \{y, \mathcal{A}\} < 0 \quad \text{on } G_{yy} = 0$$

as  $\Delta = 2\varepsilon\delta(\mathcal{A}')^2/\mathcal{A}''$  on  $G_{yy} = 0$ .

For the following we need one more assumption.

$$(H) \quad \frac{(2y\mathcal{A}''' + 3\mathcal{A}''')}{3} + \nu \left[ \frac{1 + \sqrt{1 + 8y\nu/\mathcal{A}}}{y} + \frac{2\mathcal{A}'}{\mathcal{A}} \right] > 0 \quad \text{on } [y_0, y_z].$$

We shall show in the appendix that this inequality is satisfied for Arrhenius terms with  $\gamma > 2$ . Note that at  $y_0, \nu = 0$  and the first term of (H) is positive by the Schwarzian condition.

**PROPOSITION 3.20.** *Under the assumptions (A)–(H) the only local bifurcation problems which occur in  $\Omega$  are those which appear in the universal unfolding for the winged cusp.*

*Proof.* Proposition 2.37 states that Proposition 3.20 is true if none of the following systems of equations is ever satisfied in  $\Omega$ .

$$(3.21) \quad G_{\varepsilon\varepsilon} = 0,$$

$$(3.22) \quad G = G_y = G_{yy} = G_{yyy} = 0,$$

$$(3.23) \quad G = G_y = G_\varepsilon = \det(d^2G) = d^3G(v, v, v) = 0,$$

where  $v \neq 0$  and  $(d^2G)(v, v) = 0$ .

At the end of the proof of Proposition 3.11 we showed that (3.21) and (3.22) are never satisfied in  $\Omega$ . To analyze (3.23) we need a preliminary result.

As  $G_{\varepsilon\varepsilon}$  is never zero we may solve implicitly  $G_\varepsilon(y, \varepsilon) = 0$  uniquely for  $\varepsilon = \varepsilon(y)$ . Let  $f(y) = G(y, \varepsilon(y))$ .

**LEMMA 3.24.** *The equations (3.23) are equivalent to the following system of equations:*

$$(3.25) \quad f = f' = f'' = f''' = 0.$$

*Proof.* Observe that

$$(3.26a) \quad f'(y) = G_y(y, \varepsilon(y)),$$

$$(3.26b) \quad f'' = G_{y\varepsilon}\varepsilon' + G_{yyy},$$

$$(3.26c) \quad f''' = G_{y\varepsilon\varepsilon}(\varepsilon')^2 + 2G_{yy\varepsilon}\varepsilon' + G_{yyy} + G_{y\varepsilon}\varepsilon''.$$

By (a) we have that  $f = f' = 0$  if  $G = G_y = G_\varepsilon = 0$ . Next differentiate the defining equation  $G_\varepsilon = 0$  to obtain

$$(3.27) \quad \varepsilon' = -G_{y\varepsilon}/G_{\varepsilon\varepsilon}.$$

Thus  $f'' = 0$  if  $\det d^2G = 0$ . Differentiating  $G_\varepsilon = 0$  a second time yields

$$(3.28) \quad \varepsilon'' = -(G_{yy\varepsilon} + 2G_{y\varepsilon\varepsilon}\varepsilon' + G_{\varepsilon\varepsilon\varepsilon}(\varepsilon')^2)/G_{\varepsilon\varepsilon}.$$

Substituting (3.28) and (3.27) into (3.26c) yields

$$(3.29) \quad f''' = (d^3G)(v, v, v)$$

by application of (2.18). This proves the lemma.

Of course one may use Lemma 3.10 (ii) to solve for  $\varepsilon(y)$  explicitly, obtaining

$$(3.30) \quad \varepsilon(y) = (\sqrt{B} - \sqrt{y})/\delta\mathcal{A}\sqrt{y}.$$

Thus we may take

$$(3.31) \quad f(y) = \mathcal{A}G(y, \varepsilon(y)) = (\eta - y)\mathcal{A} + (\sqrt{y} - \beta)^2/\delta,$$

where  $\beta = \sqrt{B}$ . Since  $\mathcal{A}$  is never zero on  $\Omega$  we still maintain the equivalence of (3.25) with (3.23).

To complete the proof of Proposition 3.20 we must show that (3.25) is never satisfied on  $\Omega$ . A computation shows that

$$(3.32a) \quad f'(y) = (\eta - y)\mathcal{A}' - \mathcal{A} + (\sqrt{y} - \beta)/\delta\sqrt{y},$$

$$(3.32b) \quad f''(y) = (\eta - y)\mathcal{A}'' - 2\mathcal{A}' + \beta/(2\delta y^{3/2}),$$

$$(3.32c) \quad f'''(y) = (\eta - y)\mathcal{A}''' - 3\mathcal{A}'' - 3\beta/(4\delta y^{5/2}).$$

We use the following notation:

$$(3.33) \quad \nu = \mathcal{A}' + y\mathcal{A}'', \quad \tau = 3\mathcal{A}'' + 2y\mathcal{A}''', \quad \mathcal{S} = 2\mathcal{A}'\mathcal{A}''' - 3(\mathcal{A}'')^2;$$

and make the following observations at a solution to (3.25):

$$(3.34a) \quad \eta - y < 0,$$

$$(3.34b) \quad \eta - y = 6\nu/\tau,$$

$$(3.34c) \quad \beta/\delta = 4\mathcal{S}y^{5/2}/\tau,$$

$$(3.34d) \quad \tau < 0,$$

$$(3.34e) \quad \nu > 0,$$

$$(3.34f) \quad \frac{1}{\delta} = \mathcal{A} - 6\nu\mathcal{A}'/\tau + 4y^2\mathcal{S}/\tau.$$

It is clear from (3.31) that to solve  $f = 0$  implies (3.34a). Equations (3.34b) and (3.34c) are obtained from (3.32b) and (3.32c). So (3.34d) follows from (3.34c) as  $\beta/\delta > 0$  and  $\mathcal{S} < 0$ . Now (3.34e) follows from (3.34b). Finally (3.34f) is obtained from (3.32a).

Substitution of this data into (3.31) yields

$$(3.35) \quad (\mathcal{A}\tau - 6\nu\mathcal{A}' + 4y^2\mathcal{S})6\nu\mathcal{A} + y(\tau\mathcal{A} - 6\nu\mathcal{A}')^2 = 0,$$

which is an equation in  $y$  alone. Letting

$$(3.36) \quad w = \tau/6\nu$$

we obtain from (3.35), noting that  $y\mathcal{S} = \mathcal{A}'\tau - 3\mathcal{A}''\nu$ ,

$$(3.37) \quad \left(w + \frac{\mathcal{A}'}{\mathcal{A}}\right)^2 + \frac{1}{y}\left(w + \frac{\mathcal{A}'}{\mathcal{A}}\right) - \frac{2\nu}{y\mathcal{A}} = 0.$$

As  $w < 0$  (by (3.34a) and (3.34b)),  $\mathcal{A}'/\mathcal{A} < 0$  by (A) and (B) and  $\nu > 0$  (by (3.34e)) we have

$$(3.38) \quad w = -(1 + \sqrt{1 + 8y\nu/\mathcal{A}})/2y - \mathcal{A}'/\mathcal{A}.$$

Hence

$$(3.39) \quad \tau/3 + \nu \left[ \frac{(1 + \sqrt{1 + 8y\nu/\mathcal{A}})}{y} + \frac{2\mathcal{A}'}{\mathcal{A}} \right] = 0.$$

Let  $y_f$  be a solution to (3.39) satisfying (3.34). In particular  $\nu(y_f) > 0$  implies by Lemma 3.6 that  $y_f > y_0$ . For  $y_f$  to generate a solution to (3.25) in  $\Omega$  it must also satisfy

$$(3.40) \quad 0 < \varepsilon(y_f).$$

We claim that (3.40) implies that  $y_f < y_z$  thus proving the proposition. In particular (3.30) and (3.40) together imply that  $\beta > \sqrt{y}$ . Using (3.34c) and (3.34f) one obtains

$$(3.41) \quad 1 + \frac{\tau\mathcal{A} - 6\nu\mathcal{A}'}{4\mathcal{S}y^2} < 1,$$

which holds only if

$$(3.42) \quad \tau\mathcal{A} - 6\nu\mathcal{A}' > 0$$

as  $\mathcal{S} < 0$ . Upon expanding  $\tau$  we obtain

$$(3.43) \quad -\frac{2}{3}y\mathcal{A}''' < \mathcal{A}'' - 2\nu\mathcal{A}'/\mathcal{A}.$$

Substituting this inequality in (3.39) implies

$$(3.44) \quad \sqrt{1 + 8y\nu/\mathcal{A}} < -\frac{\mathcal{A} + 4y\mathcal{A}'}{\mathcal{A}}.$$

If  $(\mathcal{A} + 4y\mathcal{A}')|_{y_f}$  is positive then  $y_f$  does not correspond to a solution to (3.35) in  $\Omega$ . So assume that it is negative and square (3.44) to obtain

$$(3.45) \quad 2(\mathcal{A}')^2 - \mathcal{A}\mathcal{A}''|_{y_f} = \mathcal{F}(y_f) > 0.$$

As  $\mathcal{F}(y) < 0$  for  $y \geq y_z$  by Lemma 3.8 we have that  $y_0 < y_f < y_z$ .

Then by (H) the proposition is proved.

We now state and prove the main result of this section. In particular this result is satisfied for Arrhenius terms when  $\gamma > 8/3$ .

**THEOREM 3.46.** *Let*

$$G(y, \varepsilon, B, \delta, \eta) = \eta - (1 + \varepsilon)y + B\varepsilon/\Delta,$$

where  $\Delta = 1 + \varepsilon\delta\mathcal{A}$  and  $\mathcal{A}$  satisfies the conditions (A)–(H). Then there exists a unique winged cusp point in  $\Omega$  and for every  $(y', \varepsilon', B', \delta', \eta')$  the bifurcation problem  $G(y, \varepsilon, B', \delta', \eta') = 0$  is contact equivalent to a bifurcation problem contained in the universal unfolding of the winged cusp point. Moreover  $B, \delta,$  and  $\eta$  form universal unfolding parameters for any such bifurcation problem.

*Proof.* The first two statements are the results of Propositions 3.11 and 3.20. The proof of the last statement uses Proposition 2.27. In fact, it is sufficient to show that

$$(a) \quad G_{\delta y}G_{B\varepsilon} - G_{By}G_{\delta\varepsilon} \neq 0 \quad \text{on } \Omega;$$

$$(b) \quad \text{rank} \begin{pmatrix} G_{By} & G_{\delta y} & G_{y\varepsilon} \end{pmatrix} = 1 \quad \text{on } \Omega;$$

$$(c) \quad \text{rank} \begin{pmatrix} G_{By} & G_{Byy} - G_{B\epsilon} \frac{G_{yyy}}{G_{y\epsilon}} \\ G_{\delta y} & G_{\delta yy} - G_{\delta\epsilon} \frac{G_{yyy}}{G_{y\epsilon}} \\ G_{y\epsilon} & G_{yy\epsilon} - G_{\epsilon\epsilon} \frac{G_{yyy}}{G_{y\epsilon}} \end{pmatrix} = 2;$$

at points where  $G_y = G_{yy} = 0$  and

$$(d) \quad \det \begin{pmatrix} G_{By} & G_{B\epsilon} & G_{By\epsilon} - \frac{G_{yy\epsilon}}{G_{yyy}} G_{Byy} \\ G_{\delta y} & G_{\delta\epsilon} & G_{\delta y\epsilon} - \frac{G_{yy\epsilon}}{G_{yyy}} G_{\delta yy} \\ 0 & G_{\epsilon\epsilon} & G_{y\epsilon\epsilon} - \frac{G_{yy\epsilon}}{G_{yyy}} G_{yy\epsilon} \end{pmatrix} \neq 0$$

at the winged cusp point.

One calculates:

$$(3.47) \quad \begin{aligned} (i) \quad & G_{By} = -\epsilon^2 \delta \mathcal{A}' / \Delta^2; \\ (ii) \quad & G_{B\epsilon} = 1 / \Delta^2; \\ (iii) \quad & G_{\delta y} = B \epsilon^2 \mathcal{A}' (\epsilon \delta \mathcal{A} - 1) / \Delta^3; \\ (iv) \quad & G_{\delta\epsilon} = -2 B \epsilon \mathcal{A} / \Delta^3. \end{aligned}$$

Thus

$$(3.48) \quad G_{\delta y} G_{B\epsilon} - G_{By} G_{\delta\epsilon} = -B \epsilon^2 \mathcal{A}' / \Delta^4 > 0.$$

So (a) is satisfied. Since  $G_{By} > 0$  on  $\Omega$  by (ii) (b) is also satisfied.

To show that (d) holds observe that if a function  $f(\epsilon, \delta)$  has the form  $g(\epsilon\delta)$  then  $\epsilon f_\epsilon \equiv \delta f_\delta$  so that  $f_\epsilon / f_\delta = \delta / \epsilon$ . Observe—using Lemma 3.10—that this is the case for  $G_\epsilon$ ,  $G_{y\epsilon}$ , and  $Q$ . Also note that  $G_{yy\epsilon} / G_{yyy} = Q_\epsilon / Q_\delta$  when  $Q = 0$ . Thus (d) holds if

$$(3.49) \quad G_{\delta y} \det \begin{pmatrix} G_{B\epsilon} & G_{By\epsilon} - \frac{G_{yy\epsilon}}{G_{yyy}} G_{Byy} \\ G_{\epsilon\epsilon} & G_{y\epsilon\epsilon} - \frac{G_{yy\epsilon}}{G_{yyy}} G_{yy\epsilon} \end{pmatrix} \neq 0.$$

Recall from (3.13) that

$$(3.50) \quad \epsilon \delta \mathcal{A} - 1 = -2 \frac{\mathcal{A} + y \mathcal{A}'}{\mathcal{A} + 2y \mathcal{A}'} = 2 / \epsilon_0 > 0$$

at the winged cusp point. Now note that  $G_{yy} = 0$  when  $Q = 0$  and that  $G_{yyB} = \epsilon^2 \delta Q / \Delta^3 = 0$ . So we need only evaluate

$$(3.51) \quad \det \begin{pmatrix} G_{B\epsilon} & G_{By\epsilon} \\ G_{\epsilon\epsilon} & G_{y\epsilon\epsilon} - G_{yy\epsilon}^2 / G_{yyy} \end{pmatrix} = -\frac{B \delta}{\Delta^5} \left( \frac{2 \mathcal{A}'}{\Delta} + \frac{(\mathcal{A}'')^2}{Q_y} \right) > 0$$

since  $\mathcal{A}' < 0$  by (B) and  $Q_y < 0$  by (3.19).



To complete the proof of the Theorem we must verify (c). Now note that (c) holds if

$$(3.52) \quad \det \begin{pmatrix} G_{By} & G_{B\epsilon} & G_{Byy} \\ G_{\delta y} & G_{\delta\epsilon} & G_{\delta yy} \\ G_{y\epsilon} & G_{\epsilon\epsilon} & G_{yy\epsilon} \end{pmatrix} \neq 0.$$

This is a sufficient though not necessary condition. Recall that  $G_{yy} = B\epsilon^2 \delta Q / \Delta^3$  and that  $Q = 0$  iff  $G_{yy} = 0$ . Hence  $G_{yyB} = 0$  when  $G_{yy} = 0$ . Now using the same observation as in the proof of (d) that  $Q$  and  $G_\epsilon$  are functions of  $\epsilon\delta$  we see that the rank of

$$\begin{pmatrix} G_{\delta\epsilon} & G_{yy\delta} \\ G_{\epsilon\epsilon} & G_{yy\epsilon} \end{pmatrix} \text{ is } 1.$$

Thus we need only compute

$$(3.53) \quad \det \begin{pmatrix} G_{By} & G_{B\epsilon} & 0 \\ G_{\delta y} & 0 & G_{yy\delta} \\ G_{y\epsilon} & 0 & G_{yy\epsilon} \end{pmatrix}.$$

Note that  $G_{B\epsilon} = 1/\Delta^2 \neq 0$  so this computation reduced to showing

$$(3.54) \quad D = \det \begin{pmatrix} G_{y\delta} & G_{yy\delta} \\ G_{y\epsilon} & G_{yy\epsilon} \end{pmatrix} \neq 0.$$

Now

$$(3.55) \quad D = \frac{B\epsilon^2 \delta}{\Delta^3} \mathcal{F} \det \begin{pmatrix} G_{y\delta} & \epsilon \\ G_{y\epsilon} & \delta \end{pmatrix},$$

where  $\mathcal{F} = 2(\mathcal{A}')^2 - \mathcal{A}\mathcal{A}''$ . Observe that  $Q = \epsilon\delta\mathcal{F} - \mathcal{A}''$  so that  $\mathcal{F} \neq 0$  when  $Q = 0$ . The problem is reduced to computing

$$(3.56) \quad \delta G_{y\delta} - \epsilon G_{y\epsilon} = \epsilon + B\epsilon^2 \delta \mathcal{A}' / \Delta^2 = -1.$$

This last equality is obtained from  $G_y = 0$ .

**Appendix A.** We now sketch a proof of:

PROPOSITION A.1. Condition (H) is satisfied for the Arrhenius terms for all  $\gamma > 2$ .

To prove this proposition we need to show that (3.39) has no solutions on  $[y_0, y_z]$ . From the derivation of (3.39) this is equivalent to showing that (3.37) has no solutions on  $[y_0, y_z]$  when  $\tau < 0$ . This is our approach.

Note that if  $\mathcal{A} = e^g$  then

$$(A.2) \quad \frac{\mathcal{A}'}{\mathcal{A}} = g', \quad \frac{\mathcal{A}''}{\mathcal{A}} = (g')^2 + g'' \quad \text{and} \quad \frac{\mathcal{A}'''}{\mathcal{A}} = (g')^3 + 3g'g'' + g'''.$$

For the Arrhenius terms  $g(y) = -\gamma y / (1 + y)$ . Thus

$$(A.3) \quad \begin{aligned} \frac{\mathcal{A}'}{\mathcal{A}} &= \frac{-\gamma}{(1+y)^2}, & \frac{\mathcal{A}''}{\mathcal{A}} &= \frac{\gamma}{(1+y)^4} [2y + \gamma + 2], \\ \frac{\mathcal{A}'''}{\mathcal{A}} &= \frac{-\gamma}{(1+y)^6} [6y^2 + (12 + 6\gamma)y + 6 + 6\gamma + \gamma^2]. \end{aligned}$$

Recall that  $\tau = 3\mathcal{A}'' + 2y\mathcal{A}'''$ , so

$$(A.4) \quad \frac{\tau}{\mathcal{A}} = -\frac{6\gamma}{(1+y)^6} \left[ y^3 + (1 + \frac{3}{2}\gamma)y^2 + \left(\frac{\gamma^2}{3} - \gamma - 1\right)y - \left(\frac{\gamma}{2} + 1\right) \right].$$

Since  $\nu = \mathcal{A}' + y\mathcal{A}'' > 0$  on  $[y_0, y_z]$  we may compute (3.37) in the form

$$(A.5) \quad \left(\frac{\tau}{6\mathcal{A}} + \frac{\mathcal{A}'}{\mathcal{A}} \frac{\nu}{\mathcal{A}}\right) \left(y \left(\frac{\tau}{6\mathcal{A}} + \frac{\mathcal{A}'}{\mathcal{A}} \frac{\nu}{\mathcal{A}}\right) + \frac{\nu}{\mathcal{A}}\right) - 2\left(\frac{\nu}{\mathcal{A}}\right)^3.$$

Compute

$$(A.6) \quad \frac{\nu}{\mathcal{A}} = \frac{\gamma}{(1+y)^4} (y^2 + \gamma y - 1).$$

Then (A.5) is given by

$$(A.7) \quad \frac{\gamma^2}{(1+y)^{12}} \left[ -\left(\frac{\gamma}{2} + 1\right)y^6 - \left(\frac{11}{12}\gamma^2 + 2\gamma + 2\right)y^5 - \left(\frac{2}{3}\gamma^3 + \gamma^2 + \frac{5}{2}\gamma - 1\right)y^4 \right. \\ \left. - \left(\frac{2}{9}\gamma^4 + \frac{\gamma^2}{6} - 4\right)y^3 + \left(\frac{2}{3}\gamma^3 - \gamma^2 + \frac{5}{2}\gamma + 1\right)y^2 - \left(\frac{11}{12}\gamma^2 - 2\gamma + 2\right)y + \left(\frac{\gamma}{2} - 1\right) \right]$$

Letting  $y = (K/\gamma)$  we now show:

LEMMA A.8. *Expression (A.7) < 0 for all  $K > 1.2$ .*

LEMMA A.9.  *$\tau > 0$  for all  $K \leq 1.2$  when  $\gamma \geq 2$ .*

These two lemmas together prove Proposition A.1. Substituting for  $y$  in (A.7) and grouping terms by powers of  $\gamma$  yields:

$$(A.10) \quad c_1\gamma + c_2 + \frac{c_3}{\gamma} + \frac{c_4}{\gamma^2} + \frac{c_5}{\gamma^3} + \frac{c_6}{\gamma^4} + \frac{c_7}{\gamma^5} + \frac{c_8}{\gamma^6},$$

where

$$(A.11) \quad \begin{aligned} c_1 &= -\left(\frac{2K^3}{9} - \frac{2K^2}{3} + \frac{11K}{12} - \frac{1}{2}\right), \\ c_2 &= -(K-1)^2, \\ c_3 &= -\left(\frac{2}{3}K^4 + \frac{K^3}{6} - \frac{5K^2}{2} + 2K\right), \\ c_4 &= -(K^4 - K^2), \\ c_5 &= -\left(\frac{11}{12}K^5 + \frac{5}{2}K^4 - 4K^3\right), \\ c_6 &= -(2K^5 - K^4), \\ c_7 &= -\left(\frac{K^6}{2} + 2K^5\right), \\ c_8 &= -K^6. \end{aligned}$$

We make the following observations:

$$\begin{aligned}
 (A.12) \quad & c_1 < 0 \quad \text{for } K > 1.2, \\
 & c_2 < 0 \quad \text{for all } K, \\
 & c_3 < 0 \quad \text{for } K > 1.2, \\
 & c_4 < 0 \quad \text{for } K > 1, \\
 & c_5 < 0 \quad \text{for } K > 1.2, \\
 & c_6 < 0 \quad \text{for } K > 1/2, \\
 & c_7 < 0 \quad \text{for } K > 0, \\
 & c_8 < 0 \quad \text{for all } K.
 \end{aligned}$$

This proves Lemma A.8.

Next we compute  $\tau(K/\gamma)$ —grouped in powers of  $K$ —obtaining

$$(A.13) \quad -\frac{K^3}{\gamma^3} - \left(\frac{1}{\gamma^2} + \frac{3}{2\gamma}\right) K^2 - \left(\frac{\gamma}{3} - 1 - \frac{1}{\gamma}\right) K + \left(\frac{\gamma}{2} + 1\right).$$

Note that for any positive  $\gamma$  (A.13) has at most one positive root by Descartes' rule of signs. Since  $\tau(0) > 0$  and  $\tau < 0$  for large  $K$ , (A.13) has exactly one positive root. So if we evaluate (A.13) at  $K = 1.2$  and obtain a positive number then Lemma A.9 is proved. This evaluation yields,

$$(A.14) \quad \frac{1}{\gamma^3} (.1\gamma^4 + 2.2\gamma^3 - .96\gamma^2 - 1.44\gamma - 1.728).$$

Again by Descartes' rule of signs (A.14) has one positive root. Since (A.14) evaluated at  $\gamma = 0$  is  $< 0$  and at  $\gamma = 2$  is 10.752, Lemma A.9 is proved and Proposition A.1 follows.

#### REFERENCES

- [1] R. ARIS, *Num in olla agitata papilio est? or, Catastrophes and chemical reactors*, Catastrophes and Other Important Matters, University of Minnesota, Minneapolis, MN, 1977.
- [2] E. CARTAN, *Leçons sur la Théorie des Espaces à Connexions Projective*, Cahiers Scientifiques Fascicule XVII, Gauthier-Villars, ed., Paris, 1937.
- [3] M. CRANDALL AND P. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Functional Analysis 8 (1971), pp. 321–340.
- [4] M. GOLUBITSKY AND D. SCHAEFFER, *A theory for imperfect bifurcation via singularity theory*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.
- [5] W. H. RAY, *Bifurcation phenomena in chemical reacting systems*, Applications of Bifurcation Theory, P. Rabinowitz, ed., Academic Press, New York, 1977, pp. 285–316.
- [6] A. UPPAL, W. H. RAY AND A. B. POORE, *On the dynamic behavior of continuous stirred tank reactors*, Chem. Eng. Sci., 29 (1974), pp. 967–985.
- [7] ———, *The classification of the dynamic behavior of continuous stirred tank reactors—influence of reactor residence time*, Chem. Eng. Sci., 31 (1976), pp. 205–214.

**POROUS MEDIA PROBLEMS\***

WAYNE T. FORD†, MARIA C. FUENTE‡ AND MARGARET C. WAID§

**Abstract.** Laminar isothermal fluid flow of two immiscible compressible fluid phases in a porous medium is formulated in terms of four unknown functions  $\rho_1, \rho_2, S_1$  and  $S_2$  in a pair of partial differential equations

$$\frac{\partial}{\partial t}[\phi(x, t)S_i\rho_i] = \frac{\partial}{\partial x}\left\{\kappa(x, t)\sigma_i(S_i)\frac{\partial}{\partial x}[\Phi_i(\rho_i)]\right\}$$

and a pair of auxiliary relations

$$S_1 = \Gamma_1(\rho_1, \rho_2) \quad \text{and} \quad S_2 = 1 - S_1.$$

The first boundary value problem is reformulated for this system as a fixed point problem involving a mapping  $\Pi$  wherein  $S_1$  implies  $S_2$ , the differential equations are used to find  $\rho_1$  and  $\rho_2$ , and  $\Pi(S_1)$  is set equal to  $\Gamma_1(\rho_1, \rho_2)$ . The mapping  $\Pi$  is shown to map a subset of  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  into  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  under appropriate conditions on the coefficients and equations.

**1. Introduction.** Source-free laminar isothermal flow of two immiscible compressible fluid phases in a linear horizontal porous medium can be described [9], [11], [14], [40] as a mathematical problem (to be physically motivated in § 2) involving four unknown functions  $\rho_1, \rho_2, S_1$ , and  $S_2$  in two partial differential equations

$$(1.1) \quad \frac{\partial}{\partial t}[\phi(x, t)S_i\rho_i] = \frac{\partial}{\partial x}\left\{\kappa(x, t)\sigma_i(S_i)\frac{\partial}{\partial x}[\Phi_i(\rho_i)]\right\}$$

and two auxiliary relations

$$(1.2) \quad S_i = \Gamma_i(\rho_1, \rho_2).$$

We adopt the convention, illustrated above, that every usage of the subscript  $i$  is assumed to apply to both values,  $i = 1$  and  $i = 2$ . Moreover, the subscript will be suppressed in discussions which apply equally to both subscripts. For example,  $\Phi$  will be defined (in § 2) as a monotone increasing map of  $E = (-\infty, \infty)$  onto itself. Also,  $\sigma$  will be defined as a map from  $(0, 1)$  into  $E^+ = (0, \infty)$ , and  $\Gamma$  will be given as a map from  $E^2$  to  $(0, 1)$ .

The coefficients,  $\phi$  and  $\kappa$ , will be given maps of  $\bar{\Omega}$  into  $E^+$ , where  $\Omega = (0, 1) \times (0, T)$ , and we consider (1.1) and (1.2) in terms of the first boundary problem wherein solutions are sought for  $(x, t)$  in  $\Omega$  subject to appropriate initial and boundary conditions on  $(0, 1) \times \{0\}$  and  $\{0, 1\} \times [0, T]$ , respectively. We let

$$(1.3) \quad \Omega_B = (0, 1) \times \{0\} \cup \{0, 1\} \times [0, T]$$

to choose combined initial-boundary conditions in the form

$$(1.4) \quad \rho(x, t) = \psi(x, t), \quad (x, t) \in \Omega_B.$$

We consider (1.1) through (1.4) in terms of solutions in the Banach space  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$ , for some  $\theta \in (0, 1)$ , with its norm,  $|\cdot|_{\bar{\Omega}}$ , written without the superscript,  $(\theta)$ . Since precise definitions are available in the literature [28, p. 7], it suffices here to

\* Received by the editors August 2, 1978, and in revised form May 29, 1979.

† Mathematics Department, Texas Tech University, Lubbock, Texas 79409. The work of this author was supported in part by Gulf Universities Research Consortium (GURC), Houston, Texas.

‡ Mathematics Department, Texas Tech University, Lubbock, Texas 79409. The work of this author was supported in part by a Texas Tech University Graduate School Summer Research Assistant Grant.

§ Mathematics Department, University of Delaware, Newark, Delaware 19711.

remark that  $u \in H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  implies that  $u_{xx}$  and  $u_t$  are Hölder continuous on  $\bar{\Omega}$  with exponent  $\theta$  in  $x$  and exponent  $\theta/2$  in  $t$ . Also, we define  $|\cdot|_0$  as the supremum of the magnitude of its argument over its domain. For example,  $|\sigma|_0$  is the supremum of  $|\sigma(S)|$  over  $(0, 1)$ , while  $|\kappa|_0$  and  $|\sigma \circ S|_0$  both represent suprema over  $\bar{\Omega}$ .

ASSUMPTION 1.1. *Each  $\psi$  belongs to  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$ , i.e.,  $\psi$  can be extended to a function that does so belong. Also,  $\psi_1$  and  $\psi_2$  substituted for  $\rho_1$  and  $\rho_2$ , satisfy (1.1) and (1.2) at the lower corners  $(0, 0)$  and  $(1, 0)$ , of  $\bar{\Omega}$ .*

Although (1.2) can be used to formally eliminate  $S_1$  and  $S_2$  from (1.1), our plan of study of (1.1) through (1.4) does not permit this. Specifically, we will consider formulation of (1.1) through (1.4) in terms of a fixed point of a mapping wherein (1.2) is used to produce improved  $S_1$  and  $S_2$  from solutions,  $\rho_1$  and  $\rho_2$ , of (1.1) and (1.4) based on estimates of  $S_1$  and  $S_2$ . It is convenient to rewrite (1.1) in the form

$$(1.5) \quad \frac{\partial}{\partial t}(M\rho) = \frac{\partial}{\partial x} \left\{ N \frac{\partial}{\partial x} [\Phi(\rho)] \right\},$$

where

$$(1.6) \quad M(x, t) = \phi(x, t)S(x, t) \quad \text{and} \quad N(x, t) = \kappa(x, t)\sigma[S(x, t)].$$

If only one phase were involved, (1.5) would describe the flow of that phase with  $S \equiv \sigma(S) \equiv 1$ . Thus, (1.5) can be thought of as a *porous medium equation* (PME), while (1.1) and (1.2) constitute a *porous medium system* (PMS). The fixed point approach allows consideration of the PMS in relation to its PME parts.

We reformulate the PME in several ways (in § 3) for completeness and for use in our discussion. One of these formulations can be used to show (in § 4) that (1.5) maps  $(S \rightarrow \rho)H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  into itself with  $S$  substituted in (1.6),  $\rho = \psi$  on  $\Omega_B$  satisfying Assumption 1.1, and appropriate assumptions of  $\phi$ ,  $\kappa$ ,  $\sigma$  and  $\Phi$  (§ 2). This leads to a fixed point formulation of the PMS wherein (1.1) and (1.4) produce the pair  $(\rho_1, \rho_2)$  from a given pair  $(S_1, S_2)$  so that (1.2) can then produce an improved pair  $(S_1, S_2)$  from this  $(\rho_1, \rho_2)$  pair. We show that the fixed point formulation is well-defined (in § 4).

We show that a solution of a PME satisfies a heat equation with low order terms. Thus, initial-boundary  $\psi > 0$  implies that the corresponding PME density is positive on  $\bar{\Omega}$ , and an a priori upper bound can be given (§ 3).

Even though (1.1) through (1.4) involve only two phases, the mathematical problem has not been studied in detail in so far as we know. Analysis is rather limited for complicated generalizations [9], [17], [18], which are often treated in a pragmatic numerical sense [26], [34], [40]. Equations rather like those in the PMS appear in fields other than petroleum engineering [7], [24], [26], [36], and singular and special cases have been studied [4], [5], [14], [16], [21], [27], [33].

Our bibliography is restricted to works consulted in the preparation of this paper. Certain authors are represented by a recent work only [4], [5], [6], [8], [14], [40], some of the entries are collections of papers [1], [13] by various authors, and some entries contain very substantial bibliographies [6], [20], [24], [36], [38].

**2. Physical motivation.** If (1.1) is rewritten in engineering terminology, the result is [6], [11]

$$(2.1) \quad \frac{\partial}{\partial t}(\phi S\rho) = \frac{\partial}{\partial x} \left( \kappa \sigma \rho \mu^{-1} \frac{\partial P}{\partial x} \right),$$

where each symbol represents a definite physical concept. For example,  $P$  represents the *pressure* in the individual phase, and the *saturation*  $S$  is the volume occupied by that

phase as a fraction of the total local available fluid volume, required to be fluid-filled by the equation

$$(2.2) \quad S_1 + S_2 = 1.$$

Darcy’s work (reprinted in [25]) dealt with a single phase incompressible fluid. His concepts of *porosity*  $\phi$  and *permeability*  $\kappa$  have been extended to more general cases. We take both  $\phi$  and  $\kappa$  to be given positive functions of  $(x, t)$  throughout  $\bar{\Omega}$ , independent of the fluids occupying the pore spaces of the medium. However, availability of flow channels to a particular fluid in a multiphase flow is taken to depend on the saturation in that phase so that each *relative permeability*  $\sigma$  is taken to be a known nondecreasing function of its own  $S$ .

Each *density*  $\rho$  and each (*dynamic*) *viscosity*  $\mu$  is given as a function of its individual phase pressure  $P$ . The expression of  $\rho$  in terms of  $P$  is known as an *equation of state*, which we take to be invertible so that we can define an additional quantity,  $\Phi$ , by the equations

$$(2.3) \quad \frac{d\Phi}{d\rho} = \frac{\rho}{\zeta(\rho)} \frac{dP}{d\rho} \quad \text{and} \quad \zeta(\rho) \equiv \mu[P(\rho)].$$

Thus, definition of the symbols in (2.1) is complete so that it can be identified with (1.1); while specification of the functions in (1.2) requires one final physical concept.

Immiscibility implies [11, p. 201] the existence of a *capillary pressure*  $P_c$ , presumed to be a known function of one saturation, say  $S_1$ , so that

$$(2.4) \quad P_1 - P_2 = P_c(S_1).$$

If this  $P_c$  maps  $(0, 1)$  onto  $E$  in a one-to-one fashion, it can be used to write

$$(2.5) \quad S_1 = P_c^{-1}(P_1 - P_2).$$

This relation is then used to define  $\Gamma_1$  in (1.2) by

$$(2.6a) \quad \Gamma_1(\rho_1, \rho_2) = P_c^{-1}[P_1(\rho_1) - P_2(\rho_2)],$$

and (2.2) is used in defining  $\Gamma_2$  by

$$(2.6b) \quad \Gamma_2(\rho_1, \rho_2) = 1 - \Gamma_1(\rho_1, \rho_2).$$

Although  $\Phi'$  could be given on  $E$ , we presume its accuracy to apply only on some closed interval  $J \subset E^+$ . Specifically, we take  $\Phi'$  to be given in  $C^3(J)$  so that it can be extended to  $E$  as an even function in  $C^3(E)$  such that

$$(2.7) \quad 0 < \Phi'(\rho) \leq |\Phi'|_0 \quad \text{for } \rho \in E, \quad \lim_{\rho \rightarrow \infty} \Phi'(\rho) = |\Phi'|_0,$$

$\Phi''$  is nonnegative for large values of  $\rho$ , and  $\Phi$ , the integral of  $\Phi'$ , is zero for zero  $\rho$ . Since we want (2.3) to apply on  $E$  with  $P' = dP/d\rho$  being positive, compatible extensions will be assumed so that  $P'$  and  $\zeta$  are even and odd, respectively.

We summarize the above discussion in the following specific mathematical assumptions:

ASSUMPTION 2.1 (Flows).  $\phi$  and  $\kappa$  belong to  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  with values in  $E^+$ .  $\sigma \in C^3[(0, 1)]$  with values in  $E^+$ .

ASSUMPTION 2.2 (States).  $\Phi' \in C^3(E)$  with values in  $E^+$ . It is even,  $\Phi''$  is nonnegative for large values of its argument, and (2.7) applies.

ASSUMPTION 2.3 (Interaction).  $\Gamma_1$  is defined on  $E^2$  with values in  $(0, 1)$ . If  $u_1$  and  $u_2$  belong to  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$ , then  $v(x, t) = \Gamma_1[u_1(x, t), u_2(x, t)]$  defines a function  $v$  in  $H^{2+\theta, 1+\theta/2}(\bar{\Omega})$ . Note that (2.6b) shows that  $\Gamma_2$  has the properties given above for  $\Gamma_1$ .

**3. Reformulations.** Since parabolic partial differential equations are often formulated in terms of solutions of the equations for the time derivatives of the dependent variables [9], [20], [28], [31], [32], [35], [41], [42], we present several transformations of (1.5) into such form for later discussion. It is convenient to introduce the definitions

$$(3.1a) \quad \alpha(x, t) = N(x, t)/M(x, t), \quad \beta(x, t) = \ln M(x, t),$$

and

$$(3.1b) \quad \gamma(x, t) = \alpha_x(x, t) + \alpha(x, t)\beta_x(x, t) = N_x(x, t)/M(x, t).$$

We use subscript differentiation and suppression of arguments to formally manipulate (1.5) in

$$(3.2) \quad \begin{aligned} (M\rho)_t &= M\rho_t + M_t\rho \\ &= (N\Phi'\rho_x)_x \\ &= (M \cdot M^{-1}N\Phi'\rho_x)_x \\ &= M(M^{-1}N\Phi'\rho_x)_x + M_xM^{-1}N\Phi'\rho_x \end{aligned}$$

to obtain the divergence form

$$(3.3) \quad \rho_t = (\alpha\Phi'\rho_x)_x + \alpha\beta_x\Phi'\rho_x - \beta_t\rho$$

which can be related to existing theory [31]. Similarly, the calculation,

$$(3.4) \quad \begin{aligned} \rho_t &= (\alpha\Phi'\rho_x)_x + (\alpha\beta_x)_x\Phi - (\alpha\beta_x)_x\Phi + \alpha\beta_x\Phi'\rho_x - \beta_t\rho \\ &= (\alpha\Phi'\rho_x + \alpha\beta_x\Phi)_x - (\alpha\beta_x)_x\Phi - \beta_t\rho, \end{aligned}$$

displays a special case for which an initial-boundary problem of the third kind has been studied [9].

Since a fully differentiated form is used in the literature [20], [28], [35], [41], [42], we rewrite (3.3) in the form

$$(3.5) \quad \rho_t = \alpha\Phi'\rho_{xx} + \alpha\Phi''\rho_x^2 + \gamma\Phi'\rho_x - \beta_t\rho.$$

The latter form, which will be used below (in § 4), motivates the definitions

$$(3.6a) \quad C(x, t, u) = \alpha(x, t)\Phi'(u)$$

and

$$(3.6b) \quad a(x, t, u, v) = -\alpha(x, t)\Phi''(u)v^2 - \gamma(x, t)\Phi'(u)v + \beta_t(x, t)u.$$

Additional forms, which can be used for understanding and for a priori estimates, can be obtained from the result below.

**THEOREM 3.1.** *Suppose  $\rho$  satisfies (1.5) with both  $M$  and  $N$  being positive and continuously differentiable on  $\bar{\Omega}$ . Then, a differentiable invertible coordinate transformation,  $(x, t) \rightarrow (\xi(x, t), \tau(t))$ , exists such that the  $(x, t)$  rectangle  $\Omega$  and the equation in (1.5) transform, respectively, to a  $(\xi, \tau)$  rectangle,  $(0, 1) \times (0, \tau(T))$ , and the partial differential equation, with  $u(\xi, \tau) = \rho(x, t)$ ,*

$$(3.7) \quad \frac{\partial u}{\partial \tau} - \frac{\partial^2}{\partial \xi^2}[\Phi(u)] = a(\xi, \tau) \frac{\partial}{\partial \xi}[\Phi(u)] + b(\xi, \tau) \frac{\partial u}{\partial \xi} + c(\xi, \tau)u.$$

*Proof.* Consider the change of independent variables given by the definitions

$$(3.8a) \quad \lambda(x, t) = \int_0^x [M(x, t)/N(x, t)]^{1/2} dx,$$

$$(3.8b) \quad \tau = \int_0^t [\lambda(1, t)]^{-2} dt,$$

and

$$(3.8c) \quad \xi = \lambda(x, t) / \lambda(1, t).$$

The hypotheses on  $M$  and  $N$  imply that (3.8) defines a differentiable invertible coordinate transformation such that  $\xi$  depends on both  $x$  and  $t$ ,  $\tau$  depends only on  $t$ ,  $\Omega \rightarrow (0, 1) \times (0, \tau(T))$ , and

$$(3.9) \quad M(x, t) d\tau/dt = N(x, t)(\partial\xi/\partial x)^2.$$

It follows that (1.5) transforms to (3.7), and the proof is complete.

**COROLLARY 3.1.** *Suppose  $\rho$  satisfies (1.5) with both  $M$  and  $N$  being positive and continuously differentiable on  $\bar{\Omega}$ , and suppose that  $\Phi'(\rho)$  is always positive. Then, (3.8) can be modified to define an invertible transformation so that (3.7) becomes, with  $u(\xi, \tau) = \rho(x, t)$ ,*

$$(3.10) \quad \frac{\partial u}{\partial \tau} - \frac{\partial^2 u}{\partial \xi^2} = b(\xi, \tau) \frac{\partial u}{\partial \tau} + c(\xi, \tau)u.$$

*Proof.* Write  $N(x, t)\Phi'[\rho(x, t)]$  for  $N(x, t)$  in (3.8a).  $\square$

Oleinik [32] reported study [33] of the initial value problem for the partial differential equation (3.7), with  $a \equiv b \equiv c \equiv 0$ , based on physical motivation given by Barenblat [5]. If  $\Phi(u)$  and  $\Phi'(u) = d\Phi/du$  are zero when  $u$  is zero, her equation has mathematically interesting properties [4], [27], [32] including finite speed of propagation of effects of initial data. Although Theorem 3.1 shows that PME densities satisfy equations with the same principal part as in Oleinik’s work [32], we recall that our present interest lies in problems where  $\Phi'(u)$  is always positive.

Although Corollary 3.1 shows that PME densities satisfy the heat equation, with respect to  $(\xi, \tau)$ , plus terms of lower order, it should be noted that this is not an uncommon fact. Specifically, if  $v$  is sufficiently differentiable on some domain  $\chi$  in the  $(\xi, \tau)$  plane, then positivity of  $v$  on  $\chi$  is a sufficient condition that  $v$  satisfy the equation

$$(3.11a) \quad \frac{\partial v}{\partial \tau} = \frac{\partial^2 v}{\partial \xi^2} + \left\{ \left[ \frac{\partial v}{\partial \tau}(\xi, \tau) - \frac{\partial^2 v}{\partial \xi^2}(\xi, \tau) \right] / v(\xi, \tau) \right\} v$$

of the form shown in (3.10). For example, we have

$$(3.11b) \quad \frac{\partial v}{\partial \tau} = \frac{\partial^2 v}{\partial \xi^2} + (\tau - \xi^2 - 1)v \quad \text{if } v(\xi, \tau) = \exp [(\xi^2 + \tau^2)/2].$$

Nevertheless, certain maximum principles apply.

**THEOREM 3.2.** *Suppose  $\rho$  satisfies (1.5) on  $\bar{\Omega}$  with  $\rho = \psi > 0$  on  $\Omega_B$ ; and suppose that both  $M$  and  $N$  are positive and continuously differentiable on  $\bar{\Omega}$  and that  $\Phi'(\rho)$  is always positive. If  $\varepsilon > |\beta_t|_0$ , then*

$$(3.12) \quad 0 < \rho(x, t) \leq e^{\varepsilon T} |\psi|_0, \quad (x, t) \in \bar{\Omega}.$$

*Proof.* Substitute  $\rho(x, t)$  in  $\Phi'$  and  $\Phi''$  to write (3.3) in the form

$$(3.13) \quad \rho_t = (A\rho_x)_x + B\rho_x - \beta_t\rho = A\rho_{xx} + (A_x + B)\rho_x - \beta_t\rho,$$

and let  $\rho = u \exp(\varepsilon t)$  to obtain

$$(3.14) \quad u_t = Au_{xx} + (A_x + B)u_x - (\beta_t + \varepsilon)u.$$

Since  $u_{xx}$  has a positive coefficient and  $u$  has a negative coefficient in (3.14), standard



methods [35] yield

$$(3.15) \quad 0 < u(x, t) \leq |e^{-\varepsilon t} \psi|_0 \leq |\psi|_0, \quad (x, t) \in \bar{\Omega},$$

(3.12) follows, and the proof is complete.

Observe that the development of (3.3) can be simultaneously applied to each equation in (1.1) to produce a simplified PMS wherein each equation has been solved for its own  $\partial \rho_i / \partial t$ . Note that Theorem 3.1 and its corollary are quite different from the development of (3.3) in that, if suppressed subscripts were written into (3.7) or (3.10),  $i$  would appear on every variable and function therein. Specifically,  $\xi_i$  and  $\tau_i$  would actually be written for  $\xi$  and  $\tau$ , respectively.

**4. Fixed point formulation.** Solution of the PMS can be studied in terms of a functional fixed point problem involving a mapping  $\Pi$  to be defined below. It is convenient to define the classes of functions

$$(4.1a) \quad R_i = \{u \in H^{2+\theta, 1+\theta/2}(\bar{\Omega}) : u|_{\Omega_B} = \psi_i\}$$

and

$$(4.1b) \quad Q_i = \{v \in H^{2+\theta, 1+\theta/2}(\bar{\Omega}) : v(\bar{\Omega}) \subset (0, 1), v|_{\Omega_B} = \Gamma_i(\psi_1, \psi_2)\}.$$

The mapping  $\Pi$  is then formally defined as follows:

- a) Choose  $S_1 \in Q_1$ ;
- b) Use (2.2) to define  $S_2 = 1 - S_1$  on  $\bar{\Omega}$ ;
- c) Solve (1.5), separately once for each  $i$ , for  $\rho_1$  and  $\rho_2$  subject to the conditions in (1.4);
- d) Define the mapping  $\Pi$  by the equation

$$(4.2) \quad \Pi(S_1)(x, t) = \Gamma_1[\rho_1(x, t), \rho_2(x, t)] \quad (x, t) \in \bar{\Omega}.$$

It is clear that a solution of the PMS can be sought in terms of the fixed point formulation

$$(4.3) \quad S_1 = \Pi(S_1).$$

**LEMMA 4.1.** *Adopt Assumptions 1.1 and 2.1 through 2.3, choose  $i = 1$  or  $i = 2$ , and let  $S \in Q_i$ . If  $S$  is used in (1.6), then there is a unique solution,  $u$ , of (1.5) in  $R_i$ .*

*Proof.* We will verify the application of Theorem 5.2 of Ladyzenskaja, Solonnikov, and Uralceva [28, p. 564] in terms of their Remark 5.2 [28, p. 565] based on our (3.5) and (3.6). Four hypotheses must be checked.

First,  $C(x, t, u) = a_{11}(x, t, u, v) = a_{11}(x, t, u, 0) = \alpha(x, t)\Phi'(u) > 0$  for  $(x, t, u) \in \bar{\Omega} \times E$ , and

$$(4.4) \quad -a(x, t, u, 0) = -\beta_i(x, t)u \leq |u|K, \quad \int_1^\infty K^{-1} dy = \infty,$$

where continuity of derivatives of  $\alpha$  and  $\beta$  permits the definition

$$(4.5) \quad K = \max \{|\alpha|_0, |\alpha_x|_0, |\beta|_0, |\beta_t|_0, |\gamma|_0\}$$

and classical solutions of (1.5) satisfy the estimate [28]

$$(4.6) \quad |u|_0 < K_1.$$

Second, note that  $y \geq 0$  implies that

$$(4.7) \quad \max \{1 + y, (1 + y)^2, y^2, y, 1\} = (1 + y)^2,$$

and calculate, noting that  $\partial C/\partial v$  is zero,

$$(4.8) \quad \begin{aligned} & |\partial C/\partial u|(1+|v|)^2 + |\partial C/\partial x|(1+|v|) + |a| \\ &= |\alpha\Phi''|(1+|v|)^2 + |\alpha_x\Phi'|(1+|v|) + |-\alpha\Phi''v^2 - \gamma\Phi'v + \beta u| \\ &\leq K(1+|v|)^2(2|\Phi''| + 2|\Phi'| + |u|) \leq q(|u|)(1+|v|)^2, \end{aligned}$$

where  $q$ , monotone increasing on the interval  $[0, K_1]$ , is defined by

$$(4.9) \quad q(z) = K \max_{0 \leq y \leq z} (2|\Phi''(y)| + 2|\Phi'(y)| + y).$$

Of course,  $C$  is differentiable, bounded with a positive lower bound, and  $a$  is continuous for  $|u|_0$  as in (4.6).

Third, for  $|u|_0$  as in (4.6),  $C$  and  $a$  are Hölder continuous in  $t$  with exponent  $\theta/2$  and in  $x$ ,  $u$ , and  $v$  with exponent  $\theta$ .

Finally, Assumption 1.1 places the appropriate conditions on  $\psi$  so that Theorem 5.2 [28, p. 564] applies, and the proof is complete.

**THEOREM 4.1.**  $\Pi$  is a well-defined map of  $Q_1 \subset H^{2+\theta, 1+\theta/2}(\bar{\Omega})$  into itself.

*Proof.* Our formal definition of  $\Pi$  places  $S_1 \in Q_1$  and implies  $S_2 \in Q_2$ . Then, Lemma 4.1 shows that  $\rho_1 \in R_1$  and  $\rho_2 \in R_2$ , Assumption 2.3 implies that  $\Pi(S_1) \in Q_1$ , and the proof is complete.

#### REFERENCES

- [1] American Chemical Society, *Flow through porous media*, Sixth State-of-the-Art Symposium, Washington, 1970.
- [2] W. F. AMES, *Nonlinear Partial Differential Equations in Engineering*, Academic Press, New York, 1965.
- [3] J. W. AMYX, D. M. BASS, JR. AND R. L. WHITING, *Petroleum Reservoir Engineering*, McGraw-Hill, New York, 1960.
- [4] D. G. ARONSON, *Regularity properties of flows through porous media: The interface*, Arch. Rational Mech. Anal., 37 (1970), pp. 1–10.
- [5] G. I. BARENBLAT, *On a class of exact solutions of horizontal one-dimensional unsteady filtration of a gas in a porous medium*, Prikl. Mat. Meh., 17 (1953), pp. 739–742.
- [6] J. BEAR, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.
- [7] G. I. BELL AND S. GLASSTONE, *Nuclear Reactor Theory*, Van Nostrand, New York, 1970.
- [8] F. E. BROWDER, *Approximation-solvability of nonlinear functional equations in normed linear spaces*, Arch. Rational Mech. Anal., 26 (1967), pp. 33–42.
- [9] J. R. CANNON, W. T. FORD AND A. V. LAIR, *Quasilinear parabolic systems*, J. Differential Equations, 20 (1976), pp. 441–472.
- [10] H. S. CARSLAW AND J. C. JAEGER, *Conduction of Heat in Solids*, 2nd ed., Oxford University Press, London, 1959.
- [11] R. E. COLLINS, *Flow of Fluids through Porous Materials*, Reinhold, New York, 1961.
- [12] J. CRANK, *The Mathematics of Diffusion*, Oxford University Press, London, 1956.
- [13] R. J. M. DE WIEST, ed., *Flow through Porous Media*, Academic Press, New York, 1969.
- [14] J. DOUGLAS, JR. AND T. DUPONT, *The numerical solution of waterflooding problems in petroleum engineering by variational methods*, Studies in Numerical Analysis 2, J. M. Ortega and W. C. Rheinboldt, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1970, pp. 53–63.
- [15] S. D. EIDEL'MAN, *Parabolic Systems*, North Holland, Amsterdam, 1969.
- [16] L. C. EVANS, *A free boundary problem: The flow of two immiscible fluids in a porous medium*, Ph.D. dissertation, Univ. of California at Los Angeles, 1976.
- [17] W. T. FORD, *Convexity in phase behavior*, SIAM J. Appl. Math., 24 (1973), pp. 545–551.
- [18] ———, *A mapping in phase behavior*, SIAM J. Appl. Math., 20 (1971), pp. 14–23.
- [19] A. R. FORSYTH, *Theory of Differential Equations*, vols. 5 and 6, Dover, New York, 1959.
- [20] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

- [21] W. FULKS AND R. B. GUENTHER, *A free boundary problem and an extension of Muskat's model*, Acta Math., 122 (1969), pp. 273–300.
- [22] P. R. GARABEDIAN, *Partial Differential Equations*, Wiley, New York, 1964.
- [23] J. L. GRAVELEAU AND P. JAMET, *A finite difference approach to some degenerate nonlinear parabolic equations*, SIAM J. Appl. Math., 20 (1971), pp. 199–223.
- [24] G. M. HIDY AND J. R. BROCK, *The Dynamics of Aerocolloidal Systems*, Pergamon, Oxford, 1970.
- [25] M. K. HUBBERT, *The Theory of Ground-Water Motion and Related Papers*, Hafner, New York, 1969.
- [26] R. KHALEEL AND D. L. REDDELL, *Simulation of pollutant movement in groundwater aquifers*, Tech. Rep. 81, Texas Water Resources Institute, Texas A & M Univ., College Station, TX, 1977.
- [27] B. F. KNERR, *Some results concerning solutions of the Cauchy problem for the porous medium equation when the initial data have compact support*, Ph.D. dissertation, Northwestern Univ., Evanston, IL, 1976.
- [28] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URALCEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Mono. 23, American Mathematical Society, Providence, RI, 1968.
- [29] M. MUSKAT, *The Flow of Homogeneous Fluids through Porous Media*, McGraw-Hill, 1937; reprinted J. W. Edwards, Ann. Arbor, MI, 1946.
- [30] M. MUSKAT, *Physical Principles of Oil Production*, McGraw-Hill, New York, 1949.
- [31] O. A. OLEINIK, *Quasi-linear second order parabolic equations with many independent variables*, Seminari dell'Istituto Nazionale di Alta Matematica 1962–63, Oderisi, Gubbio, 1964, pp. 332–354.
- [32] ———, *On some degenerate quasilinear parabolic equations*, Seminari dell'Istituto Nazionale di Alta Matematica 1962–63, Oderisi, Gubbio, 1964, pp. 355–371.
- [33] O. A. OLEINIK, A. S. KALASHNIKOV AND C. 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.
- [34] D. W. PEACEMAN, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, 1977.
- [35] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
- [36] I. REMSON, G. M. HORNBERGER AND F. J. MOLZ, *Numerical Methods in Subsurface Hydrology*, Wiley-Interscience, New York, 1971.
- [37] L. I. RUBINSTEIN, *The Stefan Problem*, Transl. Math. Mono. 27, American Mathematical Society, Providence, RI, 1971.
- [38] A. E. SCHEIDEGGER, *The Physics of Flow through Porous Media*, 3rd ed., University of Toronto Press, Toronto, Ontario, 1974.
- [39] J. T. SCHWARTZ, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.
- [40] A. SETTARI, H. S. PRICE AND T. DUPONT, *Development and application of variational methods for simulation of miscible displacement in porous media*, Soc. Petr. Eng. J., 17 (1977), pp. 228–246.
- [41] M. C. WAID, *Differential inequalities for systems of degenerate parabolic operators*, Applicable Anal., 6 (1977), pp. 229–237.
- [42] W. WALTER, *Differential and Integral Inequalities*, Springer, New York, 1970.

## A SINGULAR PARABOLIC INITIAL-BOUNDARY VALUE PROBLEM IN A NONCYLINDRICAL DOMAIN\*

VASILIOS ALEXIADES†

**Abstract.** The well-posedness of the first Fourier problem for a class of singular parabolic equations in noncylindrical domain is established in an appropriate weighted Hilbert–Sobolev space. The Green’s function is also constructed by means of potentials, and it is shown that the generalized and classical solutions coincide when the data permit the latter to exist.

**Introduction.** We study the first initial-boundary value problem for the class of singular parabolic operators depending on a real parameter  $k \geq 1$ :

$$L_k[u] + cu \equiv u_t - \left\{ u_{xx} + \frac{k}{x} u_x \right\} + cu,$$

in the noncylindrical domain

$$D \equiv D^T = \{(x, t) : 0 < t < T, 0 < x < \chi(t)\},$$

where the curve  $\Gamma: x = \chi(t)$  satisfies

$$(*) \quad \chi \in \mathcal{C}^1[0, T], \quad 0 < X_0 \leq \chi(t) \leq X_1, \quad |\dot{\chi}(t)| \leq \dot{X} < \infty \quad \text{for } 0 \leq t \leq T.$$

The operators  $L_k$  arise in axially symmetric problems and in probability theory (see [2]). The Cauchy problem for  $L_k$  has been investigated very extensively in many directions, e.g., Arena [4], Brezis–Rosenkrantz–Singer and Lax [6], Colton [8], Cholewinski–Haimo [7], Bragg [5]. In [2] the theory of potentials for  $L_k$  was developed and the main Fourier problems were studied classically. In this paper we concentrate on the first Fourier problem

$$\begin{aligned} (IBVP) \quad & L_k[u] + cu = f \quad \text{in } D, \\ & u|_{\Omega_0} \equiv u(x, 0) = g_0 \quad \text{in } \Omega_0, \\ & u|_{\Gamma} = g_1 \quad \text{in } (0, T), \end{aligned}$$

where  $\Omega_0 := \{(x, 0) : 0 < x < \chi(0)\}$ ,  $k \geq 1$  fixed, and  $c(x, t)$ ,  $f(x, t)$ ,  $g_0(x)$ ,  $g_1(t)$  are given. Note that no data are prescribed along the singular axis  $x = 0$ . In appropriate weighted Hilbert–Sobolev spaces (which are studied in § 1) we define the concept of generalized (variational) solution (§ 2) and establish the well-posedness of the problem (§ 3) under the assumptions:  $c \in L^\infty(D)$ ,  $x^{k/2}f \in L^2(D)$ ,  $x^{k/2}g_0 \in L^2(\Omega_0)$ ,  $g_1 \in H^1[0, T]$ . For the remainder of the paper we restrict ourselves to the case  $c \equiv \text{constant}$  and construct the Green’s function (§ 4) by means of potentials of first kind (the theory of which was developed in [2]). The representation of the classical solution in terms of the Green’s function allows us to show that when  $f$ ,  $g_0$ ,  $dg_1/dt$  are continuous on  $\bar{D}$ ,  $\bar{\Omega}_0$ ,  $[0, T]$  respectively, and  $f(x, t)$  is locally Hölder in  $x$  (uniformly in  $t$ ) inside  $D$ , then the generalized and the classical solutions coincide.

**1. Spaces.** From [2, § 10] we know that the classical solution  $u$  of the problem for  $L_k$  satisfies  $x^k u_x \rightarrow 0$  as  $x \rightarrow 0$ . Let  $\varphi(x, t)$  be any smooth function vanishing on  $\Gamma$  and on

\* Received by the editors October 26, 1978, and in revised form May 25, 1979.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916. This research was based on a portion of the author’s dissertation written under the supervision of Professor R. J. Weinacht at the University of Delaware.

$\Omega_T$  (i.e., for  $t = T$ ); an integration by parts yields

$$(1.1) \quad \iint_D x^k \{u_x \varphi_x - u \varphi_t + cu \varphi\} dx dt = \iint_D x^k f \varphi dx dt + \int_{\Omega_0} x^k g_0(x) \varphi(x, 0) dx.$$

Observe that (1.1) makes sense for  $u$ 's possessing only an  $x$ -derivative and  $\varphi$ 's an  $x$ - and a  $t$ -derivative. This leads us to consider appropriate spaces in which to formulate the problem in weak form.

We define the following Hilbert spaces (of equivalent classes of measurable functions as usual):

$$\begin{aligned} \mathcal{L}^2(\Omega_0, k) &:= \left\{ v(x) : \int_{\Omega_0} x^k v(x)^2 dx < \infty \right\}, \\ \mathcal{L}^2(D, k) &:= \left\{ u(x, t) : \iint_D x^k u(x, t)^2 dx dt < \infty \right\}, \end{aligned}$$

whose inner products will be denoted respectively by  $(\cdot, \cdot)_{\Omega_0, k}$  and  $(\cdot, \cdot)_{D, k}$ ;

$$\begin{aligned} W^{1,0}(D, k) &:= \{u(x, t) : u, u_x \in \mathcal{L}^2(D, k)\}, \\ W^{1,1}(D, k) &:= \{\varphi(x, t) : \varphi, \varphi_x, \varphi_t \in \mathcal{L}^2(D, k)\}, \end{aligned}$$

with inner products given respectively by  $(u, v)_{1,0;D,k} := (u, v)_{D,k} + (u_x, v_x)_{D,k}$  and  $(\varphi, \psi)_{1,1;D,k} := (\varphi, \psi)_{D,k} + (\varphi_x, \psi_x)_{D,k} + (\varphi_t, \psi_t)_{D,k}$ . The notation for norms will be similar. The derivatives are weak derivatives, in the following sense: Let  $u \in \mathcal{L}^1_{loc}(D, k) \cong L^1_{loc}(D)$ ; a function  $w \in \mathcal{L}^1_{loc}(D, k)$  is the weak  $x$ -derivative of  $u$  over  $D$ ,  $w = u_x$ , if  $\iint_D (\partial/\partial x)[x^k \zeta(x, t)]u(x, t) dx dt = -\iint_D x^k \zeta(x, t)w(x, t) dx dt \forall \zeta \in \mathcal{C}^\infty_0(D)$ . Similarly for the  $t$ -derivative. By the standard method of mollification and partition of unity arguments [1], [9] one can see that the spaces  $W$  above can equivalently be defined as closures of spaces of smooth functions; for example,  $W^{1,0}(D, k)$  is the closure with respect to the  $\|\cdot\|_{1,0;D,k}$ -norm of  $\{u \in \mathcal{C}^\infty(D) : \|u\|_{1,0;D,k} < \infty\}$ . Let us remark that the above definition as well as Theorems 1.1 and 1.2 below are valid for any  $k \in \mathbb{R}$  (our interest however is only for  $k \geq 1$ ).

Let  $X > 0$  be fixed and set  $R := (0, X) \times (0, T)$ ,  $\tilde{\Gamma} := \{X\} \times (0, T)$ ,  $\tilde{x} := (X/\chi(t))x$  for each  $t \in [0, T]$ . The (global) change of coordinates  $x \rightarrow \tilde{x}$ ,  $t \rightarrow t$  is a  $\mathcal{C}^1$ -transformation (by  $(*)$ ) with nonvanishing Jacobian, which transforms the domain  $D$  into the rectangle  $R$  by mapping  $\Gamma$  onto  $\tilde{\Gamma}$ . Then, if  $u(x, t)$  is defined on  $D$ , the function

$$(1.2) \quad \tilde{u}(\tilde{x}, t) := u\left(\frac{\chi(t)}{X}\tilde{x}, t\right) \equiv u(x, t)$$

is defined on  $R$ . The derivatives (both classical and weak) are related by  $\tilde{u}_{\tilde{x}}(\tilde{x}, t) = (\chi(t)/X)u_x(x, t)$  and  $\tilde{u}_t(\tilde{x}, t) = (\dot{\chi}(t)/\chi(t))xu_x(x, t) + u_t(x, t)$ . Over the rectangle  $R$  we define the spaces  $W^{1,0}(R, k)$  and  $W^{1,1}(R, k)$  exactly as we did over  $D$ . Various properties of the elements of the  $W$  spaces over  $D$  will be deduced from properties of elements of the corresponding spaces over  $R$ . The main reason for the transformation however is that it is needed in the uniqueness proof (§ 3). Clearly

$$(1.3) \quad u \in W^{1,j}(D, k) \quad \text{iff} \quad \tilde{u} \in W^{1,j}(R, k), \quad j = 0, 1,$$

and the identification  $u \leftrightarrow \tilde{u}$  is an isomorphism between the spaces in (1.3). We establish the existence of a trace on  $\Gamma$ :

**THEOREM 1.1.** *If  $u \in W^{1,0}(D, k)$ , then the trace  $\gamma u$  of  $u$  along  $\Gamma$  exists in  $L^2(\Gamma) \cong L^2(0, T)$  and  $u \rightarrow \gamma u$  in  $L^2(0, T)$  as  $(x, t) \rightarrow (\chi(t), t) \in \Gamma$ .*

*Proof.* Let  $0 < X_* < X$  be fixed and set  $R_* := (X_*, X) \times (0, T)$ . Now,  $u \in W^{1,0}(D, k) \Rightarrow \tilde{u}, \tilde{u}_{\tilde{x}} \in \mathcal{L}^2(R_*) \Rightarrow \tilde{u}(\cdot, t), \tilde{u}_{\tilde{x}}(\cdot, t) \in L^2(X_*, X)$  for a.a.  $t \in (0, T)$ . Hence ([13, p. 28])  $\tilde{u}(\tilde{x}, t) = \tilde{u}(X_*, t) + \int_{X_*}^{\tilde{x}} \tilde{u}_{\tilde{y}}(\tilde{y}, t) d\tilde{y}$  for a.a.  $t \in (0, T)$ , which implies that the trace  $\tilde{\gamma}\tilde{u}(t) := \lim_{\tilde{x} \rightarrow X^-} \tilde{u}(\tilde{x}, t) = \tilde{u}(X_*, t) + \int_{X_*}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t) d\tilde{y}$  exists and  $\tilde{u}(\tilde{x}, t)$  converges to it a.e. as  $\tilde{x} \rightarrow X$ . Clearly the value of  $\tilde{\gamma}\tilde{u}$  is independent of  $X_*$ . We claim that  $\tilde{\gamma}\tilde{u}(\cdot) \in L^2(0, T)$ ; indeed,  $\int_0^T \tilde{u}(\tilde{x}, t)^2 dt < \infty$  for a.a.  $\tilde{x} \in (X_*, X) \Rightarrow \exists \tilde{x}_0 \in (X_*, X)$  such that  $\int_0^T \tilde{u}(\tilde{x}_0, t) dt < \infty$ ; then  $\int_0^T |\tilde{\gamma}\tilde{u}(t)|^2 dt = \int_0^T |\tilde{u}(\tilde{x}_0, t) + \int_{\tilde{x}_0}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t) d\tilde{y}|^2 dt \leq 2 \int_0^T \tilde{u}(\tilde{x}_0, t)^2 dt + 2(X - \tilde{x}_0) \int_0^T \int_{X_*}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t)^2 d\tilde{y} dt < \infty$ . Finally, for  $X_* < \tilde{x} < X$ ,  $\int_0^T |\tilde{u}(\tilde{x}, t) - \tilde{\gamma}\tilde{u}(t)|^2 dt \leq (X - \tilde{x}) \int_0^T \int_{X_*}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t)^2 d\tilde{y} dt \rightarrow 0$  as  $\tilde{x} \rightarrow X$ . The assertions for  $u$  follow by defining  $\gamma u(t) := \tilde{\gamma}\tilde{u}(t)$ . Q.E.D.

Let us note that

$$(1.4) \quad \gamma u(t) := \tilde{\gamma}\tilde{u}(t) = \tilde{u}(\tilde{\zeta}, t) + \int_{\tilde{\zeta}}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t) d\tilde{y}$$

holds for a.a.  $t \in (0, T)$ , a.a.  $\tilde{\zeta} \in (X_*, X)$ , any fixed  $X_* \in (0, X)$ , from which one easily obtains

**THEOREM 1.2.** *The trace  $\gamma: W^{1,0}(D, k) \rightarrow L^2(0, T)$  is a continuous linear operator on  $W^{1,0}(D, k)$  with range dense in  $L^2(0, T)$ .*

These results allow us to define the space

$$(1.5) \quad \dot{W}^{1,0}(D, k) := \{u \in W^{1,0}(D, k) : \gamma u = 0\} \equiv \text{kernel of } \gamma \text{ in } W^{1,0}(D, k).$$

It follows that  $Rg(\gamma) = W^{1,0}(D, k) / \dot{W}^{1,0}(D, k)$ , in other words we have

**COROLLARY 1.** *The space of boundary values (traces) along  $\Gamma$  of elements of  $W^{1,0}(D, k)$  is the quotient  $W^{1,0}(D, k) / \dot{W}^{1,0}(D, k)$  which is a dense subset of  $L^2(\Gamma) \cong L^2(0, T)$ .*

Next we show that for functions continuous up to  $\Gamma$  the trace coincides in some sense with the restriction of the function on  $\Gamma$ . Namely we prove

**THEOREM 1.3.** *Let  $u \in W^{1,0}(D, k) \cap \mathcal{C}(D \cup \Gamma)$ . Then  $\gamma u = 0 \Leftrightarrow u|_{\Gamma} = 0$ .*

*Proof.* Equivalently, if the transformed function  $\tilde{u}$  (see (1.2)) belongs to  $W^{1,0}(R, k) \cap \mathcal{C}(R \cup \tilde{\Gamma})$  we show that  $\tilde{\gamma}\tilde{u} = 0 \Leftrightarrow \tilde{u}|_{\tilde{\Gamma}} \equiv \tilde{u}(X, t) = 0$ . Assume  $\tilde{\gamma}\tilde{u} = 0$ . From (1.4) we have

$$(1.6) \quad \tilde{u}(\tilde{x}, t) = - \int_{\tilde{x}}^X \tilde{u}_{\tilde{y}}(\tilde{y}, t) d\tilde{y}$$

for a.a.  $t \in (0, T)$ , a.a.  $\tilde{x} \in (X_*, X)$ , any fixed  $0 < X_* < X$ . Let  $(X, t_0) \in \tilde{\Gamma}$ ,  $0 < t_0 < T$ . We want  $\tilde{u}(X, t_0) = 0$ . Let  $B := (t_0 - \varepsilon, t_0 + \varepsilon)$ ,  $\varepsilon > 0$ , and for  $h > 0$  let  $S_h := (X - h, X) \times B \subset R$ . Then  $\tilde{u} \in \mathcal{C}(\tilde{S}_h)$  and (1.6) implies (by careful estimation of the constant)

$$(1.7) \quad \iint_{S_h} \tilde{x}^k \tilde{u}(\tilde{x}, t)^2 d\tilde{x} dt \leq \frac{Xh}{k+1} \iint_{S_h} \tilde{x}^k \tilde{u}_{\tilde{x}}(\tilde{x}, t)^2 d\tilde{x} dt.$$

On the other hand, by the mean value theorem there is  $\tilde{x}_0 \in (X - h, X)$  such that the left-hand side of (1.7) equals  $h \int_B \tilde{x}_0^k \tilde{u}(\tilde{x}_0, t)^2 dt$ . Thus, as  $h \rightarrow 0$  (and  $\tilde{x}_0 \rightarrow X$ ) we find  $X^k \int_B \tilde{u}(X, t)^2 dt = 0$  by continuity of  $\tilde{u}$  on  $\tilde{S}_h$ , whence  $\tilde{u}(X, t_0) = 0$ . Conversely, if  $\tilde{u}|_{\tilde{\Gamma}} \equiv \tilde{u}(X, t) = 0$  then we can write  $\tilde{u}(\tilde{x}, t)$  as in (1.6) for any  $0 < \tilde{x} < X$ , any  $0 < t < T$ , and (1.4) implies  $\tilde{\gamma}\tilde{u} = 0$ . Q.E.D.

The elements of  $W^{1,1}(D, k)$  clearly have the same trace properties along  $\Gamma$  since  $W^{1,1} \subset W^{1,0}$ . Moreover, they also possess traces on  $t = 0$  and  $t = T$  as one can easily see by proceeding as in the proof of Theorem 1.1. We state the result

**THEOREM 1.4.** *If  $\varphi \in W^{1,1}(D, k)$  then the traces  $\gamma$ ,  $\gamma_0$  and  $\gamma_T$  of  $\varphi$  along  $\Gamma$ ,  $\Omega_0$ ,  $\Omega_T$  exist respectively in  $L^2(0, T)$ ,  $\mathcal{L}^2(\Omega_0, k)$ ,  $\mathcal{L}^2(\Omega_T, k)$ . The linear operators  $\gamma$ ,  $\gamma_0$ ,  $\gamma_T$  are continuous on  $W^{1,1}(D, k)$  with dense ranges.*

This allows us to define the closed subspace of  $W^{1,1}(D, k)$

$$(1.8) \quad \dot{W}^{1,1}(D, k) := \{\varphi \in W^{1,1}(D, k) : \gamma\varphi = 0, \gamma_T\varphi = 0\},$$

which will be used as the space of test functions. Clearly  $\dot{W}^{1,1} \subset \dot{W}^{1,0}$ , so that the properties of  $\dot{W}^{1,0}$  are also valid for  $\dot{W}^{1,1}$ .

From (1.4) by repeated use of the Cauchy-Schwarz inequality one obtains

**THEOREM 1.5 (Poincaré inequality).** *For any  $u \in W^{1,0}(D, k)$  and any  $\varepsilon > 0$*

$$(1.9) \quad \iint_D x^{k-2+\varepsilon} u^2 \, dx \, dt \leq \frac{X_1^{k-1+\varepsilon}}{k-1+\varepsilon} \left\{ \|\gamma u\|_{L^2(0,T)}^2 + \frac{1}{\varepsilon X_0^{k-1}} \iint_D x^k u_x^2 \, dx \, dt \right\}.$$

As immediate consequences we have a Friedrichs inequality (take  $\varepsilon = 2$  in (1.9)), which will be used repeatedly later, and also some results about the growth of functions in  $W^{1,0}(D, k)$  as  $x \rightarrow 0$ .

**COROLLARY 1 (Friedrichs inequality).** *For  $u \in \dot{W}^{1,0}(D, k)$ ,*

$$(1.10) \quad \|u\|_{D,k} \leq C \|u_x\|_{D,k}.$$

**COROLLARY 2.** *Let  $u \in W^{1,0}(D, k)$ . For any  $\varepsilon > 0$  the function  $u^\varepsilon(x, t) := x^{\nu+\varepsilon} u(x, t)$ ,  $\nu = (k-1)/2 \geq 0$ , satisfies:*

- (i)  $\iint_D |\partial u^\varepsilon / \partial x| \, dx \, dt \leq C(k; D; \varepsilon) \{ \|\gamma u\|_{L^2(0,T)} + \|u_x\|_{D,k} \};$
- (ii) *for a.a.  $t \in (0, T)$ ,  $u^\varepsilon(x, t)$  is absolutely continuous in  $x \in [0, \chi(t)]$ ;*
- (iii)  $\lim_{x \rightarrow 0^+} u^\varepsilon(x, t) = 0$  *for a.a.  $t \in (0, T)$ .*

Consequently, for a.a.  $t \in (0, T)$

$$\lim_{x \rightarrow 0^+} x^{k-1} u(x, t) = 0 \quad \text{if } k > 1,$$

$$\lim_{x \rightarrow 0^+} x^\varepsilon u(x, t) = 0 \quad \forall \varepsilon > 0 \quad \text{if } k = 1.$$

**2. Generalized formulation of the problem.** Given  $c \in L^\infty(D)$ ,  $f \in \mathcal{L}^2(D, k)$ ,  $g_0 \in \mathcal{L}^2(\Omega_0, k)$  and  $g_1 \in L^2(0, T)$  we consider problem (IBVP) (see Introduction). Led by (1.1), we consider solutions in the space  $W^{1,0}(D, k)$  and test functions in the space  $\dot{W}^{1,1}(D, k)$ . Clearly this is the largest possible space of test functions one could choose within  $\mathcal{L}^2(D, k)$ ; there is a gain in doing so, as we shall see in the uniqueness proof, and certainly there is no loss since one could use any dense subspace whenever preferable. We define a bilinear form  $a(u, \varphi)$  on  $W^{1,0}(D, k) \times W^{1,1}(D, k)$  by

$$(2.1) \quad a(u, \varphi) := \iint_D x^k \{ u_x \varphi_x - u \varphi_t + c u \varphi \} \, dx \, dt,$$

a linear form  $\Lambda\varphi$  on  $\dot{W}^{1,1}(D, k)$  by

$$(2.2) \quad \Lambda\varphi := \iint_D x^k f \varphi \, dx \, dt + \int_{\Omega_0} x^k g_0(x) \varphi(x, 0) \, dx$$

(we have written  $\varphi(x, 0)$  for the trace  $\gamma_0\varphi$  on  $\Omega_0$ ) and we make (1.1) the basis of definition of a generalized solution.

**DEFINITION.** A function  $u(x, t)$  defined in  $D$  will be called a generalized solution of problem (IBVP) if: (i)  $u \in W^{1,0}(D, k)$ , (ii)  $\gamma u = g_1$  in  $L^2(0, T)$ , (iii)  $a(u, \varphi) = \Lambda\varphi \, \forall \varphi \in \dot{W}^{1,1}(D, k)$ .

*Remark.* This concept of generalized solution for problem (IBVP) is wider than “weak solution in Hilbert space” of the Fichera and Oleinik theory (cf. [14, p. 28]).

**3. Existence and uniqueness results.** First we consider the case  $g_1 \equiv 0$ . Let  $(IBVP)_0$  denote problem (IBVP) with boundary condition  $u|_{\Gamma} = 0$ . Then  $u$  is a generalized solution of (IBVP) if  $u \in \dot{W}^{1,0}(D, k)$  and (iii) of the definition in § 2 is satisfied.

**THEOREM 3.1.** *If (i)  $c \in L^\infty(D)$ , (ii)  $f \in \mathcal{L}^2(D, k)$ , (iii)  $g_0 \in \mathcal{L}^2(\Omega_0, k)$ , then a generalized solution of  $(IBVP)_0$  exists and satisfies*

$$(3.1) \quad \|u\|_{1,0;D,k} \leq C(k; \Gamma) \{ \|f\|_{D,k} + \|g_0\|_{\Omega_0,k} \}.$$

*Proof.* We use the Lions “variant of the projection theorem” [12, p. 37] with the following choices of spaces  $F, \Phi$  and norms: Let  $F := \dot{W}^{1,0}(D, k)$  with  $\|u\|_F := \|u_x\|_{D,k}$ ; by (1.10),  $\|\cdot\|_F$  and  $\|\cdot\|_{1,0;D,k}$  are equivalent norms on the Hilbert space  $F$ . Let  $\Phi := \dot{W}^{1,1}(D, k)$  with  $\|\varphi\|_\Phi^2 := \|\varphi_x\|_{D,k}^2 + \frac{1}{2} \int_{\Omega_0} x^k \varphi(x, 0)^2 dx$  (the trace  $\varphi(x, 0) := \gamma_0 \varphi \in \mathcal{L}^2(\Omega_0, k)$  exists by Theorem 1.4) and note that  $\Phi$  is not complete in this norm. Clearly  $\Phi \subset F$  and  $\|\varphi\|_F \leq \|\varphi\|_\Phi \forall \varphi \in \Phi$ . The form  $a(u, \varphi)$  is defined on  $F \times \Phi$  by (2.1) and satisfies:

$$\text{for each } \varphi \in \Phi, \quad |a(u, \varphi)| \leq C(\varphi) \|u\|_F \quad \forall u \in F.$$

Next, noting that without loss we can assume  $c \geq 0$  a.e. in  $D$  (if necessary, change  $u$  to  $u e^{-c_0 t}$  with  $c_0 := \text{ess sup}_D |c(x, t)|$  which has the effect of replacing  $c$  by  $c_0 + c \geq 0$  a.e. in  $D$ ) we find by an integration by parts  $a(\varphi, \varphi) \geq \|\varphi\|_\Phi^2 \forall \varphi \in \Phi$ . Finally, the linear functional  $\Lambda$  is defined on  $\Phi$  by (2.2) and thanks to (1.10) it satisfies  $|\Lambda \varphi| \leq \{C(k, \Gamma) \|f\|_{D,k} + \|g_0\|_{\Omega_0,k}\} \cdot \|\varphi\|_\Phi \forall \varphi \in \Phi$ . Thus the Lions theorem yields the existence of a  $u \in F$  such that  $a(u, \varphi) = \Lambda \varphi \forall \varphi \in \Phi$  and  $\|u\|_F \leq \|\Lambda\|$ , where the operator norm  $\|\Lambda\|$  of  $\Lambda$  is bounded by the constant in the just mentioned estimate for  $|\Lambda \varphi|$ . Q.E.D.

*Remark.* (3.1) is not an a priori bound, so uniqueness has not been proved.

**THEOREM 3.2.** *If (i), (ii), (iii) of Theorem 3.1 hold and if (iv):  $g_1 \in L^2(0, T)$  is the trace on  $\Gamma$  of some  $g \in W^{1,1}(D, k)$  (in particular, if  $g_1 \in H^1(0, T)$ ) then a generalized solution of (IBVP) exists and satisfies*

$$(3.2) \quad \|u\|_{1,0;D,k} \leq C(k; \Gamma) \{ \|f\|_{D,k} + \|g_0\|_{\Omega_0,k} + \|g\|_{1,1;D,k} + \|c\|_{L^\infty(D)} \|g\|_{D,k} + \|\gamma_0 g\|_{\Omega_0,k} \}.$$

*Proof.* The proof of Theorem 3.1 applies with the same  $F, \Phi, a(\cdot, \cdot)$  but with  $\Lambda$  replaced by  $\bar{\Lambda}$  where  $\bar{\Lambda} \varphi := \Lambda \varphi - a(g, \varphi)$  and yields the existence of a  $v \in \dot{W}^{1,0}(D, k)$  satisfying  $a(v, \varphi) = \bar{\Lambda} \varphi \forall \varphi \in \dot{W}^{1,1}(D, k)$ . Then  $u := v + g$  is a generalized solution of (IBVP). In particular, if  $g_1 \in H^1(0, T)$  then, for example, take  $g(x, t) := g_1(t), (x, t) \in D$  which clearly is in  $W^{1,1}(D, k)$ . Q.E.D.

**THEOREM 3.3 (Uniqueness).** *If  $c \in L^\infty(D)$  then problem (IBVP) has at most one generalized solution.*

*Proof.* By linearity and employing the coordinate transformation introduced in § 1 (see (1.2)) we have to show that

$$(3.3) \quad \tilde{u} \in \dot{W}^{1,0}(R, k) \quad \text{and} \quad \tilde{a}(\tilde{u}, \tilde{\varphi}) = \tilde{\Lambda} \tilde{\varphi} \quad \forall \tilde{\varphi} \in \dot{W}^{1,1}(R, k)$$

with  $\tilde{f} = 0, \tilde{g}_0 = 0$ , imply  $\tilde{u} = 0$ ; here  $\tilde{a}, \tilde{\Lambda}$ , etc. denote the transformed quantities referring to the rectangle  $R = (0, X) \times (0, T)$ . We shall prove that this is indeed the case for some appropriate choice of  $R$ , i.e., of  $X > 0$ :

**LEMMA 3.1.** *If  $u \in \dot{W}^{1,0}(D, k)$  and  $a(u, \varphi) = A \varphi \forall \varphi \in \dot{W}^{1,1}(D, k)$  with  $f \in \mathcal{L}^2(D, k), g_0 \in \mathcal{L}^2(\Omega_0, k)$  (and  $c \in L^\infty(D)$ ), then there exist  $X > 0$  and  $T_1 \in (0, T]$  such*



that the transformed solution  $\tilde{u}$  defined in  $R := (0, X) \times (0, T)$  satisfies

$$(3.4) \quad \iint_{R^\tau} \tilde{x}^k \tilde{u}(\tilde{x}, t)^2 d\tilde{x} dt \leq C \left\{ \|\tilde{f}\|_{R^\tau, k}^2 + \int_0^X \tilde{x}^k \tilde{g}_0(\tilde{x})^2 d\tilde{x} \right\},$$

$0 \leq \tau \leq T_1$ , where the constant  $C = C(k; \Gamma; X; \tau)$  is a continuous and increasing function of  $\tau$ , and  $R^\tau = (0, X) \times (0, \tau)$ .

Uniqueness is obtained from this lemma as follows:  $f = 0 = g_0 \Rightarrow \tilde{f} = 0 = \tilde{g}_0$ , so the lemma implies  $\tilde{u} = 0$  for a.a.  $(\tilde{x}, t) \in R^\tau \forall \tau \in [0, T_1]$ . Hence, if  $T_1 = T$  then  $\tilde{u} = 0$  a.e. in  $R$ ; if  $T_1 < T$ , we repeat the argument over  $[T_1, 2T_1]$ , etc. until  $T$  is reached, to show that  $\tilde{u} = 0$  a.e. in  $R^T \equiv R$  and therefore also  $u = 0$  a.e. in  $D$ . Thus, it only remains to prove the lemma.

*Proof of the lemma.* Let  $X > 0$  be fixed (to be chosen later) and set as before  $R^\tau := (0, X) \times (0, \tau)$ , with  $R := R^T$ . By transforming from  $D$  to  $R$ , the hypotheses of the lemma are equivalent to (3.3). The basic idea of the method we shall use is due to Ladyzenskaya [11, p. 127]. All quantities below refer to  $R$  but the  $\sim$  will be omitted for simplicity in the notation. Let  $\tau \in (0, T]$  be fixed (to be chosen) and set

$$(3.5) \quad \psi(x, t) := \begin{cases} \int_t^\tau u(x, s) ds & \text{for } 0 \leq t \leq \tau, \\ 0 & \text{for } \tau \leq t \leq T. \end{cases}$$

Then  $\psi \in \dot{W}^{1,1}(R, k)$  is an acceptable test function in (3.3) and at the same time  $u = -\psi_x$ ,  $u_x = -\psi_{xt}$ . Thus (3.3) gives

$$(3.6) \quad \begin{aligned} & -\frac{X^2}{2} \iint_{R^\tau} (x^k \chi(t)^{k-1} \psi_x^2)_t dx dt + \iint_{R^\tau} x^k \chi(t)^{k+1} \psi_t^2 dx dt \\ & = -\frac{(k-1)X^2}{2} \iint_{R^\tau} x^k \chi(t)^{k-2} \dot{\chi}(t) \psi_x^2 dx dt + \iint_{R^\tau} x^{k+1} \chi(t)^k \dot{\chi}(t) \psi_x \psi_t dx dt \\ & \quad + \iint_{R^\tau} x^k \chi(t)^{k+1} c \psi \psi_t dx dt + \iint_{R^\tau} x^k \chi(t)^{k+1} f \psi dx dt \\ & \quad + \int_0^X x^k \chi(0)^{k+1} g_0(x) \psi(x, 0) dx. \end{aligned}$$

We integrate the first term on the left and estimate each term on the right using (\*), Cauchy-Schwarz, an arithmetic-geometric mean inequality and (1.10). The result is

$$(3.7) \quad \begin{aligned} & \frac{X^2 \chi(0)^{k-1}}{2} \left[ 1 - \varepsilon \frac{\chi(0)^2}{2(k+1)} \right] \int_0^X x^k \psi_x(x, 0)^2 dx + A_1 \iint_{R^\tau} x^k \chi(t)^{k+1} \psi_t^2 dx dt \\ & \leq (\text{positive const.}) \iint_{R^\tau} x^k \psi_x^2 dx dt + F_1(\tau), \quad \forall \varepsilon > 0, 0 \leq \tau \leq T, \end{aligned}$$

where  $A_1 := 1 - (\varepsilon/2)(X \dot{X} X_1^{1/2} + \|c\|_{L^\infty})$ , and

$$(3.8) \quad F_1(\tau) := \frac{\varepsilon}{2} X_1^{k+1} \|f\|_{R^\tau, k} + \frac{X(0)^{k+1}}{2\varepsilon} \int_0^X x^k g_0(x)^2 dx.$$

Now, the parameters  $X > 0$  and  $\varepsilon > 0$  can be chosen so that the coefficients in (3.7) are all positive. Then, dividing through by the first coefficient we can write (3.7) in the form

$$(3.9) \quad \int_0^X x^k \psi_k(x, 0)^2 dx + A \iint_{R^\tau} x^k \psi_t^2 dx dt \leq B \iint_{R^\tau} x^k \psi_x^2 dx dt + F(\tau),$$

where  $A, B$  are positive constants depending only on  $k, \|c\|_{L^\infty}$  and (the parameters of)  $\Gamma$ , and  $F(\tau)$  (similar to  $F_1(\tau)$  in (3.8)) is a continuous and increasing function of  $\tau, 0 \leq \tau \leq T$ .

Next, the arbitrariness of  $\tau$  is utilized as follows: Let  $\zeta(x, t) := \int_0^t u(x, s) ds$  for  $0 \leq t \leq \tau$ . Because of (3.5) we have  $\psi(x, t) = \zeta(x, \tau) - \zeta(x, t)$ , hence

$$(3.10) \quad \|\psi_x\|_{R^\tau, k}^2 \leq 2\|\zeta_x\|_{R^\tau, k}^2 + 2\tau \int_0^x x^k \zeta(x, \tau)^2 dx, \quad 0 \leq \tau \leq T.$$

Substituting  $\psi$  in terms of  $\zeta$  in (3.9) and using (3.10) we obtain

$$(3.11) \quad (1 - 2B\tau) \int_0^x x^k \zeta_x(x, \tau)^2 dx + A\|\zeta_t\|_{R^\tau, k}^2 \leq 2B \int_0^\tau \left( \int_0^x x^k \zeta_x(x, t)^2 dx \right) dt + F(\tau), \quad 0 \leq \tau \leq T.$$

Letting  $T_1 := \min \{1/(4B), T\}$ , we have  $1 - 2B\tau \geq \frac{1}{2}$  for  $0 \leq \tau \leq T_1$ . We apply the Gronwall inequality [16, p. 14] to (3.11) and disregarding unneeded terms, we find

$$A \int \int_{R^\tau} x^k \zeta_t(x, t)^2 dx dt \leq F(\tau) + 4B e^{4B\tau} \int_0^\tau F(s) ds \leq [1 + 4B e^{4B\tau} \cdot \tau]F(\tau), \quad 0 \leq \tau \leq T_1.$$

This is the assertion of the lemma (recall that  $\zeta_t = u$ , (3.8), and that we have been deleting  $\sim$  over  $x, u, f, g_0$ ). Q.E.D.

**4. Green's function.** From now on we restrict our attention to the case  $c(x, t) =$  constant, i.e. (after changing  $u$  to  $ue^{-ct}$ ) to the equation

$$(4.1) \quad L_k[u] \equiv u_t - \left\{ u_{xx} + \frac{k}{x} u_x \right\} = f,$$

and (IBVP)\* will refer to (IBVP) with the equation replaced by (4.1).

We shall employ the ingenious method of Pogorzelski [15] to construct the Green's function by means of a potential of first kind (instead of the second kind potential usually needed). The theory of potentials for the operator  $L_k$  was developed in [2], and the classical solution of (IBVP)\* was represented there in terms of potentials. The representation of the classical solution via Green's function, which we shall obtain here, allows us to prove that the generalized and the classical solution coincide when the data permit.

We begin with the construction of the Green's function  $G(x, t; y, s)$  for the operator  $L_k$  in the domain  $D$ . It is a function defined and continuous for  $(x, t) \in \bar{D}, (y, s) \in D \cup \Omega_0, 0 \leq s < t \leq T$  and of the form

$$(4.2) \quad G(x, t; y, s) \equiv E(x, t; y, s) - Z(x, t; y, s),$$

where  $y^k E(x, t; y, s)$  is the fundamental solution of  $L_k$  (see [2]) and  $Z$  has the following properties: For each  $(y, s) \in D \cup \Omega_0$

$$(Z1) \quad Z(\cdot, \cdot; y, s) \in \mathcal{C}(\bar{D}_s), \quad \text{where } D_s = \{(x, t) : s < t < T, 0 < x < \chi(t)\},$$

$$(Z2) \quad L_k[Z](x, t) = 0 \quad \text{for } (x, t) \in D, 0 \leq s < t \leq T,$$

$$(Z3) \quad Z(x, s; y, s) := \lim_{t \downarrow s} Z(x, t; y, s) = 0, \quad 0 \leq x \leq \chi(s),$$

$$(Z4) \quad Z(x, t; y, s)|_{(x, t) \in \Gamma} = E(x, t; y, s)|_{(x, t) \in \Gamma}, \quad 0 \leq s < t \leq T.$$

We recall that

$$E(x, t; y, s) = \frac{(xy)^{-\nu}}{2(t-s)} I_\nu \left( \frac{xy}{2(t-s)} \right) \exp \left\{ \frac{x^2 + y^2}{4(t-s)} \right\}, \quad t > s, \quad xy \neq 0,$$

and  $E = 0$  for  $t \leq s$ ,  $x \neq y$ , whereas for  $t > s$ ,  $xy = 0$ ,  $E$  takes the asymptotic form  $[2^k \Gamma(\nu + 1)]^{-1} (t-s)^{-(\nu+1)} \exp \{-(x^2 + y^2)/(4(t-s))\}$ ; here  $\nu = (k-1)/2 \geq 0$ ,  $I_\nu$  is the modified Bessel function of order  $\nu$  and  $\Gamma(\cdot)$  is the gamma function. Note that by (uniqueness) Theorem 2 of [2, § 6], properties (Z1)–(Z4) determine  $Z$  uniquely for each  $(y, s) \in D \cup \Omega_0$ . Now,  $Z$  can be found in the form of a second kind potential by solving a problem of type (I) $_{k \geq 1}$  (see [2, § 9]). Pogorzelski's method however determines  $Z$  in the form of a first kind potential which is always smoother. The method was devised for uniformly parabolic operators [15] and is also described in [10]. Fix  $(y, s) \in D \cup \Omega_0$  and define  $Z$  as the potential of first kind ([2, § 4]):

$$(4.3) \quad Z(x, t; y, s) := \int_0^t \mathcal{H}(x, t; \tau) \omega(\tau; y, s) d\tau, \quad x > 0, \quad s < t < T,$$

where

$$(4.4) \quad \mathcal{H}(x, t; \tau) := \chi(\tau)^k E(x, t; \chi(\tau), \tau),$$

and  $\omega(\cdot; y, s)$  is the unique continuous solution of the Volterra integral equation

$$(4.5) \quad \frac{1}{2} \omega(t; y, s) + E_x(\chi(t), t; y, s) = \int_s^t \mathcal{H}_x(\chi(t), t; \sigma) \omega(\sigma; y, s) d\sigma,$$

$s < t \leq T$  (see [2, § 9] for the solvability of (4.5)). Thanks to the properties of potentials of first kind established in [2, § 4], the function  $Z(x, t; y, s)$  is continuous for  $x \geq 0$ ,  $s \leq t \leq T$  and satisfies (Z1)–(Z4) above, for each  $(y, s) \in D \cup \Omega_0$ . Then (4.2) determines  $G$ . Using the strong (Nirenberg) maximum principle as in the proof of Theorem 2 of [2, § 6] one can show that for each  $(y, s) \in D \cup \Omega_0$ ,  $G(x, t; y, s) > 0$  for  $(x, t) \in D_s$  (see [3, Chap. 4] for details).

As one would expect the Green's function is no more singular than the fundamental solution itself. Indeed,  $G$  satisfies estimates identical with those of  $E$  [2, § 1], namely we have

**THEOREM 4.1.** For  $(x, t) \in D$ ,  $(y, s) \in D \cup \Omega_0$ ,  $0 \leq s < t \leq T$ ,

$$(4.6)_G \quad |G(x, t; t, s)| \leq C_1 \cdot (xy)^{-k/2} (t-s)^{-1/2} \exp \{-(x-y)^2/(32(t-s))\},$$

$$(4.7)_G \quad |G_x(x, t; y, s)| \leq C_2 \cdot (xy)^{-k/2} (t-s)^{-1} \exp \{-(x-y)^2/(32(t-s))\};$$

the constants  $C_1, C_2$  depend only on  $k$  and (the parameters of)  $\Gamma$ .

These follow from the corresponding estimates on  $E$  [2, § 1] and Lemma 4.2 below:

$$\text{LEMMA 4.1.} \quad |\omega(t; y, s)| \leq C \cdot [\chi(t)y]^{-k/2} (t-s)^{-1} \exp \left\{ -\frac{[\chi(t)-y]^2}{16(t-s)} \right\}, \quad t > s.$$

**LEMMA 4.2.**  $Z$  and  $Z_x$  satisfy the estimates of Theorem 4.1.

*Proof of Lemma 4.1.*  $E$  and  $E_x$  admit the bounds (4.6) and (4.7) respectively (with 8 in place of 32, see [2, § 1]) and then the same is true for  $\mathcal{H}$  and  $\mathcal{H}_x$  by (4.4); let us refer to these estimates as (4.6) $_E$ , (4.6) $_{\mathcal{H}}$ , etc. Apply the Gronwall inequality [16, p. 14] to (4.5) to get

$$(4.8) \quad \frac{1}{2} |\omega(t; y, s)| \leq |E_x| + \int_s^t |E_x(\chi(\sigma), \sigma; y, s)| |\mathcal{H}_x(\chi(t), t; \sigma)| e^{C\sqrt{T}} d\sigma,$$

$s \leqq t \leqq T$ . Use (4.7)<sub>E</sub>, (4.7)<sub>ℳ</sub> and on the resulting exponential the estimate (recall that  $\chi$  is Lipschitz by (\*))

$$(4.9) \quad \exp \left\{ -\frac{[\chi(t) - \chi(\sigma)]^2}{8(t - \sigma)} - \frac{[\chi(\sigma) - y]^2}{8(\sigma - s)} \right\} \leqq \text{const. } e^{-p^2/(t-s)} \cdot e^{-q^2/(\sigma-s)},$$

where  $p := \frac{1}{4}[\chi(t) - y]$ ,  $q := \frac{1}{4}[\chi(s) - y]$ , to find

$$(4.10) \quad \text{integral in (4.8)} \leqq C[\chi(t)y]^{-k/2} e^{-p^2/(t-s)} \int_s^t (t-s)^{-1/2} (\sigma-s)^{-1} e^{-q^2/(\sigma-s)} d\sigma.$$

By a change of variable the last integral is recognized as a Laplace transform involving the Bessel function  $I_0$  and it can be estimated by  $(t-s)^{-1/2} e^{-q^2/(t-s)}$ . Thus the lemma follows. Q.E.D.

*Proof of Lemma 4.2.* Use (4.3), (4.6)<sub>ℳ</sub>, Lemma 4.1, (4.9) (with obvious changes) and finally estimate the resulting integral as above to prove (4.6)<sub>Z</sub>. Similarly for (4.7)<sub>Z</sub>. Q.E.D.

**THEOREM 4.2.** *If (i)  $f \in \mathcal{C}(\bar{D})$  and is locally Hölder continuous in  $x$ , uniformly in  $t$ , for  $(x, t) \in D$ ; (ii)  $g_0 \in \mathcal{C}(\bar{\Omega}_0)$ ; (iii)  $g_1 \in \mathcal{C}^1[0, T]$ ; (iv)  $g_0(\chi(0)) = g_1(0)$ , then the solution of (IBVP)\* is given by*

$$(4.11) \quad u(x, t) = g_1(t) + \iint_{D'} y^k G(x, t; y, s) \bar{f}(y, s) dy ds + \int_{\Omega_0} y^k G(x, t; y, 0) \bar{g}_0(y) dy,$$

with  $\bar{f}(x, t) := f(x, t) - g_1'(t)$ ,  $\bar{g}_0(x) := g_0(x) - g_1(0)$ .

*Proof.* By Theorem 4.1, the integrals in (4.11) are analogous to the area potential  $U[\bar{f}]$  and the initial potential  $i[\bar{g}_0]$  respectively (see [2]); their properties and those of  $Z$  allow one to show that  $u$  is a classical solution. Uniqueness follows from [2, § 6, Thm. 2]. Details can be found in [3]. Q.E.D.

One of the advantages of representation (4.11) over the one in terms of potentials, given in [2, § 9], is that it does not involve the potential of second kind which is singular. This enables us to show that the classical solution is also a generalized solution (§ 2) and that the two coincide.

**THEOREM 4.3.** *The classical solution of (IBVP)\* given in (4.11) belongs to  $W^{1,0}(D, k)$  and for any  $\varphi \in \dot{W}^{1,1}(D, k)$  it satisfies*

$$(4.12) \quad \lim_{x \rightarrow 0} \int_0^T x^k u_x \varphi dt = 0.$$

*Outline of proof.* Since  $u \in \mathcal{C}(\bar{D})$ , we only have to prove  $u_x \in \mathcal{L}^2(D, k)$  and (4.12). Now,  $u_x$  can be computed from (4.11) and the resulting integrals behave respectively like  $U_x[\bar{f}]$  and  $i_x[\bar{g}_0]$ , [2, § 5 and 2]; from the estimate

$$|u_x(x, t)| \leqq C \cdot x^{-k/2} \{t^m M_1 + t^{-1/2} M_2\}, \quad x > 0, \quad 0 < t \leqq T, \quad \text{any } 0 < m \leqq \frac{1}{2},$$

$M_1, M_2$  constants, we see that the trace of  $u_x$  along  $\Gamma$  is integrable and also that (4.12) holds for  $\varphi$  bounded on  $\bar{D}$ . An energy type estimate from the equation shows  $u_x \in \mathcal{L}^2(D, k)$ . Next,  $\varphi \in \dot{W}^{1,1}(D, k)$  implies that for a.a.  $x \in [0, X_0]$  and any  $0 < \varepsilon < 1$ ,  $\int_0^T t^{-1/2} |\varphi| dt \leqq \text{const. } \|\varphi_t\|_{L^2(0,T)} < \infty$ ; also, the above estimate on  $|u_x(x, t)|$  gives  $|\int_0^T x^k u_x \varphi dt| \leqq C \cdot x^{k/2} \int_0^T t^{-1/2} |\varphi| dt$ ,  $x > 0$ , and from these (4.12) follows. Q.E.D.

Multiplying equation (4.1) by  $x^k \varphi$  and integrating over  $D$  using (4.12), one sees that the classical solution  $u$  also satisfies  $a(u, \varphi) = \Lambda \varphi$ ,  $\varphi \in \dot{W}^{1,1}(D, k)$ , therefore it is a generalized solution. By uniqueness of the latter (Theorem 3.3) the two must coincide.

Thus we have

**THEOREM 4.4.** *Under the hypotheses of Theorem 4.2, the generalized solution of (IBVP)\* is classical.*

#### REFERENCES

- [1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] V. ALEXIADES, *Generalized axially symmetric heat potentials and singular parabolic initial boundary value problems*, Arch. Rational Mech. Anal., to appear.
- [3] ———, *Moving boundary problems for a class of singular parabolic partial differential equations*, Ph.D. dissertation, Univ. of Delaware.
- [4] O. ARENA, *On a singular parabolic equation related to axially symmetric heat potentials*, Ann. Mat. Pura Appl., Ser. IV, 105 (1975), pp. 347–393.
- [5] L. BRAGG, *The radial heat polynomials and related functions*, Trans. Amer. Math. Soc., 119 (1965), pp. 270–290.
- [6] H. BREZIS, W. ROSENKRANTZ AND B. SINGER with an appendix by P. LAX, *On a degenerate elliptic-parabolic equation occurring in the theory of probability*, Comm. Pure Appl. Math., 24 (1971), pp. 395–416.
- [7] F. CHOLEWINSKI AND D. HAIMO, *The Weirstrass–Hankel convolution transform*, J. Analyse Math., 17 (1966), pp. 1–58.
- [8] D. COLTON, *Cauchy's problem for a singular parabolic differential equation*, J. Differential Equations, 8 (1970), pp. 250–257.
- [9] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
- [10] A. M. IL'IN, A. S. KALACHNIKOV AND O. A. OLEINIK, *Linear equations of the second order of parabolic type*, Russian Math. Surveys, 17 (1963), pp. 1–143.
- [11] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Mono., vol. 23, American Mathematical Society, Providence, RI, 1968.
- [12] J. L. LIONS, *Equations Différentielles Opérationnelles et Problèmes aux Limites*, Springer-Verlag, Berlin, 1961.
- [13] S. G. MIKHLIN, *Mathematical Physics, An Advanced Course*, North-Holland, Amsterdam, 1970.
- [14] O. A. OLEINIK AND E. V. RADKEVIC, *Second Order Equations with Nonnegative Characteristic Form*, American Mathematical Society, Providence, RI, and Plenum, New York, 1973.
- [15] W. POGORZELSKI, *Étude d'une fonction de Green et du Problème aux Limites pour l'Équation Parabolique Normale*, Ann. Polon. Math., 4 (1958), pp. 288–307.
- [16] W. WALTER, *Differential and Integral Inequalities*, Springer-Verlag, New York, 1970.

## ORTHOGONAL POLYNOMIALS WHOSE DISTRIBUTION FUNCTIONS HAVE FINITE POINT SPECTRA\*

T. S. CHIHARA†

**Abstract.** A counterexample is given to an assertion by K. M. Case that if the coefficients in the three-term recurrence formula for orthogonal polynomials converge as fast as  $n^{-2}$ , the corresponding distribution function has only finitely many discrete points in its spectrum. Some positive results concerning this situation are also given, and continuity of the distribution function is investigated.

**1. Introduction.** The classical three-term recurrence formula

$$(1.1) \quad \begin{aligned} P_n(x) &= (x - c_n)P_{n-1}(x) - \lambda_n P_{n-2}(x), & n \geq 1, \\ P_{-1}(x) &= 0, \quad P_0(x) = 1, & c_n \text{ real, } \lambda_{n+1} > 0, \end{aligned}$$

defines a sequence of monic polynomials which are orthogonal with respect to a real distribution  $d\psi$  on a subset of the real line. Conversely, of course, every monic orthogonal polynomial sequence satisfies such a recurrence.

According to an often overlooked theorem of O. Blumenthal [1], if

$$(1.2) \quad \lim_{n \rightarrow \infty} c_n = c, \quad \lim_{n \rightarrow \infty} \lambda_n = \lambda,$$

then the numbers

$$(1.3) \quad \sigma = c - 2\sqrt{\lambda}, \quad \tau = c + 2\sqrt{\lambda}$$

are the smallest and largest limit points of  $\mathcal{S}(\psi)$ , the spectrum of  $\psi$  (=support of  $d\psi$ ). Moreover, the zeros of all  $P_n(x)$  form a dense subset of  $[\sigma, \tau]$ .

Blumenthal also asserted that there are at most finitely many points of  $\mathcal{S}(\psi)$  on the complement of  $[\sigma, \tau]$ , but this has been shown to be incorrect [6]. In recent years, there has been important work by K. M. Case and others in which (1.1) is studied from the viewpoint of scattering theory in physics (see [2], [3], [4], [9]). In particular, this work has renewed some interest in the question of when there are, in fact, at most finite many spectral points outside  $[\sigma, \tau]$ .

J. S. Geronimo and Case [9] have recently proved that if

$$\sum_{n=1}^{\infty} n(|c_n - c| + |\lambda_n - \lambda|) < \infty,$$

then there are at most finitely many spectral points on the complement of  $[\sigma, \tau]$ , and  $\psi$  is continuous at  $\sigma$  and  $\tau$ . (For two different and simpler proofs, see [8].) Also, P. Nevai has shown [10, Thm. 40] that if

$$\sum_{n=1}^{\infty} (|c_n - c| + |\lambda_n - \lambda|) < \infty,$$

then  $\psi$  is absolutely continuous on  $(\sigma, \tau)$  and  $\psi$  is positive and continuous on  $(\sigma, \tau)$ .

Earlier, Case [2], [3] had claimed that

$$(1.4) \quad c_n - c = O(n^{-2}), \quad \lambda_n - \lambda = O(n^{-2})$$

is sufficient for  $\psi$  to have at most finitely many discrete spectral points. However, his proof [3] was in error so the question of whether in fact the assertion is correct has

\* Received by the editors January 12, 1979, and in revised form May 10, 1979.

† Department of Mathematics, Purdue University Calumet Campus, Hammond, Indiana 46323.

aroused some interest. It is the purpose of this note to provide a counterexample. In addition, we will obtain a few positive results involving the hypothesis (1.4), and will also investigate the continuity of  $\psi$  at  $\sigma$  and  $\tau$ .

**2. Counterexamples.** There is no loss of generality if we assume

$$(2.1) \quad \lim_{n \rightarrow \infty} c_n = 0, \quad \lim_{n \rightarrow \infty} \lambda_n = \frac{1}{4},$$

so that  $[\sigma, \tau] = [-1, 1]$ .

For brevity's sake, we will denote by  $\mathcal{A}$  the set of all distribution functions for which  $[\sigma, \tau] = [-1, 1]$  and which have at most finitely many spectral points outside  $[-1, 1]$ . We then recall the connection between (1.1) and chain sequences [5] (see also [7]). We set

$$(2.2) \quad \alpha_n(x) = \frac{\lambda_{n+1}}{(x - c_n)(x - c_{n+1})}, \quad n \geq 1.$$

**THEOREM 1.** *A necessary and sufficient condition for  $\psi \in \mathcal{A}$  is that there exists  $N \geq 0$  such that  $|c_n| < 1$  for  $n > N$  and  $\{\alpha_{n+N}(t)\}_{n=1}^\infty$  is a chain sequence for  $t = -1$  and  $t = 1$ .*

*Proof.* Let  $\{P_n^{(N)}(x)\}_{n=1}^\infty$  denote the orthogonal polynomial sequence determined by (1.1) after replacing  $c_n$  and  $\lambda_n$  by  $c_{n+N}$  and  $\lambda_{n+N}$ . If  $|c_{n+N}| < 1$  ( $n \geq 1$ ) and  $\{\alpha_{n+N}(t)\}_{n=1}^\infty$  is a chain sequence for  $t = \pm 1$ , then the true interval of orthogonality for  $P_n^{(N)}(x)$  is a subset of  $[-1, 1]$  [5, Lemma 5]. Hence by [5, Lemma 7],  $\psi$  has at most  $N$  spectral points smaller than  $-1$  and at most  $N$  larger than  $1$ .

The converse is given by [6, Thm. 1].

We will have several occasions to refer to the following fundamental results of H. S. Wall [12, pp. 82, 84] (see also [7]).

**THEOREM 2.** *Let  $0 < a_n \leq b_n, n \geq 1$ .*

(a) *If  $\{b_n\}$  is a chain sequence, so is  $\{a_n\}$ . Moreover, if  $m_n$  and  $M_n$  are the  $n$ -th minimal and maximal parameters of  $\{a_n\}$  and if  $g_n$  is any  $n$ -th parameter for  $\{b_n\}$ , then*

$$m_n \leq g_n \leq M_n, \quad n = 0, 1, 2, \dots$$

(b)  *$\{g_n\}$  is the maximal parameter sequence for  $\{b_n\}$  if and only if*

$$(2.3) \quad \sum_{m=1}^\infty \frac{g_1 \cdots g_m}{(1 - g_1) \cdots (1 - g_m)} = \infty.$$

Now consider

$$(2.4) \quad a_n = \frac{1}{4} + \frac{1}{16n(n+1)}, \quad n \geq 1.$$

We have  $a_n = (1 - H_{n-1})H_n$  where

$$H_n = \frac{2n+1}{4(n+1)}, \quad n \geq 0.$$

Then  $\{a_n\}$  is a chain sequence and Wall's criterion (2.3) shows that  $\{H_n\}$  is its maximal parameter sequence.

On the other hand, if we set

$$(2.5) \quad b_n \equiv b_n(\gamma) = \frac{1}{4} + \frac{1}{4\gamma n(n+1)}, \quad n \geq 1,$$

then direct calculation shows that  $\{b_n\}$  is not a chain sequence if  $0 < \gamma < 4$ . We will now show that, at least for  $0 < \gamma < 1$ ,  $\{b_{n+N}\}_{n=1}^\infty$  is not a chain sequence for any  $N \geq 0$ .

For any numerical sequence  $\{f_n\}$ , we write

$$f_n^{(k)} = f_{n+k}.$$

Then let  $0 < \gamma < 1$  and assume that  $\{b_n^{(N)}(\gamma)\}_{n=1}^\infty$  is a chain sequence. Let  $m_n$  and  $M_n$  denote its minimal and maximal parameters.

Since  $\{\frac{1}{4}\}_{n=1}^\infty$  is a chain sequence whose  $n$ th minimal parameter is  $n/(2n+2)$ , we have by Theorem 2,  $M_n \geq n/(2n+2)$ . On the other hand,  $\{a_n^{(N)}\}_{n=1}^\infty$  is a chain sequence whose maximal parameters are  $H_n^{(N)}$  [7, p. 94]. Since  $b_n^{(N)} > a_n^{(N)}$ , Theorem 2 now yields

$$(2.6) \quad \frac{n}{2(n+1)} \leq M_n \leq H_n^{(N)} = \frac{2(n+N)+1}{4(n+N+1)}.$$

Next set  $d_n = b_n^{(N)}$ . Then for every integer  $P \geq 0$ ,  $\{d_n^{(P)}\}_{n=1}^\infty$  is also a chain sequence and its maximal parameter sequence is  $\{M_n^{(P)}\}_{n=0}^\infty$ . We have

$$(2.7) \quad \begin{aligned} M_k^{(P)} - M_{k-1}^{(P)} &= \frac{d_k^{(P)} - (1 - M_{k-1}^{(P)})M_{k-1}^{(P)}}{1 - M_{k-1}^{(P)}} \\ &\cong \frac{d_{k+P} - \frac{1}{4}}{1 - M_{k+P-1}}. \end{aligned}$$

Therefore by (2.6),

$$\begin{aligned} M_{n+P} - M_P &\geq \sum_{k=1}^n \frac{2(k+P)}{k+P+1} (d_{k+P} - \frac{1}{4}), \\ \frac{1}{2} - \frac{P}{2(P+1)} &\geq \frac{P+1}{2\gamma(P+2)} \sum_{k=1}^n [(k+N+P)(k+N+P+1)]^{-1}. \end{aligned}$$

Hence

$$\gamma \geq \frac{(P+1)^2}{P+2} \left[ \frac{1}{N+P+1} - \frac{1}{n+N+P+1} \right].$$

Letting  $n \rightarrow \infty$ , we conclude that we can choose  $P$  sufficiently large to arrive at a contradiction. Thus  $\{b_n^{(N)}(\gamma)\}_{n=1}^\infty$  is not a chain sequence for any  $N \geq 0$ .

To obtain a counterexample to Case's assertion, we can choose an arbitrary sequence  $\{c_n\}$  such that  $c_n = O(n^{-2})$ , and then set

$$\lambda_{n+1} = b_n(\gamma)(1 - c_n)(1 - c_{n+1}), \quad 0 < \gamma < 1.$$

Then  $\{\alpha_n^{(N)}(1)\} = \{b_n^{(N)}(\gamma)\}$  is not a chain sequence for any  $N \geq 0$ ; hence by Theorem 1, the corresponding distribution function  $\psi$  has denumerably many spectral points larger than  $\tau = 1$ . There may be at most finitely many spectral points smaller than  $\sigma = -1$ , but it is possible to choose  $c_n$  small enough (e.g.  $c_n = 0$ ) so that  $\{\alpha_n^{(N)}(-1)\}$  is not a chain sequence for any  $N$  either.

We note a specific, simple example involving (2.4). Let

$$(2.8) \quad c_n = \frac{1}{4n^2 - 1}, \quad \lambda_{n+1} = \frac{n(n+1)}{(2n-1)(2n+3)}.$$

Then

$$\alpha_n(-1) = \frac{(2n+1)^2}{16n(n+1)} = a_n$$



so  $\{\alpha_n(-1)\}$  is a chain sequence and there are no spectral points smaller than  $-1$ . (In fact, since  $\{a_n\}$  does not determine its parameters uniquely, the distribution will have the form  $d\psi(x) = (x + 1) d\phi(x)$ , where the spectrum of  $\phi$  lies in  $[-1, \infty)$  [5, Thm. 1].)

By contrast,

$$\alpha_n(1) = \frac{n(n+1)(2n+1)^2}{4(2n^2-1)(2n^2+4n+1)} > b_n(\gamma)$$

for all  $n$  sufficiently large if  $\frac{4}{5} < \gamma < 1$ . Thus there are denumerably many spectral points larger than 1.

*Remark.* The inequality (2.7) can easily be improved and  $\{b_n^{(N)}\}$  shown to be not a chain sequence if  $0 < \gamma < \frac{4}{3}$ . However, this slight improvement seems not to be worth the effort since we conjecture that this conclusion holds for  $0 < \gamma < 4$ .

**3. Some positive results.** We next note a few positive results involving the condition (1.4). In many cases, the conclusion  $\psi \in \mathcal{A}$  can be obtained simply by comparing  $\alpha_n(\pm 1)$  with  $a_n$  in (2.4). However, it seems desirable to have conditions expressed directly in terms of the asymptotic properties of  $c_n$  and  $\lambda_n$ .

**THEOREM 3.** *Let  $c_n = O(n^{-2})$ ,  $\lambda_n - \frac{1}{4} = O(n^{-2})$ , and*

$$(3.1) \quad L(t) = \limsup_{n \rightarrow \infty} n^2 \left[ \left( \lambda_{n+1} - \frac{1}{4} \right) + \frac{t}{4} (c_n + c_{n+1}) \right].$$

If  $L(t) < \frac{1}{16}$  for  $t = -1$  and  $t = 1$ , then  $\psi \in \mathcal{A}$ .

*Proof.* Let  $\delta_n(x) = \alpha_n(x) - \frac{1}{4}$ . Then for  $t = \pm 1$ ,

$$(3.2) \quad \delta_n(t) = \frac{4\lambda_{n+1} - 1 + t(c_n + c_{n+1}) - c_n c_{n+1}}{4(t - c_n)(t - c_{n+1})}.$$

Thus if  $c_n = O(n^{-2})$ , then referring to (3.1),

$$\limsup_{n \rightarrow \infty} n^2 \delta_n(t) = L(t).$$

Therefore, if  $L(t) < \frac{1}{16}$ , there exists an  $N \geq 0$  such that  $|c_n| < 1$  and

$$0 < \alpha_n(t) < \frac{1}{4} + \frac{1}{16n(n+1)} = a_n$$

for  $n \geq N$ . By Theorem 2,  $\{\alpha_n^{(N)}(t)\}$  is a chain sequence so  $\psi \in \mathcal{A}$ .

For the (monic) Jacobi polynomials, we have

$$L(-1) = (1 - 4\alpha^2)/16, \quad L(1) = (1 - 4\beta^2)/16,$$

where  $\alpha$  and  $\beta$  are the usual parameters. Thus Theorem 3 does not apply to the Legendre polynomials ( $\alpha = \beta = 0$ ). However, in this case

$$c_n = 0, \quad \lambda_{n+1} = \frac{1}{4} + \frac{1}{4(4n^2 - 1)}.$$

Thus  $\alpha_n^{(1)}(\pm 1) = \lambda_{n+2} < a_n$  so  $\psi \in \mathcal{A}$ . We can, in fact, draw the conclusion that the spectrum of  $\psi$  has at most one point smaller than  $-1$  and one larger than 1. However, we cannot conclude that the true interval of orthogonality is precisely  $[-1, 1]$  by a comparison. This is because  $\{n^2/(4n^2 - 1)\}_{n=1}^\infty$  is a chain sequence that uniquely determines its parameters,  $M_n = n/(2n + 1)$ . Thus no other chain sequence can dominate it (Theorem 2).

We further note we can write

$$\frac{n^2}{4n^2-1} = \frac{1}{4} + \frac{1}{16n^2} \left[ 1 + \frac{1}{4n^2-1} \right].$$

This suggests the following:

LEMMA. *Let*

$$\gamma_n = \frac{1}{4} + \frac{1 + \varepsilon_n}{16n^2}, \quad n \geq 1.$$

If either (i)  $\varepsilon_n = O(n^{-1})$ ,  
 or (ii)  $\sum \varepsilon_n$  converges,  
 then there exists an  $N$  such that  $\{\gamma_n^{(N)}\}_{n=1}^\infty$  is a chain sequence.

*Proof.* In case (i), simple inequalities show that for  $N$  sufficiently large,  $\gamma_n^{(N)} < a_n$  ( $n \geq 1$ ). In case (ii), choose  $N \geq 1$  so that

$$\left| \sum_{\nu=N}^{n+N} \varepsilon_\nu \right| \leq 1, \quad n \geq 0.$$

Define

$$b_n = \frac{1}{2} \sum_{\nu=N}^{n+N} \varepsilon_\nu, \quad g_n = \frac{1}{2} - \frac{1}{4(n+N+b_n)}.$$

We have  $0 \leq g_0 < \frac{1}{2}$ ,  $0 < g_n < \frac{1}{2}$  ( $n \geq 1$ ). Hence we form the chain sequence

$$\begin{aligned} \beta_n &= (1 - g_{n-1})g_n = \frac{1}{4} + \frac{1 + 2(b_n - b_{n-1})}{16(n+N+b_{n-1})(n+N+b_n)} \\ &\geq \frac{1}{4} + \frac{1 + \varepsilon_{n+N}}{16(n+N)^2}. \end{aligned}$$

Thus  $\beta_n \geq \gamma_n^{(N)}$ .

THEOREM 4. *Let  $c_n = O(n^{-2})$ . If, for  $t = \pm 1$ ,*

$$(3.3) \quad 16n^2(\lambda_{n+1} - \frac{1}{4}) + 4tn^2(c_n + c_{n+1}) \leq 1 + r_n(t),$$

where  $r_n(t) = O(n^{-1})$  or  $\sum r_n(t)$  converges, then  $\psi \in \mathcal{A}$ .

*Proof.* If  $c_n = O(n^{-2})$ , we can write

$$[(t - c_n)(t - c_{n+1})]^{-1} = 1 + \theta_n(t), \quad t = \pm 1,$$

where  $\theta_n(t) = O(n^{-2})$ . Therefore, referring to (3.2), we have

$$(3.4) \quad 16n^2\delta_n(t) = 16n^2(\lambda_{n+1} - \frac{1}{4}) + 4tn^2(c_n + c_{n+1}) + F_n(t)$$

where  $F_n(t) = -4n^2c_nc_{n+1} + G_n(t)\theta_n(t)$ , and  $G_n(t)$  is bounded.

Thus if (3.3) holds,  $16n^2\delta_n(t) \leq 1 + \varepsilon_n(t)$ , where  $\varepsilon_n \equiv \varepsilon_n(t) = r_n(t) + F_n(t)$ . Since  $F_n(t) = O(n^{-2})$ ,  $\varepsilon_n$  satisfies the conditions in the preceding lemma. Hence by Theorem 1,  $\psi \in \mathcal{A}$ .

**4. Continuity of  $\psi$  at  $\pm 1$ .** We conclude with a look at conditions that yield the conclusion that  $\psi$  is continuous at  $\pm 1$ .

If  $\{\alpha_n^{(N)}(x)\}_{n=1}^\infty$  is a chain sequence, it follows from [6, Thm. 1] that it has a parameter sequence given by

$$(4.1) \quad g_n(x) = 1 - \frac{P_{N+n+1}(x)}{(x - c_{N+n+1})P_{N+n}(x)}, \quad n \geq 0.$$

In terms of the orthonormal polynomials

$$p_n(x) = (\lambda_1 \lambda_2 \cdots \lambda_{n+1})^{-1/2} P_n(x)$$

(where  $\lambda_1 = (\psi(\infty) - \psi(-\infty))$ ), (4.1) yields

$$\left[ \frac{p_{N+n+1}(x)}{p_{N+n}(x)} \right]^2 = \frac{[1 - g_n(x)]^2 (x - c_{N+n+1})^2}{\lambda_{N+n+2}} = \frac{1 - g_n(x)}{g_{n+1}(x)} \cdot \frac{x - c_{N+n+1}}{x - c_{N+n+2}}.$$

Thus

$$(4.2) \quad \left[ \frac{p_{N+n+1}(x)}{p_{N+n}(x)} \right]^2 = \left[ 1 + \frac{1 - g_n(x) - g_{n+1}(x)}{g_{n+1}(x)} \right] \left[ 1 + \frac{c_{N+n+2} - c_{N+n+1}}{x - c_{N+n+2}} \right].$$

For the remaining theorems we will maintain the preceding notation but will not require the hypothesis (1.4). In fact, the following apply, to a limited extent, to unbounded coefficients in (1.1).

**THEOREM 5.** *Let  $c_n - c_{n-1} = o(n^{-1})$ ,  $\lambda_n \geq \lambda > 0$ , and let  $\{\alpha_n^{(N)}(x)\}_{n=1}^\infty$  be a chain sequence. If*

$$L^*(x) \equiv \limsup_{n \rightarrow \infty} \frac{n[1 - g_n(x) - g_{n+1}(x)]}{g_{n+1}(x)} < -1,$$

then  $\psi$  is continuous at  $x$ . If

$$L_*(x) \equiv \liminf_{n \rightarrow \infty} \frac{n[1 - g_n(x) - g_{n+1}(x)]}{g_{n+1}(x)} > -1,$$

then  $\psi$  has a positive jump at  $x$ .

*Proof.* We note that  $\{c_n\}$  may be unbounded. However, we can write  $c_n = \sum_{k=1}^n k^{-1} \varepsilon_k$  where  $\varepsilon_k \rightarrow 0$ . Thus  $|c_n| \leq M + \log n$  ( $M > 0$ ). Since  $0 < \alpha_{N+n}(x) < 1$  ( $n \geq 1$ ),  $\lambda_{n+1} < [x + M + \log(n+1)]^2$  for  $n \geq N$ . Thus  $\sum \lambda_n^{-1/2} = \infty$  so by a theorem of Carleman (see [11, p. 58]), the associated Hamburger moment problem is determined.

Now since  $\lambda_n$  is bounded away from 0,  $x - c_n$  is bounded away from 0. Hence referring to (4.2), we can write

$$n \left[ \frac{p_{N+n+1}^2(x)}{p_{N+n}^2(x)} - 1 \right] = \frac{n[1 - g_n(x) - g_{n+1}(x)]}{g_{n+1}(x)} [1 + o(1)] + o(1).$$

It now follows from Raabe's test that  $\sum p_n^2(x)$  converges if  $L^*(x) < -1$  and diverges if  $L_*(x) > -1$ . But according to a classical theorem from the problem of moments [11, Cor. 2.6], if the moment problem is determined, then the jump of  $\psi$  at  $x$  is  $\rho(x) \equiv \{\sum_{n=0}^\infty p_n^2(x)\}^{-1}$ .

**THEOREM 6.** *Let  $c_n - c_{n-1} = o(n^{-1})$ ,  $\lambda_n \geq \lambda > 0$ , and let  $\{\alpha_n^{(N)}(x)\}$  be a chain sequence satisfying*

$$\alpha_n^{(N)}(x) \geq \frac{1}{4} - \frac{a}{16n(n+1)}, \quad n \geq 1,$$

where  $0 \leq a < 3$ . Then  $\psi$  is continuous at  $x$ .

*Proof.* Let

$$\sqrt{1+a} - 1 \leq 2s < 1,$$

and set

$$M_{n-1} = \frac{n+s}{2n}, \quad \beta_n = (1 - M_{n-1})M_n, \quad n \geq 1.$$

Then  $\{M_n\}_{n=0}^\infty$  is the maximal parameter sequence for the chain sequence  $\{\beta_n\}_{n=1}^\infty$ . Also

$$\beta_n = \frac{1}{4} - \frac{(1+s)s}{4n(n+1)} \leq \frac{1}{4} - \frac{a}{16n(n+1)}.$$

Therefore  $\alpha_n^{(N)}(x) \geq \beta_n$  so, referring to (4.1) and Theorem 2, we conclude  $g_n(x) \leq M_n$ . Thus

$$\frac{n[1 - g_n(x) - g_{n+1}(x)]}{g_{n+1}(x)} \geq \frac{n}{M_{n+1}}[1 - M_n - M_{n+1}] = -\frac{sn(2n+3)}{(n+1)(n+s+2)}.$$

It now follows that

$$\liminf_{n \rightarrow \infty} \frac{n[1 - g_n(x) - g_{n+1}(x)]}{g_{n+1}(x)} \geq -2s > -1.$$

Thus by Theorem 5,  $\psi$  is continuous at  $x$ .

We note that in the case of the Jacobi polynomials, Theorem 6 yields the conclusion that  $\psi$  is continuous at  $\pm 1$  only for  $|\alpha| < 1, |\beta| < 1$ .

#### REFERENCES

- [1] O. BLUMENTHAL, *Über die Entwicklung einer willkürlichen funktion nach den Nennern des Kettenbruches für  $\int_{-\infty}^\infty [\phi(\xi)/(z - \xi)] d\xi$* , Inaugural Dissertation, Göttingen, 1898.
- [2] K. M. CASE, *Orthogonal polynomials revisited*, Theory and Application of Special Functions, R. Askey, ed., Academic Press, NY, 1975.
- [3] ———, *Orthogonal polynomials from the viewpoint of scattering theory*, J. Math. Phys., 15 (1974), pp. 2166–2174.
- [4] K. M. CASE AND M. KAC, *A discrete version of the inverse scattering problem*, J. Math. Phys., 14 (1973), pp. 594–603.
- [5] T. S. CHIHARA, *Chain sequences and orthogonal polynomials*, Trans. Amer. Math. Soc., 104 (1962), pp. 1–16.
- [6] ———, *Orthogonal polynomials whose zeros are dense in intervals*, J. Math. Anal. Appl., 24 (1968), pp. 362–371.
- [7] ———, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.
- [8] T. S. CHIHARA AND P. G. NEVAI, *Orthogonal polynomials and measures with finitely many point masses*, Numer. Math., submitted.
- [9] J. S. GERONIMO AND K. M. CASE, *Scattering theory and polynomials orthogonal on the real line*, J. Math. Phys., to appear.
- [10] P. G. NEVAI, *Orthogonal Polynomials*, Memoirs Amer. Math. Soc., 18 (1979), 185 pp.
- [11] J. A. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, Mathematical Surveys No. 1, Amer. Math. Soc., New York, 1943.
- [12] H. S. WALL, *Analytic Theory of Continued Fractions*, Van Nostrand, Princeton, NJ, 1948.

## SECONDARY BIFURCATION NEAR A DOUBLE EIGENVALUE\*

M. SHEARER†

**Abstract.** General conditions are formulated under which secondary bifurcation is rigorously established for a family of bifurcation problems depending continuously on a real auxiliary parameter. With more specific conditions, it is shown that, although the presence of secondary bifurcation renders the problem a priori degenerate, a full local bifurcation analysis is still possible.

The results of this paper demonstrate the prime importance of symmetry (or more generally, invariance) to the mechanism by which secondary bifurcation points are created as the auxiliary parameter is varied.

**1. Introduction.** In this paper, we investigate the following suggestion made by Bauer, Keller and Reiss [2] in connection with bifurcation problems for which primary bifurcation points are continuous functions of a real parameter  $\mu$ : If a multiple primary bifurcation point occurs for  $\mu = \mu_0$ , and splits into two or more simple primary bifurcation points as  $\mu$  varies from  $\mu_0$ , then secondary bifurcation may occur for values of  $\mu \neq \mu_0$  near  $\mu_0$ .

We show that for a wide class of problems, the splitting of primary bifurcation points does occur (Theorem 3.2), and formulate conditions under which secondary bifurcation points are created in the process of this splitting (Theorem 3.3). We also give conditions under which no secondary bifurcation points are created, even though the multiple primary bifurcation point splits into two simple primary bifurcation points (Corollary 3.6, and Case CII ( $k = 2$ ), § 4).

Let  $X, Y$  be real Banach spaces, and let  $F: \mathbb{R}^2 \times X \rightarrow Y$  be a mapping of class  $C^n$  for some  $n \geq 2$ . That is,  $F$  is  $n$  times continuously Fréchet differentiable at each point of  $\mathbb{R}^2 \times X$ . We consider such mappings  $F$  for which there is a known solution  $x = \tilde{x}(\lambda, \mu)$  for each  $(\lambda, \mu) \in \mathbb{R}^2$ , of the equation

$$(1.1) \quad F(\lambda, \mu, x) = 0, \quad (\lambda, \mu, x) \in \mathbb{R}^2 \times X$$

and such that  $\tilde{x}: \mathbb{R}^2 \rightarrow X$  is of class  $C^n$ . Without loss of generality, we assume  $\tilde{x}(\lambda, \mu) = \theta$  identically, where  $\theta \in X$  denotes the zero of  $X$ .

$$(H1) \quad F(\lambda, \mu, \theta) = 0 \quad \text{for all } (\lambda, \mu) \in \mathbb{R}^2.$$

Indeed, set  $G(\lambda, \mu, x) = F(\lambda, \mu, \tilde{x}(\lambda, \mu) + x)$ . Then  $G: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  and  $G(\lambda, \mu, \theta) = 0$  identically.

Our next assumption is that for fixed  $\mu = \mu_0$ , the linear operator  $F_x(\lambda, \mu_0, \theta)$  (the Fréchet derivative at  $x = \theta$  of the map  $X \rightarrow Y: x \mapsto F(\lambda, \mu_0, x)$ ) possesses a double degeneracy for some  $\lambda = \lambda_0$ . Without loss of generality, we assume  $\lambda_0 = 0$  and  $\mu_0 = 0$ , as this is equivalent to relabelling  $\lambda - \lambda_0$  as  $\lambda$  and  $\mu - \mu_0$  as  $\mu$ .

(H2)  $F_x(0, 0, \theta): X \rightarrow Y$  is a Fredholm operator with Fredholm index zero and two-dimensional null space.

In order to define secondary bifurcation for (1.1), we specify that  $\lambda$  is the bifurcation parameter, and  $\mu$  is an auxiliary, or perturbation, parameter. With this understanding, we shall principally be concerned with solutions  $(\lambda, x) \in \mathbb{R} \times X$  near  $(0, \theta)$  of (1.1), for each fixed  $\mu \in \mathbb{R}$  near zero. In the following definitions,  $\mu \in \mathbb{R}$  is fixed.

\* Received by the editors September 28, 1978, and in final revised form June 25, 1979.

† Fluid Mechanics Research Institute, University of Essex, Colchester, England. Now at Department of Mathematics, Duke University, Durham, NC 27706. This work was supported by the United States Army under Contract DAAG29-75-C-0024.

A curve  $\mathcal{C}_\mu = \{(\hat{\lambda}(s), \hat{x}(s)) : s \in (-1, 1)\} \subset \mathbb{R} \times X$  is called a *branch of solutions* of (1.1) if  $(\hat{\lambda}, \hat{x}) : (-1, 1) \rightarrow \mathbb{R} \times X$  is one-to-one and continuous, and  $F(\lambda, \mu, x) = 0$  for each  $(\lambda, x) \in \mathcal{C}_\mu$ .

Let  $\mathcal{C}_\mu \subset \mathbb{R} \times X$  be a branch of solutions of (1.1). A point  $(\lambda_1, x_1) \in \mathcal{C}_\mu$  is called a *bifurcation point* (with respect to  $\mathcal{C}_\mu$  and (1.1)) if, for each neighborhood  $U \subset \mathbb{R} \times X$  of  $(\lambda_1, x_1)$ , there exists a solution  $(\lambda, x) \in U \setminus \mathcal{C}_\mu$  of (1.1). In particular, bifurcation points on the set  $\Gamma_\mu = \{(\lambda, \theta) : \lambda \in \mathbb{R}\}$  of *trivial solutions* of (1.1) are called *primary*, and if  $\mathcal{C}_\mu \cap \Gamma_\mu$  is a single (primary bifurcation) point, then  $\mathcal{C}_\mu$  is called a *primary branch* of solutions of (1.1). Bifurcation points on  $\mathcal{C}_\mu \setminus \Gamma_\mu$  are then referred to as *secondary bifurcation points*, and if  $\mathcal{D}_\mu$  is a branch of solutions of (1.1) which intersects  $\mathcal{C}_\mu \setminus \Gamma_\mu$  at a single (secondary bifurcation) point, then  $\mathcal{D}_\mu$  is called a *secondary branch* of solutions of (1.1).

In § 3, we give sufficient conditions for secondary bifurcation. The principal assumption is that  $F$  should satisfy an invariance hypothesis (H3). With this assumption, there are natural nondegeneracy conditions on the lower order derivatives of  $F$  at  $(0, 0, \theta)$ , under which secondary bifurcation is guaranteed for each  $\mu \neq 0$  in an open interval whose closure includes zero.

The conditions for secondary bifurcation in § 3 are intended to cover a wide range of possible symmetry properties for  $F$ . In § 4, we consider various types of such symmetry assumptions on  $F$ , each of which is shown to facilitate a full local bifurcation analysis of (1.1) in a neighborhood of  $(0, 0, \theta)$ . The effect of some of these assumptions is illustrated by an example in § 5. Section 2 consists of preliminary results.

Symmetry properties, of equations of the form (1.1) satisfying (H1), (H2), have been used explicitly in bifurcation analyses of specific buckling problems by List [13], Mallet–Paret [16], and Shearer [18]. In each of these applications, secondary bifurcation occurs for each  $\mu \neq 0$  near zero. Nonsymmetric cases were studied in [13] and [16], by introducing a third parameter, variation of which removes the symmetry. In [13], variation of the third parameter also removes hypothesis (H1), so that secondary bifurcation is not so clearly defined. It is worth noting however, that there are locally no bifurcation points (in our sense), except when the symmetry is present. In [16], variation of the third parameter does not destroy either (H1) or (H2), but does remove the secondary bifurcation.

Further examples of (1.1) satisfying (H1), (H2) and exhibiting secondary bifurcation, are discussed using formal methods in [4], [5], [10]–[12], [15], [20]. Each of these applications possesses some form of symmetry, but this is not fully used in the bifurcation analyses.

The use of symmetry in bifurcation problems has been explored in some generality by Sattinger [17], who remarks that multiple degeneracy (such as (H2)) is often the result of some inherent symmetry in the application being considered.

*Notation.* Subscripts will be used to indicate partial (Fréchet) derivatives. The symbols  $\mathcal{N}(A)$ ,  $\mathcal{R}(A)$  will denote respectively the null space and range of a linear operator  $A$  between Banach spaces. If  $M$  is a  $p$ -linear operator,  $M(x_1, x_2, \dots, x_p)$  will sometimes be written as  $Mx_1x_2 \cdots x_p$ .

*Remark.* While completing this paper, the author learned of the work of Golubitsky and Schaeffer [7]–[9] on bifurcation problems with several parameters. These papers emphasize the role of symmetry in many problems involving bifurcation from a double eigenvalue.

**2. Preliminary results.** To discuss secondary bifurcation for (1.1), we need first to establish a primary branch of solutions for each  $\mu$  near zero. This is achieved by adapting the result of Crandall and Rabinowitz [6, Thm. 1.7] on one-parameter bifurcation from a simple eigenvalue, to include the additional parameter  $\mu$ . We require

the following hypotheses on  $F$ .

- (I1). There exist closed linear subspaces  $X_1$  of  $X$  and  $Y_1$  of  $Y$  such that
  - (i)  $F(U) \subset Y_1$  for some neighborhood  $U \subset \mathbb{R}^2 \times X_1$  of  $(0, 0, \theta)$ .
  - (ii) The restriction  $L: X_1 \rightarrow Y_1$ , of  $F_x(0, 0, \theta)$  to  $X_1$ , is a Fredholm operator with Fredholm index zero and one-dimensional null space.

Since the theorem is concerned only with solutions of (1.1) in  $U$ , we may suppose that  $U = \mathbb{R}^2 \times X_1$ .

- (I2). If  $F$  satisfies (I1), then  $F_{\lambda x}(0, 0, \theta)\phi \notin \mathcal{R}(F_x(0, 0, \theta))$ , where  $\phi \in X_1$  spans  $\mathcal{N}(L)$ .

**THEOREM 2.1.** Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  for some  $n \geq 2$ , and satisfies (H1), (I1), (I2). Let  $Z_1$  be a closed linear subspace of  $X_1$  complementary to  $\text{span}\{\phi\}$ .

Then there exist  $\varepsilon > 0, \delta > 0$  and functions  $\bar{\lambda}, \bar{z}$  from  $D(\delta) = \{(\mu, a) \in \mathbb{R}^2 : |\mu| < \delta, |a| < \delta\}$  into  $\mathbb{R}, Z_1$  respectively such that

- (i) For fixed  $\mu \in (-\delta, \delta)$ , the curve

$$\mathcal{C}_\mu = \{(\bar{\lambda}(\mu, a), a(\phi + \bar{z}(\mu, a))) : |a| < \delta\}$$

is a primary branch of solutions of (1.1).

- (ii)  $\bar{z}(0, 0) = \theta, \bar{\lambda}(0, 0) = 0$ .
- (iii)  $\bar{\lambda}$  and  $\bar{z}$  are of class  $C^{n-1}$ , and of class  $C^n$  on  $\tilde{D}(\delta) = \{(\mu, a) \in D(\delta) : a \neq 0\}$ .
- (iv) If  $(\lambda, \mu, x) \in \mathbb{R}^2 \times X_1$  is a solution of (1.1) satisfying  $|\lambda| < \varepsilon, |\mu| < \delta, \|x\| < \varepsilon$ , then either  $x = \theta$ , or  $(\lambda, x) \in \mathcal{C}_\mu$ .

The proof of the theorem may be found in [19].

Given the hypotheses (H1), (H2), (I1) and (I2), it is convenient to discuss secondary bifurcation from the branches  $\mathcal{C}_\mu$  of Theorem 2.1, by considering the bifurcation equations for (1.1). These equations are derived as follows.

Let  $\mathcal{N} \subset X, \mathcal{R} \subset Y$  denote respectively the null space and range of  $F_x(0, 0, \theta)$ , and let  $Z$  be a closed linear subspace of  $X$  complementary to  $\mathcal{N}$ ,

$$(2.1) \quad X = \mathcal{N} \oplus Z.$$

By (H2), there exist elements  $\psi_1, \psi_2$  of  $Y$  such that

$$(2.2) \quad \text{span}\{\psi_1, \psi_2\} \oplus \mathcal{R} = Y.$$

Set  $Y_0 = \text{span}\{\psi_1, \psi_2\}$ , and let  $\psi'_1, \psi'_2$  be continuous linear functionals on  $Y$  such that

$$(2.3) \quad \langle \psi_i, \psi'_j \rangle = \delta_{ij} \quad \text{and} \quad \langle \psi, \psi'_j \rangle = 0 \quad (j = 1, 2) \quad \text{if } \psi \in \mathcal{R}.$$

Next, define the projection  $P: Y \rightarrow Y_0$  by

$$Py = \langle y, \psi'_1 \rangle \psi_1 + \langle y, \psi'_2 \rangle \psi_2.$$

In later sections, we define more specific  $\psi_1, \psi_2, Z$  satisfying (2.1), (2.2), which in turn define  $Y_0, \psi'_1, \psi'_2$  and  $P$ .

Equation (1.1) is equivalent to the system of equations

$$(2.4) \quad (I - P)F(\lambda, \mu, v + z) = 0, \quad (\lambda, \mu, v, z) \in \mathbb{R}^2 \times \mathcal{N} \times Z,$$

$$(2.5) \quad \langle F(\lambda, \mu, v + z), \psi'_j \rangle = 0, \quad (j = 1, 2) \quad (\lambda, \mu, v, z) \in \mathbb{R}^2 \times \mathcal{N} \times Z.$$

For  $\delta > 0$ , set  $A(\delta) = \{(\lambda, \mu, v) \in \mathbb{R}^2 \times \mathcal{N} : |\lambda| < \delta, |\mu| < \delta, \|v\| < \delta\}$  and  $B(\delta) = \{z \in Z : \|z\| < \delta\}$ .

LEMMA 2.2. Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$ , with  $n \geq 1$ , and satisfies (H1), (H2). Then there exist  $\delta > 0, \varepsilon > 0$  and a function  $\hat{z}: A(\delta) \rightarrow Z$  of class  $C^n$  such that

- (i)  $\hat{z}(\lambda, \mu, \theta) = \theta$  for each  $(\lambda, \mu) \in \mathbb{R}^2, |\lambda| < \delta, |\mu| < \delta$ .
- (ii)  $(I - P)F(\lambda, \mu, v + \hat{z}(\lambda, \mu, v)) = 0$  for each  $(\lambda, \mu, v) \in A(\delta)$ .
- (iii) If  $(\lambda, \mu, v) \in A(\delta)$  and  $z \in B(\varepsilon)$  satisfy (2.4), then  $z = \hat{z}(\lambda, \mu, v)$ .
- (iv)  $\hat{z}_v(0, 0, \theta)v = \theta$  for all  $v \in \mathcal{N}$ .

*Proof.* Since  $(I - P)F(\lambda, \mu, \theta) = 0$  for all  $(\lambda, \mu) \in \mathbb{R}^2$ , and  $F_x(0, 0, \theta)$  is one-to-one and onto between  $Z$  and  $\mathcal{R}$ , (i), (ii), (iii) follow from the implicit function theorem. Property (iv) follows by differentiating (ii) with respect to  $v$ , and setting  $\lambda = \mu = 0, v = \theta$ .

Let  $\phi_1, \phi_2$  span  $\mathcal{N}$ , and set  $v = \alpha\phi_1 + \beta\phi_2$ , with  $(\alpha, \beta) \in \mathbb{R}^2$ . Substituting  $z = \hat{z}(\lambda, \mu, v)$  into (2.5), we obtain the bifurcation equations:

$$(2.6) \quad \langle F(\lambda, \mu, \alpha\phi_1 + \beta\phi_2 + \hat{z}(\lambda, \mu, \alpha\phi_1 + \beta\phi_2)), \psi'_j \rangle = 0, \quad (j = 1, 2).$$

Define  $\mathcal{A}(\delta) = \{(\lambda, \mu, \alpha, \beta) \in \mathbb{R}^4 : (\lambda, \mu, \alpha\phi_1 + \beta\phi_2) \in A(\delta)\}$ , and let  $f_j(\lambda, \mu, \alpha, \beta) \in \mathbb{R}$  denote the left-hand side of (2.6), for  $j = 1, 2$  and  $(\lambda, \mu, \alpha, \beta) \in \mathcal{A}(\delta)$ .

THEOREM 2.3. Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$ , with  $n \geq 1$ , and satisfies (H1), (H2). Then there exists a neighborhood  $U \subset \mathbb{R}^2 \times X$  of  $(0, 0, \theta)$ , and  $\delta_1 > 0$  such that: If  $(\lambda, \mu, x) \in U$  is a solution of (1.1), then there exists a unique element  $(\alpha, \beta) \in \mathbb{R}^2$  such that  $(\lambda, \mu, \alpha, \beta) \in \mathcal{A}(\delta_1), x = \alpha\phi_1 + \beta\phi_2 + \hat{z}(\lambda, \mu, \alpha\phi_1 + \beta\phi_2)$ , and  $f_j(\lambda, \mu, \alpha, \beta) = 0$  ( $j = 1, 2$ ). Moreover,  $F_x(\lambda, \mu, x): X \rightarrow Y$  has a bounded inverse if and only if the  $2 \times 2$  matrix  $[\partial(f_1, f_2)/\partial(\alpha, \beta)]$ , evaluated at  $(\lambda, \mu, \alpha, \beta)$ , is nonsingular.

The first statement follows immediately from Lemma 2.2. The second statement is proved in [19].

**3. Secondary bifurcation.** Throughout this section, we assume  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  for some  $n \geq 2$ , and satisfies (H1), (H2). In the investigation into the possibility of secondary bifurcation for (1.1), the following assumption plays a central role.

(H3). There exist closed subspaces  $X_1, X_2$  of  $X$ , and  $Y_1, Y_2$  of  $Y$  such that

- (i)  $X_1 \oplus X_2 = X, Y_1 \oplus Y_2 = Y$ ;
- (ii)  $F(\mathbb{R}^2 \times X_1) \subset Y_1$ ;
- (iii)  $\mathcal{N} \cap X_i \neq \{\theta\}, (i = 1, 2)$ ;
- (iv)  $Y_i \cap \mathcal{R} \neq Y_i, (i = 1, 2)$ ;
- (v)  $F_x(0, 0, \theta)X_2 \subset Y_2$ .

Note that (H3) implies (I1). However, we wish to apply Theorem 2.1, so we make the further assumption

$$(H4) \quad F_{\lambda x}(0, 0, \theta)v \in \mathcal{R} \text{ and } v \in \mathcal{N} \text{ together imply } v = \theta.$$

Assumption (H4) establishes that  $\lambda$  is to be considered the bifurcation parameter in (1.1). If (H1)–(H4) are satisfied, then the conditions of Theorem 2.1 all hold. We shall freely refer to the primary branches of solutions  $\mathcal{C}_\mu$  predicted by Theorem 2.1.

Since (H3) implies  $F_x(0, 0, \theta)X_i \subset Y_i, (i = 1, 2)$ , let  $A_i: X_i \rightarrow Y_i$  be defined by  $A_i x = F_x(0, 0, \theta)x, x \in X_i, i = 1, 2$ . Then (H2), (H3) imply that  $A_i$  is a Fredholm operator with index zero and null space spanned by an element  $\phi_i \in X_i$  ( $i = 1, 2$ ). Let  $\psi_1 \in Y_1, \psi_2 \in Y_2$  satisfy  $\text{span}\{\psi_i\} \oplus (Y_i \cap \mathcal{R}) = Y_i$  ( $i = 1, 2$ ), and define  $\psi'_1, \psi'_2$  in  $Y'$  by (2.3), which in turn define  $P: Y \rightarrow \text{span}\{\psi_1, \psi_2\}$ . Let  $Z_1, Z_2$  be closed linear subspaces of  $X_1, X_2$  respectively such that  $\text{span}\{\phi_i\} \oplus Z_i = X_i$  ( $i = 1, 2$ ), and set  $Z = Z_1 \oplus Z_2$ . Finally, let  $P_1: Y_1 \rightarrow \text{span}\{\psi_1\}$  be the restriction of  $P$  to  $Y_1$  given by

$$P_1 y = \langle y, \psi'_1 \rangle \psi_1 \quad (y \in Y_1).$$



Since  $(I - P_1)F(\lambda, \mu, x) = (I - P)F(\lambda, \mu, x)$  whenever  $(\lambda, \mu, x) \in \mathbb{R}^2 \times X_1$ , Lemma 2.2 implies that

$$(3.1) \quad \hat{z}(\lambda, \mu, a\phi_1) \in Z_1 \quad \text{for each } (\lambda, \mu, a\phi_1) \in A(\delta).$$

Setting  $\beta = 0$  in (2.6), we have

$$\langle F(\lambda, \mu, \alpha\phi_1 + \hat{z}(\lambda, \mu, \alpha\phi_1)), \psi'_2 \rangle = 0 \quad \text{whenever } (\lambda, \mu, \alpha\phi_1) \in A(\delta),$$

so that

$$(3.2) \quad f_2(\lambda, \mu, \alpha, 0) = 0 \quad \text{identically.}$$

Recall that, provided  $|\lambda|$  and  $|\mu|$  are small enough  $(\lambda, \theta) \in \Gamma_\mu$  is a bifurcation point only if  $F_x(\lambda, \mu, \theta): X \rightarrow Y$  does not possess a bounded inverse. Theorem 2.1 tells us that there is at least one primary bifurcation point  $(\bar{\lambda}(\mu, 0), \theta) \in \Gamma_\mu$  near  $(0, \theta)$ , for each  $\mu \in (-\delta, \delta)$ . The following theorem states that, if (H1)–(H4) are satisfied, then for  $\rho > 0$  ( $\rho < \delta$ ) small enough, and each  $\mu \in (-\rho, \rho)$ , there is at most one other primary bifurcation point near  $(0, \theta)$ .

**THEOREM 3.1.** *Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  with  $n \geq 2$ , and satisfies (H1)–(H4). Then there exist  $\rho > 0$  and functions  $\lambda_1, \lambda_2$  from  $(-\rho, \rho)$  to  $\mathbb{R}$ , each of class  $C^{n-1}$ , such that  $\mathcal{N}(F_x(\lambda_i(\mu), \mu, \theta)) \neq \{\theta\}$  for each  $\mu \in (-\rho, \rho)$  ( $i = 1, 2$ ). If  $(\lambda, \mu) \in \mathbb{R}^2$  satisfies  $|\lambda| < \rho, |\mu| < \rho$ , and does not lie in the set  $\{(\lambda_i(\mu), \mu) : |\mu| < \rho, i = 1, 2\}$ , then  $F_x(\lambda, \mu, \theta): X \rightarrow Y$  has a bounded inverse.*

*Proof.* By Theorem 2.3, provided  $|\lambda| < \delta, |\mu| < \delta, F_x(\lambda, \mu, \theta): X \rightarrow Y$  has a bounded inverse if and only if the matrix  $[\partial(f_1, f_2)/\partial(\alpha, \beta)]$ , evaluated at  $(\lambda, \mu, 0, 0)$ , is nonsingular. But (3.2) implies  $(\partial f_2/\partial \alpha)(\lambda, \mu, 0, 0) = 0$  identically. Therefore, for  $|\lambda| < \delta$  and  $|\mu| < \delta, \mathcal{N}(F_x(\lambda, \mu, \theta)) \neq \{\theta\}$  if and only if either

$$(3.3) \quad \frac{\partial f_1}{\partial \alpha}(\lambda, \mu, 0, 0) = 0$$

or

$$(3.4) \quad \frac{\partial f_2}{\partial \beta}(\lambda, \mu, 0, 0) = 0.$$

By (H2),  $\lambda = \mu = 0$  satisfies both (3.3) and (3.4). Moreover, (H3)(ii) implies  $F_{x\lambda}(0, 0, \theta)\phi_1 \in Y_1$ , so that  $\langle F_{x\lambda}(0, 0, \theta)\phi_1, \psi'_2 \rangle = 0$ , and

$$\frac{\partial^2 f_1}{\partial \alpha \partial \lambda}(0, 0, 0, 0) = \langle F_{x\lambda}(0, 0, \theta)\phi_1, \psi'_1 \rangle \neq 0 \quad \text{by (H4).}$$

But (H4) also implies that  $\psi'_2$  does not annihilate  $F_{x\lambda}(0, 0, \theta)\mathcal{N}$ . Therefore,

$$\frac{\partial^2 f_2}{\partial \beta \partial \lambda}(0, 0, 0, 0) = \langle F_{x\lambda}(0, 0, \theta)\phi_2, \psi'_2 \rangle \neq 0.$$

The result now follows immediately from the implicit function theorem.

Theorem 3.1 does not rule out situations which we wish to consider as being degenerate. For example, set  $X = Y = \mathbb{R}^2$  and define  $F: \mathbb{R}^2 \times X \rightarrow Y$  by

$$F(\lambda, \mu, \alpha, \beta) = \begin{bmatrix} \lambda & \mu \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Then  $F$  is real analytic, and satisfies (H1)–(H4). But the only points for which

$$F_x(\lambda, \mu, 0, 0) = \begin{bmatrix} \lambda & \mu \\ 0 & \lambda \end{bmatrix}$$

is singular are given by  $\lambda = 0$ . So,  $\lambda_i(\mu) = 0, (i = 1, 2)$  for all  $\mu \in \mathbb{R}$ , and  $\mathcal{N}(F_x(0, \mu, 0, 0)) = \text{span} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, (\mu \neq 0)$ .

We now consider hypotheses which guarantee that the primary bifurcation point  $(0, \theta) \in \Gamma_0$  splits into exactly two primary bifurcation points as  $\mu$  varies from zero. This splitting is central to the assertion of Bauer, Keller and Reiss [2] concerning secondary bifurcation.

Define the linear mapping  $z_0: \mathcal{N} \rightarrow \mathcal{Z}$  by  $z_0(v) = \hat{z}_{\mu v}(0, 0, \theta)v$ , or equivalently, differentiating (2.4) with respect to  $\mu$  and  $v$ , with  $z = \hat{z}$  we have

$$(3.5) \quad F_x(0, 0, \theta)z_0(v) + (I - P)F_{\mu x}(0, 0, \theta)v = 0 \quad (v \in \mathcal{N}).$$

Note that  $z_0(\phi_1) \in Z_1$ .

To simplify the notation in the following conditions on  $F$ , set  $L_1 = F_{\lambda x}(0, 0, \theta)$ ,  $M_1 = F_{\mu x}(0, 0, \theta)$ ,  $M_2 = \frac{1}{2}F_{\mu\mu x}(0, 0, \theta)$ , each of which is a linear operator from  $X$  to  $Y$ .

The conditions (D1), (D2) below are clearly mutually exclusive, and are intended to serve two purposes. Firstly, (Dk) ( $k = 1$  or  $2$ ) asserts that  $f_1(\lambda, \mu, \alpha, \beta)$  and  $f_2(\lambda, \mu, \alpha, \beta)$  contain terms involving  $\mu^k \alpha, \mu^k \beta$  respectively, as the lowest order coupled terms between  $\mu$  and  $(\alpha, \beta)$  which are linear in  $(\alpha, \beta)$ . Secondly, (D1) and (D2) are a convenient form of transversality condition which will enable us to use the implicit function theorem. These conditions are intended to be quite general, but they are important to the results in the rest of this section, as the above degenerate example illustrates.

$$(D1) \quad \begin{bmatrix} \langle L_1\phi_1, \psi'_1 \rangle & \langle M_1\phi_1, \psi'_1 \rangle \\ \langle L_1\phi_2, \psi'_2 \rangle & \langle M_1\phi_2, \psi'_2 \rangle \end{bmatrix} \text{ is nonsingular.}$$

$$(D2) \quad \langle M_1v, \psi'_i \rangle = 0 \quad (i = 1, 2) \text{ for all } v \in \mathcal{N} \text{ and}$$

$$\begin{bmatrix} \langle L_1\phi_1, \psi'_1 \rangle & \langle M_2\phi_1 + M_1z_0(\phi_1), \psi'_1 \rangle \\ \langle L_1\phi_2, \psi'_2 \rangle & \langle M_2\phi_2 + M_1z_0(\phi_2), \psi'_2 \rangle \end{bmatrix} \text{ is nonsingular.}$$

As in Theorem 2.1, define  $D(\delta) = \{(\mu, a) \in \mathbb{R}^2 : |\mu| < \delta, |a| < \delta\}$ .

**THEOREM 3.2.** *Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$ , with  $n \geq 3$ , and satisfies (H1)–(H4), together with either (D1) or (D2).*

*Then there exist  $\delta > 0, \varepsilon > 0$  and continuous functions  $\hat{\Lambda}, \hat{x}$  from  $D(\delta)$  to  $\mathbb{R}, X$  respectively such that*

(i)  $\hat{\Lambda}(0, 0) = 0, \hat{x}(0, 0) = \theta$ .

(ii) *For each  $\mu \in (-\delta, \delta), \mu \neq 0$ , the curve  $\mathcal{C}'_\mu = \{(\hat{\Lambda}(\mu, \nu), \hat{x}(\mu, \nu)) : |\nu| < \delta\}$  is a primary branch of solutions of (1.1). In fact  $\hat{x}(\mu, \nu) = \theta$  if and only if  $\nu = 0$ .*

(iii)  $\hat{\Lambda}$  and  $\hat{x}$  are of class  $C^{n-1}$  in the region defined by  $\mu \neq 0$ , of class  $C^N$  in the region defined by  $\nu\mu \neq 0$ , and of class  $C^{n-k-1}$  everywhere if (Dk) is satisfied ( $k = 1$  or  $2$ ).

(iv) *For  $|\mu| < \delta$ , if  $(\lambda, \theta) \in \Gamma_\mu$  is a bifurcation point for (1.1) and  $|\lambda| < \varepsilon$ , then  $\lambda = \bar{\lambda}(\mu, 0)$  or  $\lambda = \hat{\Lambda}(\mu, 0)$ , where  $\bar{\lambda}$  is given in Theorem 2.1.*

(v)  $\mathcal{C}_\mu \cap \mathcal{C}'_\mu = \phi$  if  $0 < |\mu| < \delta, (0, \theta) \in \mathcal{C}_0$ .

*Proof.* For  $i, j = 1, 2$ , set

$$(3.6) \quad a_{ij} = \langle L_1\phi_i, \psi'_j \rangle$$

and

$$(3.7) \quad b_{ij} = \langle M_k \phi_i + (k-1)M_1 z_0(\phi_i), \psi_j' \rangle$$

if (Dk) is satisfied ( $k = 1$  or  $2$ ). Now set  $a = a_{21}b_{22} - a_{22}b_{21}$  and  $b = a_{22}b_{11} - a_{11}b_{22}$ . Then  $b \neq 0$  by assumption.

If (Dk) is satisfied ( $k = 1, 2$ ), let  $h = (h_1, h_2): \mathbb{R}^4 \rightarrow \mathbb{R}^2$  be defined by

$$h_1(\xi, \eta, \mu, \nu) = \begin{cases} \nu^{-1} \mu^{-2k} f_1(\mu^k \xi, \mu, \nu \mu^k(a + \eta), \nu \mu^k b) & \text{if } \nu \mu \neq 0, \\ \mu^{-k} \left[ \frac{\partial f_1}{\partial \alpha}(\mu^k \xi, \mu, 0, 0)(a + \eta) + \frac{\partial f_1}{\partial \beta}(\mu^k \xi, \mu, 0, 0)b \right] & \text{if } \nu = 0, \mu \neq 0, \\ (a_{11}\xi + b_{11})(a + \eta) + (a_{21}\xi + b_{21})b & \text{if } \mu = 0, \end{cases}$$

$$h_2(\xi, \eta, \mu, \nu) = \begin{cases} \nu^{-1} \mu^{-2k} f_2(\mu^k \xi, \mu, \nu \mu^k(a + \eta), \nu \mu^k b) & \text{if } \nu \mu \neq 0, \\ \mu^{-k} \frac{\partial f_2}{\partial \beta}(\mu^k \xi, \mu, 0, 0)b & \text{if } \nu = 0, \mu \neq 0, \\ (a_{22}\xi + b_{22})b & \text{if } \mu = 0. \end{cases}$$

Then  $h(\xi, \eta, \mu, \nu)$  is defined for all  $\xi, \eta$  in bounded intervals, and for all  $\mu, \nu$  sufficiently small. In such a domain,  $h$  is continuous, of class  $C^n$  with respect to  $(\xi, \eta)$ , of class  $C^{n-1}$  away from  $\mu = 0$ , of class  $C^n$  in the region defined by  $\mu\nu \neq 0$ , and of class  $C^{n-k-1}$  everywhere.

Now,  $h_i(-b_{22}/a_{22}, 0, 0, 0) = 0$  ( $i = 1, 2$ ) and  $[(\partial(h_1, h_2)/\partial(\xi, \eta))(-b_{22}/a_{22}, 0, 0, 0)]$  has determinant  $-(a_{11}b_{22} - a_{22}b_{11})^2 \neq 0$ , (by (Dk)). The existence of  $\delta > 0$  and  $\hat{\Lambda}, \hat{x}$  satisfying properties (i)–(iii) follows from the implicit function theorem. Property (iv) is a consequence of Theorem 3.1, once (i)–(iii) have been established, and property (v) is obvious, since if  $0 < |\mu| < \delta$ , then  $\hat{x}(\mu, \nu) \in X_1$  if and only if  $\nu = 0$ . But  $\hat{x}(\mu, 0) = \theta$  identically, and  $\hat{\Lambda}(\mu, 0) = \bar{\lambda}(\mu, 0)$  only if  $\mu = 0$  (by construction). This completes the proof.

To determine secondary bifurcation points on  $\mathcal{C}_\mu$ , define a function  $g_2: \mathbb{R}^4 \rightarrow \mathbb{R}$  by

$$(3.8) \quad g_2(\lambda, \mu, \alpha, \beta) = \begin{cases} \beta^{-1} f_2(\lambda, \mu, \alpha, \beta) & \text{if } \beta \neq 0, \\ \frac{\partial f_2}{\partial \beta}(\lambda, \mu, \alpha, 0), & \text{if } \beta = 0. \end{cases}$$

By (3.2),  $g_2$  is of class  $C^{n-1}$  on  $\mathcal{A}(\delta)$  (with  $\delta > 0$  as in Lemma 2.2), of class  $C^n$  away from  $\beta = 0$ , and of class  $C^n$  with respect to  $(\lambda, \mu, \alpha)$  everywhere.

Suppose that, for fixed  $\mu \in (-\delta, \delta)$ , we can find a sequence  $\{(\lambda_m, \alpha_m, \beta_m)\} \subset \mathbb{R}^3$  of solutions of the equations

$$(3.9) \quad f_1(\lambda, \mu, \alpha, \beta) = 0, \quad g_2(\lambda, \mu, \alpha, \beta) = 0,$$

such that  $|\lambda_m| < \delta$ ,  $\|\alpha_m \phi_1 + \beta_m \phi_2\| < \delta$ ,  $\beta_m \neq 0$  for each  $m$ , and  $(\lambda_m, \alpha_m, \beta_m) \rightarrow (\lambda_0, \alpha_0, 0)$  as  $m \rightarrow \infty$ , with  $\alpha_0 \neq 0$ . Then  $x_0 = \alpha_0 \phi_1 + \hat{z}(\lambda_0, \mu, \alpha_0 \phi_1) \in X_1$ , and  $(\lambda, x) = (\lambda_0, x_0)$  is a solution of (1.1). Therefore, by Theorem 2.1, provided  $\delta > 0$  is small enough,  $(\lambda_0, x_0) \in \mathcal{C}_\mu$ . But  $x_m = \alpha_m \phi_1 + \beta_m \phi_2 + \hat{z}(\lambda_m, \mu, \alpha_m \phi_1 + \beta_m \phi_2)$  is not in  $X_1$  for any  $m$ , and  $(\lambda, x) = (\lambda_m, x_m)$  is a solution of (1.1). Since  $(\lambda_m, x_m) \rightarrow (\lambda_0, x_0) \in \mathcal{C}_\mu$ ,  $x_0 \neq \theta$  and  $(\lambda_m, x_m) \notin \mathcal{C}_\mu$ ,  $(\lambda_0, x_0)$  must be a secondary bifurcation point.

For this argument,  $\beta$  may be considered small compared with  $\alpha$ , in (3.9). We need to examine the lower order terms of the Taylor expansion of  $(f_1, g_2)(\lambda, \mu, \alpha, \beta)$  about  $(0, 0, 0, 0)$ . The Taylor expansion of  $F(\lambda, \mu, x)$  about  $(0, 0, \theta)$  may be written in the

form

$$F(\lambda, \mu, x) = L_0x + \lambda L_1x + \mu M_1x + \mu^2 M_2x + Q_0x^2 + \mu Q_1x^2 + Cx^3 + R(\lambda, \mu, x),$$

where

$$\begin{aligned} L_0 &= F_x(0, 0, \theta), & Q_0 &= \frac{1}{2}F_{xx}(0, 0, \theta), \\ Q_1 &= \frac{1}{2}F_{\mu xx}(0, 0, \theta), & C &= \frac{1}{6}F_{xxx}(0, 0, \theta). \end{aligned}$$

The linear operators  $L_1, M_1, M_2$  were defined earlier. The function  $R: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$ , and represents the remaining terms in the Taylor expansion. In particular,

$$R(\lambda, \mu, x) = o(\|x\|(|\lambda| + \mu^2 + |\mu|\|x\| + \|x\|^2)).$$

Let  $z_1: \mathcal{N} \rightarrow Z$  be the quadratic mapping defined by  $z_1(v, v) = \hat{z}_{vv}(0, 0, \theta)v^2$ , or equivalently, differentiating (2.4) we have

$$(3.10) \quad F_x(0, 0, \theta)z_1(v, v) + (I - P)F_{xx}(0, 0, \theta)v^2 = 0 \quad (v \in \mathcal{N}).$$

Assuming (H1)–(H3) and (Dk) ( $k = 1$  or  $2$ ), (3.9) takes the form

$$(3.11) \quad (a_1\lambda + b_1\mu^k)\alpha + A_1\alpha^2 + B_1\mu\alpha^2 + p\alpha^3 + R_1(\lambda, \mu, \alpha, \beta) = 0,$$

$$(3.12) \quad (a_2\lambda + b_2\mu^k) + A_2\alpha + B_2\mu\alpha + q\alpha^2 + R_2(\lambda, \mu, \alpha, \beta) = 0.$$

The coefficients  $a_i = a_{ii}, b_i = b_{ii}, (i = 1, 2)$  are defined by (3.6), (3.7),

$$\begin{aligned} A_1 &= \langle Q_0\phi_1^2, \psi'_1 \rangle, & A_2 &= \langle Q_0\phi_1\phi_2, \psi'_2 \rangle, \\ B_1 &= \langle Q_1\phi_1^2 + 2Q_0(\phi_1, z_0(\phi_1)) + \frac{1}{2}M_1z_1(\phi_1, \phi_1), \psi'_1 \rangle, \\ B_2 &= 2\langle Q_1\phi_1\phi_2 + Q_0(\phi_1, z_0(\phi_2)) + Q_0(\phi_2, z_0(\phi_1)) + \frac{1}{2}M_1z_1(\phi_1, \phi_2), \psi'_2 \rangle, \\ p &= \langle C\phi_1^3 + Q_0(\phi_1, z_1(\phi_1, \phi_1)), \psi'_1 \rangle, \\ q &= \langle 3C\phi_1^2\phi_2 + Q_0(\phi_1, z_1(\phi_1, \phi_2)) + 2Q_0(\phi_2, z_1(\phi_1, \phi_1)), \psi'_1 \rangle. \end{aligned}$$

$R_1: \mathcal{A}(\delta) \rightarrow \mathbb{R}$  is of class  $C^n$ , and  $R_2: \mathcal{A}(\delta) \rightarrow \mathbb{R}$  has the same differentiability properties as  $g_2: \mathcal{A}(\delta) \rightarrow \mathbb{R}$  defined by (3.8). The term  $(R_1(\lambda, \mu, \alpha, \beta), R_2(\lambda, \mu, \alpha, \beta))$  represents the remaining terms in the Taylor expansion of  $(f_1, g_2)(\lambda, \mu, \alpha, \beta)$  about  $(\lambda, \mu, \alpha, \beta) = (0, 0, 0, 0)$ . In particular, when  $k = 1$ , the terms involving  $(\mu^2\alpha, \mu^2)$  are included in  $(R_1(\lambda, \mu, \alpha, \beta), R_2(\lambda, \mu, \alpha, \beta))$ .

The aim is to obtain quite general sufficient conditions for secondary bifurcation, in essence independent of whether the nonlinearity in  $F$  is quadratic or cubic. The details of the analysis are however different for these two cases, so we distinguish between them with the following mutually exclusive nondegeneracy conditions.

$$(E1) \quad a_1A_2 - a_2A_1 \neq 0,$$

$$(E2) \quad A_1 = A_2 = 0 \quad \text{and} \quad a_1q - a_2p \neq 0.$$

The functions  $z_0, z_1$  defined by (3.5), (3.10) will in general be difficult to calculate, so that it is important to recognize circumstances in which these functions do not affect the hypotheses (D2), (E2) respectively. Such conditions are as follows:

$$(i) \quad PM_1x = 0 \quad \text{for all } x \in X, \quad \text{or} \quad M_1v = 0 \quad \text{for all } v \in \mathcal{N}.$$

Condition (D2) becomes

$$\begin{vmatrix} \langle L_1\phi_1, \psi'_1 \rangle & \langle M_2\phi_1, \psi'_1 \rangle \\ \langle L_1\phi_2, \psi'_2 \rangle & \langle M_2\phi_2, \psi'_2 \rangle \end{vmatrix} \neq 0$$

(ii)  $PQ_0x^2 = 0$  for all  $x \in X$ , or  $Q_0v^2 = 0$  for all  $v \in \mathcal{N}$ .

Condition (E2) then reduces to

$$\begin{vmatrix} \langle L_1\phi_1, \psi'_1 \rangle & \langle C\phi_1^3, \psi'_1 \rangle \\ \langle L_1\phi_2, \psi'_2 \rangle & \langle 3C\phi_1^2\phi_2, \psi'_2 \rangle \end{vmatrix} \neq 0.$$

Conditions like (i) and (ii) have been assumed by other authors (notably in [14]) to avoid having to include implicitly defined functions in hypotheses to guarantee bifurcation.

As hypotheses (Dk), (Ek), ( $k = 1, 2$ ) are concerned only with coefficients in the bifurcation equations, we shall assume that these coefficients may be determined, at least to a degree of accuracy to guarantee the particular set of conditions required of a specific  $F$ . The example considered in § 5 includes a situation in which the function  $z_1: \mathcal{N} \rightarrow Z$  plays an important part in the bifurcation analysis.

**THEOREM 3.3.** *Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  with  $n \geq 3$ , and satisfies (H1)–(H4), (D1) or (D2) and (E1) or (E2). If  $F$  satisfies (D2) and (E2), assume the additional condition*

$$(3.13) \quad (a_2B_1 - a_1B_2)^2 > 4(a_2b_1 - a_1b_2)(a_2p - a_1q).$$

*Then there exists an open interval  $I$  containing zero and continuous functions  $\hat{\lambda}: I \times I \rightarrow \mathbb{R}$ ,  $\hat{\mu}: I \rightarrow \mathbb{R}$ ,  $w: I \times I \rightarrow X$  such that*

(i)  $|\hat{\mu}(\nu)| = \nu^2$  if (D1) and (E2) are satisfied, otherwise  $\hat{\mu}(\nu) = \nu$ , ( $\nu \in I$ ).

(ii) *For fixed  $\nu \in I$ ,  $\nu \neq 0$ , set  $\mu = \hat{\mu}(\nu)$ . The curve  $\mathcal{D}_\mu = \{(\hat{\lambda}(\nu, y), w(\nu, y)) : y \in I\}$  is a secondary branch of solutions of (1.1), intersecting  $\mathcal{C}_\mu$  at the single point  $(\hat{\lambda}(\nu, 0), w(\nu, 0))$ .*

(iii)  $(0, \theta) \in \mathcal{D}_0$  (corresponding to  $\nu = 0$ ).

*Proof.* For  $j = 1$  or  $2$  and  $k = 1$  or  $2$ , suppose (Dj) and (Ek) are satisfied. Let  $(G_1, G_2): \mathbb{R}^3 \rightarrow \mathbb{R}^2$  be the polynomial mapping defined by

$$G_1(\Lambda, \tau, x) = (a_1\Lambda + b_1\tau^j)x + A_1x^2 + (j-1)(k-1)B_1\tau x^2 + (k-1)px^3,$$

$$G_2(\Lambda, \tau, x) = (a_2\Lambda + b_2\tau^j) + A_2x + (j-1)(k-1)B_2\tau x + (k-1)qx^2.$$

To express the fact that  $G_1(\lambda, \mu, \alpha)$ ,  $G_2(\lambda, \mu, \alpha)$  represent the dominant terms in (3.11), (3.12) when  $\beta = 0$ , we introduce the following rescaling functions.

$$\lambda^*(\nu, \Lambda) = \begin{cases} \nu\Lambda & \text{if (D1), (E1),} \\ \nu^2\Lambda & \text{otherwise,} \end{cases}$$

$$\mu^*(\nu, \tau) = \begin{cases} \nu^2\tau & \text{if (D1), (E2),} \\ \nu\tau & \text{otherwise,} \end{cases}$$

$$\alpha^*(\nu, x) = \begin{cases} \nu^2x & \text{if (D2), (E1),} \\ \nu x & \text{otherwise,} \end{cases}$$

$$\beta^*(\nu, y) = \begin{cases} \nu^2y & \text{if (D2), (E1),} \\ \nu y & \text{otherwise.} \end{cases}$$

Here,  $\nu \in \mathbb{R}$  is the rescaling parameter, and  $\Lambda, \tau, x, y$  are all real.

Now substitute  $\lambda = \lambda^*, \mu = \mu^*, \alpha = \alpha^*, \beta = \beta^*$  into (3.11), (3.12).

$$(3.14) \quad \begin{aligned} f_1(\lambda^*(\nu, \Lambda), \mu^*(\nu, \tau), \alpha^*(\nu, x), \beta^*(\nu, y)) \\ = \nu^M \{G_1(\Lambda, \tau, x) + h_1(\Lambda, \tau, x, y)y + \eta_1(\Lambda, \tau, x, y, \nu)\}, \end{aligned}$$

$$(3.15) \quad \begin{aligned} g_2(\lambda^*(\nu, \Lambda), \mu^*(\nu, \tau), \alpha^*(\nu, x), \beta^*(\nu, y)) \\ = \nu^N \{G_2(\Lambda, \tau, x) + h_2(\Lambda, \tau, x, y)y + \eta_2(\Lambda, \tau, x, y, \nu)\}. \end{aligned}$$

Here,

$$M = \begin{cases} 2 & \text{if (D1), (E1),} \\ 3 & \text{if (E2),} \\ 4 & \text{if (D2), (E1)} \end{cases}$$

and

$$N = \begin{cases} 1 & \text{if (D1), (E1),} \\ 2 & \text{otherwise.} \end{cases}$$

$h_1(\Lambda, \tau, x, y), h_2(\Lambda, \tau, x, y)$  are polynomials, and  $\eta_i: \mathbb{R}^5 \rightarrow \mathbb{R} \ (i = 1, 2)$  is defined on a set of the form

$$S = \{(\Lambda, \tau, x, y, \nu) \in \mathbb{R}^5 : (\lambda^*(\nu, \Lambda), \mu^*(\nu, \tau), \alpha^*(\nu, x), \beta^*(\nu, y)) \in \mathcal{A}(\delta)\}.$$

$(\eta_1, \eta_2): S \rightarrow \mathbb{R}^2$  is continuous, and possesses continuous partial derivatives with respect to  $(\Lambda, x)$ . Moreover,  $\eta_1, \eta_2$  are of class  $C^n, C^{n-1}$  respectively away from  $\nu = 0$ , and  $\eta_2$  is of class  $C^n$  away from  $\nu y = 0$ . We also have  $\eta_1(\Lambda, \tau, x, y, 0) = \eta_2(\Lambda, \tau, x, y, 0) = 0$  identically.

Now divide the expressions (3.14), (3.15) by  $\nu^M, \nu^N$  respectively. That is, define  $(H_1, H_2): S \rightarrow \mathbb{R}^2$  by

$$\begin{aligned} H_1(\Lambda, \tau, x, y, \nu) &= \begin{cases} \nu^{-M} f_1(\lambda^*(\nu, \Lambda), \mu^*(\nu, \tau), \alpha^*(\nu, x), \beta^*(\nu, y)) & \text{if } \nu \neq 0, \\ G_1(\Lambda, \tau, x) + y h_1(\Lambda, \tau, x, y) & \text{if } \nu = 0, \end{cases} \\ H_2(\Lambda, \tau, x, y, \nu) &= \begin{cases} \nu^{-N} g_2(\lambda^*(\nu, \Lambda), \mu^*(\nu, \tau), \alpha^*(\nu, x), \beta^*(\nu, y)) & \text{if } \nu \neq 0 \\ G_2(\Lambda, \tau, x) + y h_2(\Lambda, \tau, x, y) & \text{if } \nu = 0. \end{cases} \end{aligned}$$

Then  $H_i: S \rightarrow \mathbb{R} \ (i = 1, 2)$  has the same differentiability properties as described above for the corresponding  $\eta_i$ .

Under the conditions of the theorem, we shall show that the equations

$$(3.16) \quad G_1(\Lambda, \tau, x) = 0, \quad G_2(\Lambda, \tau, x) = 0$$

possess a solution  $(\Lambda_0, \tau_0, x_0)$  such that  $x_0 \neq 0$  and the matrix  $[\partial(G_1, G_2)/\partial(\Lambda, x)]$ , evaluated at  $(\Lambda_0, \tau_0, x_0)$ , is nonsingular.

But  $H_i(\Lambda, \tau, x, 0, 0) = G_i(\Lambda, \tau, x)$  identically, so we may apply the implicit function theorem to the equations

$$H_1(\Lambda, \tau, x, y, \nu) = 0, \quad H_2(\Lambda, \tau, x, y, \nu) = 0.$$

This implies that there exist  $\rho > 0$  and continuous functions  $\hat{\Lambda}, \hat{x}$  from  $I \times I$  to  $\mathbb{R}$  (where

$I = (-\rho, \rho)$ , satisfying

- (i)  $\hat{\Lambda}(0, 0) = \Lambda_0, \hat{x}(0, 0) = x_0$ ;
- (ii)  $\hat{\Lambda}$  and  $\hat{x}$  are of class  $C^{n-M}$  (if  $n \geq M$ ), of class  $C^{n-1}$  on  $(I \times I) \setminus (\{0\} \times I)$ , and of class  $C^n$  on  $(I \times I) \setminus [(\{0\} \times I) \cup (I \times \{0\})]$ ;
- (iii)  $H_i(\hat{\Lambda}(\nu, y), \tau_0, \hat{x}(\nu, y), y, \nu) = 0, (i = 1, 2)$  for all  $(\nu, y) \in I \times I$ .

Let  $\hat{\lambda}(\nu, y) = \lambda^*(\nu, \hat{\Lambda}(\nu, y)), \hat{\mu}(\nu) = \mu^*(\nu, \tau_0), \hat{v}(\nu, y) = \alpha^*(\nu, \hat{x}(\nu, y))\phi_1 + \beta^*(\nu, y)\phi_2$  and  $w(\nu, y) = \hat{v}(\nu, y) + \hat{z}(\hat{\lambda}(\nu, y), \hat{\mu}(\nu), \hat{v}(\nu, y))$ . Then  $F(\hat{\lambda}(\nu, y), \hat{\mu}(\nu), w(\nu, y)) = 0$  for all  $(\nu, y) \in I \times I$ , and  $w(\nu, y) \in X_1$  if and only if  $\nu y = 0$ . In particular, if  $\rho > 0$  is chosen sufficiently small, Theorem 2.1 implies  $(\hat{\lambda}(\nu, 0), w(\nu, 0)) \in \mathcal{C}_\mu$  when  $\mu = \hat{\mu}(\nu)$  and  $|\nu| < \rho$ . Moreover,  $w(\nu, 0) = \theta, |\nu| < \rho$  if and only if  $\nu = 0$ , so that  $(\hat{\lambda}(\nu, 0), w(\nu, 0))$  is a secondary bifurcation point whenever  $0 < |\nu| < \rho$ . Note also that  $\mathcal{D}_\mu = \{(\hat{\lambda}(\nu, y), w(\nu, y)) : |y| < \rho\}$  (with  $\mu = \hat{\mu}(\nu)$ ) is  $C^{n-1}$  if  $\mu \neq 0$ , and  $\mathcal{D}_\mu \setminus \{(\hat{\lambda}(\nu, 0), w(\nu, 0))\}$  is  $C^n$ . Also,  $\mathcal{D}_0 = \{(0, \theta)\}$ .

It remains only to solve (3.16) in the manner indicated above, and with  $\tau_0 = 1$  unless (D1) and (E2) hold, in which case  $|\tau_0| = 1$ . If  $(\Lambda_0, \tau_0, x_0) \in \mathbb{R}^3$  is a solution of (3.16), let  $J_0$  denote the determinant of the  $2 \times 2$  matrix  $[\partial(G_1, G_2)/\partial(\Lambda, x)]$ , evaluated at  $(\Lambda_0, \tau_0, x_0)$ . We consider separately the different cases covered by the theorem.

Cases (D1), (E1) and (D2), (E1) differ only in the definition of  $b_1, b_2$ . Set  $\tau_0 = 1$ . Then (3.16) is solved by  $\Lambda_0 = (A_2b_1 - A_1b_2)/(A_2a_1 - A_1a_2), x_0 = (b_1a_2 - b_2a_1)/(A_2a_1 - A_1a_2)$ . By assumption,  $x_0 \neq 0$ , and  $J_0 = b_1a_2 - b_2a_1 \neq 0$ .

If (D1) and (E2) hold, set  $\tau_0 = -\text{sgn}(b_1a_2 - b_2a_1)(pa_2 - qa_1)$ . Then  $\Lambda_0 = \tau_0(b_1q - b_2p)/(pa_2 - qa_1), x_0 = |(b_1a_2 - b_2a_1)/(pa_2 - qa_1)|^{1/2}$ . Again,  $x_0 \neq 0$  and  $J_0 = 2(a_1b_2 - a_2b_1)\tau_0 \neq 0$ .

Finally, suppose (D2) and (E2) hold, together with (3.13). Then  $d = (a_2B_1 - a_1B_2)^2 - 4(a_2b_1 - a_1b_2)(a_2p - a_1q) > 0$ . Setting  $\tau_0 = 1$ , we require  $x_0 = \frac{1}{2}(a_1B_2 - a_2B_1 \pm \sqrt{d})/(a_2p - a_1q)$  and  $\Lambda_0 = (b_1q - b_2p + x_0(B_1q - B_2p))/(a_2p - a_1q)$ . Then  $x_0 \neq 0$  and  $J_0 = 2(a_2b_1 - a_1b_2) + (a_2B_1 - a_1B_2)x_0$ . In particular, (D2) and (3.13) imply  $J_0 \neq 0$ .  $\square$

If (D2), (E2) and (3.13) hold, then the definition of  $x_0$  in the proof of Theorem 3.3 involves a choice of sign. Therefore, we have the immediate corollary.

**COROLLARY 3.4.** *Under the conditions of Theorem 3.3, if (D2), (E2) and (3.13) all hold, then there are two functions  $(\lambda_1, w_1), (\lambda_2, w_2)$  from  $I \times I$  to  $\mathbb{R} \times X$  satisfying the conclusions of Theorem 3.3. In particular, there are two secondary bifurcation points on  $\mathcal{C}_\mu$  for each  $\mu \neq 0$  in  $I$ .*

The proof of the following completeness result [19] is fairly straightforward, but we omit it here, as it is also quite long and technical.

**COROLLARY 3.5.** *Under the conditions of Theorem 3.3, there exists  $\gamma > 0$  such that if  $|\mu| < \gamma, |\lambda| < \gamma, \|x\| < \gamma$ , and  $(\lambda, x) \in \mathcal{C}_\mu$  is a secondary bifurcation point, then  $(\lambda, x)$  is one of the secondary bifurcation points determined in Theorem 3.3 and Corollary 3.4.*

If the inequality (3.13) is reversed, then

$$(a_2b_1 - a_1b_2) + (a_2B_1 - a_1B_2)x + (a_2p - a_1q)x^2 = 0$$

has no solutions. Consequently, under (D2), (E2), there are no solutions of (3.16) with  $\tau = 1$ . This leads to the following result.

**COROLLARY 3.6.** *Suppose  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^3$ , and satisfies (H1)–(H4), (D2), (E2), together with*

$$(3.17) \quad (a_2B_1 - a_1B_2)^2 < 4(a_2b_1 - a_1b_2)(a_2p - a_1q)$$

*then there exists  $\rho > 0$  such that if  $(\lambda, x) \in \mathcal{C}_\mu$ , with  $|\lambda| < \rho, \|x\| < \rho, |\mu| < \rho$ , then  $(\lambda, x)$  is not a secondary bifurcation point.*

**4. Bifurcation and symmetry.** In applications, condition (H3) is often satisfied as a consequence of a symmetry in the physical problem associated with (1.1). In this section we list various such symmetries as conditions on  $F$ . These conditions, together with appropriate nondegeneracy conditions (such as (E1), (E2)), facilitate a full bifurcation analysis of (1.1) in a neighborhood of  $(0, 0, \theta)$ . Throughout the section we assume that  $F: \mathbb{R}^2 \times X \rightarrow Y$  is of class  $C^n$  with  $n \geq 4$ , and that  $F$  satisfies (H1), (H2).

(S1). There exist linear operators  $S: X \rightarrow X, T: Y \rightarrow Y$  such that

- (i)  $S(Sx) = x, T(Ty) = y$  for all  $x \in X, y \in Y$ ;
- (ii)  $F(\lambda, \mu, Sx) = TF(\lambda, \mu, x)$  for all  $(\lambda, \mu, x) \in \mathbb{R}^2 \times X$ ;
- (iii) there exist  $\phi_1, \phi_2$  which span  $\mathcal{N}$  such that

$$S\phi_1 = \phi_1, \quad S\phi_2 = -\phi_2;$$

(iv) there exist  $\psi_1, \psi_2$  in  $Y \setminus \mathcal{R}$  such that

$$T\psi_1 = \psi_1, \quad T\psi_2 = -\psi_2.$$

If  $F$  satisfies (S1), set

$$X_{\pm} = \{x \in X: Sx = \pm x\}, \quad Y_{\pm} = \{y \in Y: Ty = \pm y\}.$$

Then  $X = X_+ \oplus X_-, Y = Y_+ \oplus Y_-$ . Set  $Z = Z_+ \oplus Z_-$ , where  $Z_+ \oplus \text{span}\{\phi_1\} = X_+, Z_- \oplus \text{span}\{\phi_2\} = X_-$ .

Let  $T': Y' \rightarrow Y'$  be the adjoint of  $T$ ,

$$\langle Ty, \psi' \rangle = \langle y, T'\psi' \rangle \quad \text{if } y \in Y, \psi' \in Y'$$

and note that  $\text{span}\{\psi'_1, \psi'_2\}$ , defined by (2.3), is invariant under  $T'$ . In fact,  $T'\psi'_1 = \psi'_1$  and  $T'\psi'_2 = -\psi'_2$ . Therefore,  $P: Y \rightarrow \text{span}\{\psi_1, \psi_2\}$  commutes with  $T: PT = TP$ . Lemma 2.2 now implies

$$(4.1) \quad S\hat{z}(\lambda, \mu, v) = \hat{z}(\lambda, \mu, Sv) \quad \text{if } (\lambda, \mu, v) \in A(\delta)$$

which in turn implies that the bifurcation equations (2.6) satisfy

$$(4.2) \quad f_1(\lambda, \mu, \alpha, -\beta) = f_1(\lambda, \mu, \alpha, \beta),$$

$$(4.3) \quad f_2(\lambda, \mu, \alpha, -\beta) = -f_2(\lambda, \mu, \alpha, \beta)$$

for all  $(\lambda, \mu, \alpha, \beta) \in \mathcal{A}(\delta)$ .

*Remark.* In the bifurcation analysis under (S1), we shall only use the fact that (S1) implies (4.2), (4.3). Therefore, we could replace (S1) by the weaker condition that (4.2), (4.3) hold. However,  $(f_1, f_2)$  is only implicitly defined, so that it seems more satisfactory to interpret the symmetry (4.2), (4.3) in terms of a corresponding symmetry (S1) in  $F$ . A similar remark applies to the ‘‘double’’ symmetry assumption (S2) that follows.

(S2) There exist linear operators  $S_1, S_2$  on  $X$ , and  $T_1, T_2$  on  $Y$  such that

- (i)  $S_i(S_i x) = x, T_i(T_i y) = y$  for all  $x \in X, y \in Y, i = 1, 2$ ;
- (ii)  $S_1 S_2 = S_2 S_1, T_1 T_2 = T_2 T_1$ ;
- (iii)  $F(\lambda, \mu, S_i x) = T_i F(\lambda, \mu, x), i = 1, 2$ , for all  $(\lambda, \mu, x)$  in  $\mathbb{R}^2 \times X$ ;
- (iv) there exist  $\phi_1, \phi_2$  which span  $\mathcal{N}$  and satisfy

$$S_i \phi_j = (-1)^{i+j} \phi_j \quad (i, j = 1, 2);$$

(v) there exist  $\psi_1, \psi_2$  in  $Y \setminus \mathcal{R}$  such that

$$T_i \psi_j = (-1)^{i+j} \psi_j \quad (i, j = 1, 2);$$



If  $F$  satisfies (S2), set

$$X_{\rho\sigma} = \{x \in X : S_1x = \rho x, S_2x = \sigma x\},$$

$$Y_{\rho\sigma} = \{y \in Y : T_1y = \rho y, T_2y = \sigma y\} \quad (\rho, \sigma = \pm).$$

Then  $X = X_{++} \oplus X_{+-} \oplus X_{-+} \oplus X_{--}$ , and  $Y = Y_{++} \oplus Y_{+-} \oplus Y_{-+} \oplus Y_{--}$ . Additionally,  $\phi_1 \in X_{+-}$ ,  $\phi_2 \in X_{-+}$ ,  $\psi_1 \in Y_{+-}$ ,  $\psi_2 \in Y_{-+}$ . Set  $Z = Z_1 \oplus Z_2 \oplus X_{++} \oplus X_{--}$  where  $Z_1 \oplus \text{span}\{\phi_1\} = X_{+-}$ ,  $Z_2 \oplus \text{span}\{\phi_2\} = X_{-+}$ . Repeating the argument under (S1), we have

$$(4.4) \quad \hat{z}(\lambda, \mu, S_i v) = S_i \hat{z}(\lambda, \mu, v) \quad (i = 1, 2), \quad (\lambda, \mu, v) \in A(\delta)$$

so that (4.2), (4.3) hold, together with

$$(4.5) \quad f_1(\lambda, \mu, -\alpha, \beta) = -f_1(\lambda, \mu, \alpha, \beta),$$

$$(4.6) \quad f_2(\lambda, \mu, -\alpha, \beta) = f_2(\lambda, \mu, \alpha, \beta).$$

A particular case of (S2) occurs when (S1) is satisfied, and  $F(\lambda, \mu, x)$  is odd with respect to  $x$ ,

$$(4.7) \quad F(\lambda, \mu, -x) = -F(\lambda, \mu, x)$$

for all  $(\lambda, \mu, x) \in \mathbb{R}^2 \times X$ . In fact, set  $S_1 = S$ ,  $S_2 = -S$ ,  $T_1 = T$ ,  $T_2 = -T$ . Then (S1), (4.7) imply (S2) with  $X_{++} = X_{--} = \{\theta\}$ ,  $Y_{++} = Y_{--} = \{0\}$ , and  $\hat{z}(\lambda, \mu, -v) = -\hat{z}(\lambda, \mu, v)$  for all  $(\lambda, \mu, v) \in A(\delta)$ .

Now, if  $F$  satisfies (S1) or (S2), then (H4) is satisfied if and only if

$$(H4') \quad \langle F_{\lambda x}(0, 0, \theta)\phi_i, \psi'_i \rangle \neq 0 \quad (i = 1, 2).$$

For the rest of this section, we assume (H4').

Note that if  $F$  satisfies (S1), then (H3) is satisfied (which implies (I1)), with  $X_1 = X_+$ ,  $X_2 = X_-$ ,  $Y_1 = Y_+$ ,  $Y_2 = Y_-$ . Therefore, by Theorem 2.1 there is a primary branch  $\mathcal{C}_\mu \subset \mathbb{R} \times X_+$  of solutions of (1.1), for each small  $|\mu|$ . These primary branches correspond to the fact that  $f_2(\lambda, \mu, \alpha, 0) = 0$  identically, by (4.3).

Similarly, if  $F$  satisfies (S2), then (H3) is satisfied with  $X_1 = X_{++} \oplus X_{+-}$ ,  $X_2 = X_{-+} \oplus X_{--}$  (and similarly for  $Y_1, Y_2$ ) and also with  $X_1 = X_{++} \oplus X_{-+}$ ,  $X_2 = X_{+-} \oplus X_{--}$  (and similarly for  $Y_1, Y_2$ ). Therefore, for each  $\mu$  near zero, Theorem 2.1 establishes the existence of two primary branches,  $\mathcal{C}_\mu \subset \mathbb{R} \times (X_{++} \oplus X_{+-})$ , and  $\mathcal{C}'_\mu \subset \mathbb{R} \times (X_{-+} \oplus X_{--})$ . These branches correspond to setting  $\beta = 0$  and  $\alpha = 0$  respectively in the bifurcation equations

$$(4.8) \quad f_1(\lambda, \mu, \alpha, \beta) = 0, \quad f_2(\lambda, \mu, \alpha, \beta) = 0.$$

To obtain the branches  $\mathcal{C}_\mu, \mathcal{C}'_\mu$ , we need only assure that

$$(4.9) \quad f_1(\lambda, \mu, \alpha, 0) = 0 \quad \text{and} \quad f_2(\lambda, \mu, 0, \beta) = 0 \quad \text{identically.}$$

This condition is provided by the following assumption.

(J1). There exist closed linear subspaces  $X^{(0)}, X^{(1)}, X^{(2)}$  of  $X$  and  $Y^{(0)}, Y^{(1)}, Y^{(2)}$  of  $Y$  such that

$$X = X^{(0)} \oplus X^{(1)} \oplus X^{(2)}, \quad Y = Y^{(0)} \oplus Y^{(1)} \oplus Y^{(2)},$$

$$F(\mathbb{R}^2 \times (X^{(0)} \oplus X^{(i)})) \subset Y^{(0)} \oplus Y^{(i)} \quad (i = 1, 2),$$

$$\mathcal{N} = \text{span}\{\phi_1, \phi_2\} \quad \text{with} \quad \phi_i \in X^{(i)} \quad (i = 1, 2),$$

$$\mathcal{R} \oplus \text{span}\{\psi_1, \psi_2\} = Y \quad \text{with} \quad \psi_i \in Y^{(i)} \quad (i = 1, 2).$$

Suppose  $F$  satisfies (J1), and set  $Z = X^{(0)} \oplus Z^{(1)} \oplus Z^{(2)}$ , where  $Z^{(i)} \oplus \text{span}\{\phi_i\} = X^{(i)}$  ( $i = 1, 2$ ). Restricting (2.4) to  $\mathbb{R}^2 \times \text{span}\{\phi_i\} \times (X^{(0)} \oplus Z^{(i)})$ , for  $i = 1, 2$  in turn, we see that  $\hat{z}(\lambda, \mu, \alpha\phi_i) \in X^{(0)} \oplus Z^{(i)}$ , ( $i = 1, 2$ ) for all  $\lambda, \mu, \alpha$  near zero. This implies (4.9).

For a full bifurcation analysis under (J1), we require the following additional conditions.

$$(J2) \quad F(\lambda, \mu, -x) = -F(\lambda, \mu, x) \quad \text{for all } (\lambda, \mu, x) \in \mathbb{R}^2 \times X,$$

$$(J3) \quad \langle F_{xxx}(0, 0, \theta)\phi_i^2\phi_j\psi'_i \rangle = 0 \quad \text{if } (i, j) = (1, 2) \text{ or } (i, j) = (2, 1).$$

**Bifurcation analysis under (S1).** Suppose  $F$  satisfies (H1), (H2), (S1), and (Dk),  $k = 1$  or  $2$ .

The bifurcation equations (4.8) satisfy (4.2), (4.3), and are of the form

$$(4.10) \quad (a_1\lambda + b_1\mu^k)\alpha + A\alpha^2 + C\beta^2 + h_1(\lambda, \mu, \alpha, \beta) = 0$$

$$(4.11) \quad (a_2\lambda + b_2\mu^k)\beta + B\alpha\beta + \beta h_2(\lambda, \mu, \alpha, \beta) = 0,$$

where  $a_i, b_i$ , ( $i = 1, 2$ ),  $A = A_1$ ,  $B = A_2$  are defined in § 3 and  $C = \langle Q_0\phi_2^2, \psi'_1 \rangle$ ; the mapping  $(h_1, \beta h_2): \mathcal{A}(\delta) \rightarrow \mathbb{R}^2$  is of class  $C^n$ , and represents higher order terms in  $(\lambda, \mu, \alpha, \beta)$ ,

$$|\beta^{i-1}h_i(\lambda, \mu, \alpha, \beta)| \leq \text{const.}\{(|\alpha, \beta|)^3 + (|\lambda| + |\mu|)(|\alpha, \beta|)^2 + (|\lambda|^2 + |\mu|^{k+1})(|\alpha, \beta|)\}$$

( $i = 1, 2$ ), where  $|\alpha, \beta| = (\alpha^2 + \beta^2)^{1/2}$ .

In terms of (4.10), (4.11), we assume

$$(S1) \quad h_i(\lambda, \mu, \alpha, -\beta) = h_i(\lambda, \mu, \alpha, \beta) \quad (i = 1, 2),$$

$$(H4) \quad a_1a_2 \neq 0,$$

$$(Dk) \quad a_1b_2 \neq a_2b_1,$$

$$(E1) \quad a_1B \neq a_2A \quad \text{and the additional condition } BC \neq 0.$$

*Remark.* We do not assume  $A \neq 0$ . However, if  $A \neq 0$ , this provides information about the direction of bifurcation of the primary branch  $\mathcal{E}_\mu$  from  $(\bar{\lambda}(\mu, 0), \theta) \in \Gamma_\mu$ , where  $\mathcal{E}_\mu, \bar{\lambda}$  are given by Theorem 2.1. In fact,

$$\text{sgn } \bar{\lambda}'_\alpha(\mu, 0) = -\text{sgn}(a_1A) \neq 0$$

provided  $A \neq 0$  and  $|\mu|$  is sufficiently small. So,  $A \neq 0$  implies that, for each  $\mu$  near zero, primary bifurcation from  $(\bar{\lambda}(\mu, 0), \theta)$  is *transcritical* (or *two-sided*, or *asymmetric*).

The solution  $\beta = 0$  of (4.11) corresponds to the trivial solution  $\Gamma_\mu$  and the primary branches  $\mathcal{E}_\mu$  of solutions of (1.1). By Theorem 2.1, we lose no additional solutions of (1.1) near  $(0, 0, \theta)$ , by dividing (6.11) by  $\beta$ . Disregarding the higher order terms  $(h_1, h_2)$ , we are left with

$$(4.12) \quad \begin{aligned} (a_1\lambda + b_1\mu^k)\alpha + A\alpha^2 + C\beta^2 &= 0, \\ (a_2\lambda + b_2\mu^k) + B\alpha &= 0. \end{aligned}$$

Let  $(G_1, G_2)$  denote the left-hand side of (4.12). In order that the structure of zeros of  $(G_1, G_2)$  should be qualitatively unaffected by the addition of  $(h_1, h_2)$ , we need only observe that, under our assumptions, if  $(\lambda, \mu, \alpha, \beta)$  is a solution of (4.12), then (i)  $\beta \neq 0$  implies  $|\partial(G_1, G_2)/\partial(\alpha, \beta)| \neq 0$ , whereas (ii)  $\beta = 0$  and  $(\lambda, \mu, \alpha) \neq (0, 0, 0)$  together

imply  $|\partial(G_1, G_2)/\partial(\lambda, \alpha)| \neq 0$ . This procedure is now familiar in bifurcation theory, although the full details (omitted here) are often rather technical [13], [16], [18], [19]. The solutions of (4.12) are conveniently written in the form

$$(4.13) \quad \begin{aligned} \alpha &= -(a_2\lambda + b_2\mu^k)/B, \\ \beta^2 &= (B^2C)^{-1}(a_2\lambda + b_2\mu^k)[(a_1B - a_2A)\lambda + (b_1B - b_2A)\mu^k]. \end{aligned}$$

We can now describe bifurcation diagrams for (1.1), under assumption (S1), by drawing the corresponding diagrams for the bifurcation equations (4.10), (4.11). One bifurcation diagram for each of  $\mu < 0, \mu = 0, \mu > 0$  describes the structure of solutions. We distinguish two cases as follows. For fixed  $\mu$ , let  $y_\mu(\lambda)$  be the quadratic

$$y_\mu(\lambda) = C(a_2\lambda + b_2\mu^k)((a_1B - a_2A)\lambda + (b_1B - b_2A)\mu^k).$$

*Case QI:*  $a_2C(a_1B - a_2A) > 0$ . For each  $\mu$ ,  $y_\mu(\lambda)$  has a minimum. Consequently, the primary branch of solutions  $\mathcal{C}'_\mu$  (corresponding to  $\alpha = \beta = 0$  in (4.13)), and the secondary branch  $\mathcal{D}_\mu$  (corresponding to  $\beta = 0, \alpha \neq 0$ ) have no point of intersection if  $\mu \neq 0$ . As  $\mu \rightarrow 0$ , the secondary branch becomes a primary branch through  $(0, \theta)$ , so that there are exactly three primary branches of solutions of (1.1) through  $(0, \theta)$  when  $\mu = 0$ . An example is represented in Fig. 1.

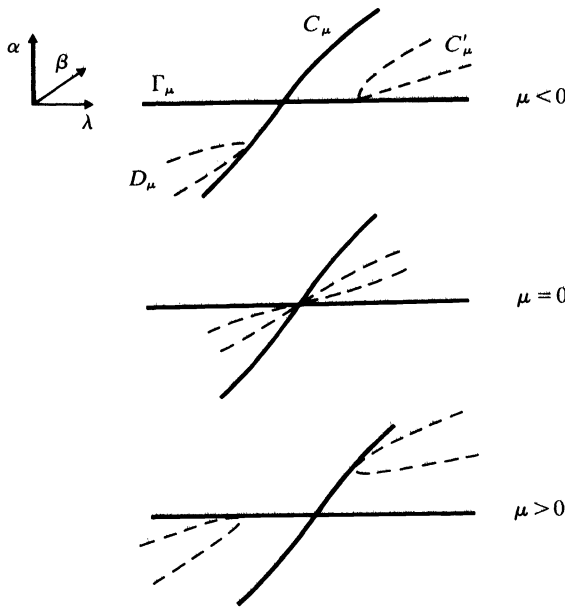


FIG. 1. QI:  $a_2C(a_1B - a_2A) > 0$ . Example illustrated:  $k = 1; 0 > B/a_2 > A/a_1; b_2/a_2 > b_1/a_1$ .

*Case QII:*  $a_2C(a_1B - a_2A) < 0$ . For each  $\mu$ ,  $y_\mu(\lambda)$  has a maximum. Consequently, for  $\mu \neq 0$ ,  $\mathcal{C}'_\mu$  and  $\mathcal{D}_\mu$  are coincident. As  $\mu \rightarrow 0$ , these branches collapse into the point  $(0, \theta)$ , so there is just one primary branch  $\mathcal{C}_0$  through  $(0, \theta)$  when  $\mu = 0$ . An example of Case QII is represented in Fig. 2.

*Remark.* Our case QI, with  $k = 1$  corresponds to cases I-III for quadratic nonlinearities in Keener's formal study of a pair of one-dimensional reaction diffusion equations [11]. Case QII ( $k = 1$ ) corresponds to Keener's cases IV-VI.

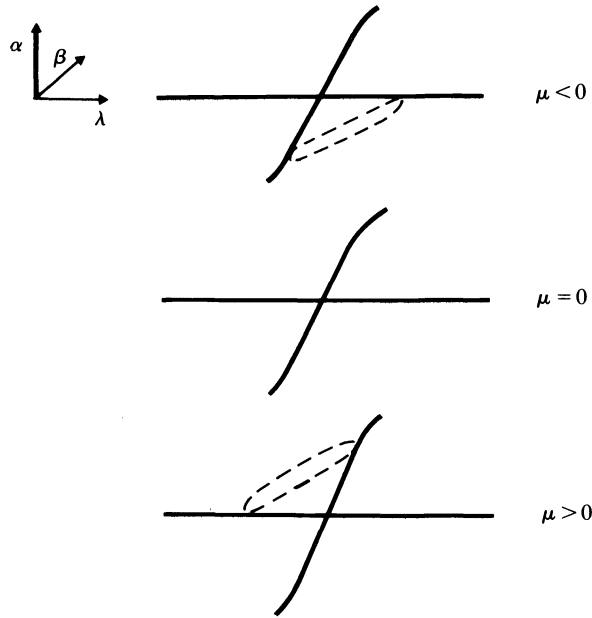


FIG. 2. QII:  $a_2C(a_1B - a_2A) < 0$ . Example illustrated: As for Fig. 1.

**Bifurcation analysis under (S2) or (J1)–(J3).** Suppose  $F$  is of class  $C^n$  with  $n \geq 4$ , and satisfies (H1), (H2), (S2) or (J1)–(J3), and (Dk),  $k = 1$  or 2. The bifurcation equations (4.8) satisfy (4.9) and are of the form

$$(4.14) \quad (a_1\lambda + b_1\mu^k)\alpha + p\alpha^3 + r\alpha\beta^2 + \alpha\tilde{h}_1(\lambda, \mu, \alpha, \beta) = 0,$$

$$(4.15) \quad (a_2\lambda + b_2\mu^k)\beta + q\alpha^2\beta + s\beta^3 + \beta\tilde{h}_2(\lambda, \mu, \alpha, \beta) = 0,$$

where  $p = \langle C\phi_1^3, \psi_1' \rangle$ ,  $q = \langle 3C\phi_1^2\phi_2, \psi_2' \rangle$ ,  $r = \langle 3C\phi_1\phi_2^2, \psi_1' \rangle$ ,  $s = \langle C\phi_2^3, \psi_2' \rangle$ ; the mapping  $(\tilde{h}_1, \tilde{h}_2): \mathcal{A}(\delta) \rightarrow \mathbb{R}^2$  is of class  $C^{n-1}$ , and represents higher order terms in  $(\lambda, \mu, \alpha, \beta)$ ,

$$|\tilde{h}_i(\lambda, \mu, \alpha, \beta)| \leq \text{const.} \{(|\alpha, \beta|)^3 + (|\lambda| + |\mu|^k)(|\alpha, \beta|) + \lambda^2 + |\mu|^{k+1} + |\mu|(|\alpha, \beta|)^2\}.$$

Note that if (S2) holds, then  $(\tilde{h}_1, \tilde{h}_2)(\lambda, \mu, \alpha, \beta)$  is even in both  $\alpha$  and  $\beta$  for all  $\lambda$  and  $\mu$ , whereas under (J1)–(J3),  $(\tilde{h}_1, \tilde{h}_2)(\lambda, \mu, \alpha, \beta)$  is even in  $(\alpha, \beta)$  for all  $\lambda$  and  $\mu$ .

In terms of (4.14), (4.15), we assume

$$(H4) \quad a_1a_2 \neq 0,$$

$$(Dk) \quad a_1b_2 \neq a_2b_1,$$

$$(E2) \quad a_1q \neq a_2p$$

and the additional conditions

$$(E2') \quad a_1s \neq a_2r \quad \text{and} \quad ps \neq rq.$$

Let  $\lambda = \bar{\lambda}(\mu, \alpha)$  be the solution of (4.14) with  $\beta = 0$ , and let  $\lambda = \hat{\lambda}(\mu, \beta)$  be the solution of (4.15) with  $\alpha = 0$ . Both  $\bar{\lambda}$  and  $\hat{\lambda}$  are guaranteed by Theorem 2.1, and correspond respectively to the branches  $\mathcal{C}_\mu, \mathcal{C}'_\mu$  of solutions of (1.1). A simple cal-

ulation shows that

$$(4.16) \quad \bar{\lambda}_\alpha(\mu, 0) = 0 = \hat{\lambda}_\beta(\mu, 0) \quad \text{for all } \mu$$

and  $\bar{\lambda}_{\alpha\alpha}(0, 0) = -2p/a_1$ ;  $\hat{\lambda}_{\beta\beta}(0, 0) = -2s/a_2$ , so that

$$(4.17) \quad \begin{aligned} \operatorname{sgn} \bar{\lambda}_{\alpha\alpha}(\mu, 0) &= -\operatorname{sgn} pa_1, \\ \operatorname{sgn} \hat{\lambda}_{\beta\beta}(\mu, 0) &= -\operatorname{sgn} sa_2. \end{aligned}$$

for all  $\mu$  near zero, provided  $p$  and  $s$  are nonzero. The formulae (4.16), (4.17) determine the direction of bifurcation of the primary solution branches  $\mathcal{C}_\mu$  and  $\mathcal{C}'_\mu$ .

Dividing (4.14) by  $\alpha$ , (4.15) by  $\beta$ , and disregarding the higher order terms  $(\tilde{h}_1, \tilde{h}_2)$ ,

$$(4.18) \quad \begin{aligned} a_1\lambda + b_1\mu^k + p\alpha^2 + r\beta^2 &= 0, \\ a_2\lambda + b_2\mu^k + q\alpha^2 + s\beta^2 &= 0. \end{aligned}$$

Let  $(G_1, G_2)$  denote the left-hand side of (4.18), and suppose  $(\lambda, \mu, \alpha, \beta) \neq (0, 0, 0, 0)$  is a solution of (4.18). Then  $(\alpha, \beta) \neq (0, 0)$  and

- (i)  $\alpha\beta \neq 0$  implies  $\left| \frac{\partial(G_1, G_2)}{\partial(\alpha, \beta)} \right| = 2(ps - rq)\alpha\beta \neq 0,$
- (ii)  $\alpha \neq 0, \quad \beta = 0$  implies  $\left| \frac{\partial(G_1, G_2)}{\partial(\alpha, \lambda)} \right| = 2(a_2p - a_1q)\alpha \neq 0,$
- (iii)  $\alpha = 0, \quad \beta \neq 0$  implies  $\left| \frac{\partial(G_1, G_2)}{\partial(\beta, \lambda)} \right| = 2(a_2r - a_1s)\beta \neq 0.$

Consequently, the addition of  $(\tilde{h}_1, \tilde{h}_2)$  to  $(G_1, G_2)$  does not affect the structure of the zeros of  $(G_1, G_2)$  near  $(0, 0, 0, 0)$ . Corresponding solutions of (1.1) are obtained from Theorem 2.3.

Two distinct types of bifurcation diagram arise, depending on the coefficients in (4.18).

Case CI:  $(a_1s - a_2r)(a_1q - a_2p) < 0$ . For fixed  $\mu$ , the lines

$$\begin{aligned} t_1(\lambda) &= -(ps - rq)\{(a_1s - a_2r)\lambda + (b_1s - b_2r)\mu^k\}, \\ t_2(\lambda) &= (ps - rq)\{(a_1q - a_2p)\lambda + (b_1q - b_2p)\mu^k\} \end{aligned}$$

have gradients with the same sign.

Consequently, there are exactly two secondary bifurcation points for each  $\mu \neq 0$ . Both secondary branches of solutions of (1.1) meet

$$\begin{cases} \mathcal{C}_\mu & \text{if } (a_1b_2 - a_2b_1)(a_1q - a_2p)\mu^k < 0, \\ \mathcal{C}'_\mu & \text{if } (a_1b_2 - a_2b_1)(a_1q - a_2p)\mu^k > 0 \end{cases}$$

only at the secondary bifurcation points. When  $\mu = 0$ , there are exactly four primary branches of solutions of (1.1) passing through  $(0, \theta)$ .

Case CII:  $(a_1s - a_2r)(a_1q - a_2p) > 0$ . For each  $\mu$ ,  $t_1(\lambda)$  and  $t_2(\lambda)$  have gradients with opposite signs.

If  $(a_1b_2 - a_2b_1)(a_1q - a_2p)\mu^k < 0$ , there are precisely two secondary bifurcation points on each of  $\mathcal{C}_\mu, \mathcal{C}'_\mu$ . The corresponding secondary branches of solutions of (1.1) connect all these secondary bifurcation points in a loop in  $\mathbb{R} \times X$ . As  $\mu \rightarrow 0$ , this loop collapses onto the point  $(0, \theta)$ , so that there are just two primary branches of solutions of (1.1) through  $(0, \theta)$  when  $\mu = 0$ .

If  $(a_1b_2 - a_2b_1)(a_1q - a_2p)\mu^k > 0$ , then there are no secondary bifurcation points. More precisely, there is a neighborhood  $V \subset \mathbb{R}^2 \times X$  of  $(0, 0, \theta)$  for which the set  $\{(\lambda, \mu, x) \in V : (a_1b_2 - a_2b_1)(a_1q - a_2p)\mu^k \cong 0\}$  contains no secondary bifurcation points for (1.1).

Clearly, the cases  $k = 1$  and  $k = 2$  are different.

CII ( $k = 1$ ). There are four (respectively, zero) secondary bifurcation points if  $(a_1b_2 - a_2b_1)(a_1q - a_2p)\mu$  is negative (respectively, positive).

CII ( $k = 2$ ). There are four (respectively, zero) secondary bifurcation points for each  $\mu \neq 0$  if  $(a_1b_2 - a_2b_1)(a_1q - a_2p)$  is negative (respectively, positive).

Examples of cases CI, CII are represented in Figs. 3-6.

**5. An example.** For  $a > 0$  let  $\Omega_a = (0, a) \times (0, \pi)$ , and consider the following nonlinear boundary value problem

$$(5.1) \quad \begin{aligned} u_{xx}(x, y) + u_{yy}(x, y) + \lambda g(u(x, y)) &= 0 & (x, y) \in \Omega_a, \\ u(x, y) &= 0 & (x, y) \in \partial\Omega_a, \end{aligned}$$

where  $\lambda \in \mathbb{R}$ ;  $g \in C^4(I, \mathbb{R})$ , for some open interval  $I$ ,  $0 \in I$ ,  $g(0) = 0$ ,  $g'(0) = 1$ .

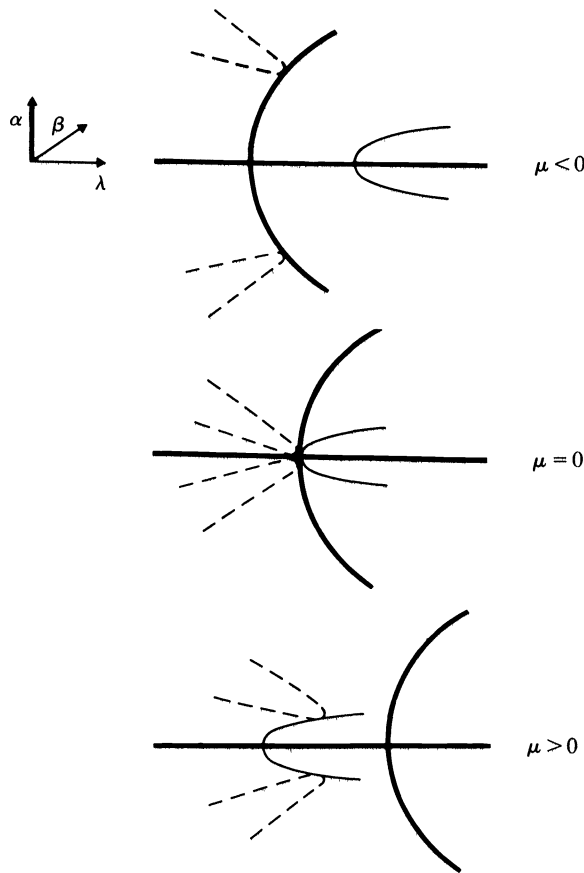


FIG. 3. CI:  $(a_1s - a_2r)(a_1q - a_2p) < 0$ . Example illustrated:  $k = 1$ ;  $b_2/a_2 > b_1/a_1$ ;  $q/a_2 > 0 > p/a_1$ ;  $(a_1b_2 - a_2b_1)(ps - rq) < 0$ ;  $r/a_1 > 0 > s/a_2$ .

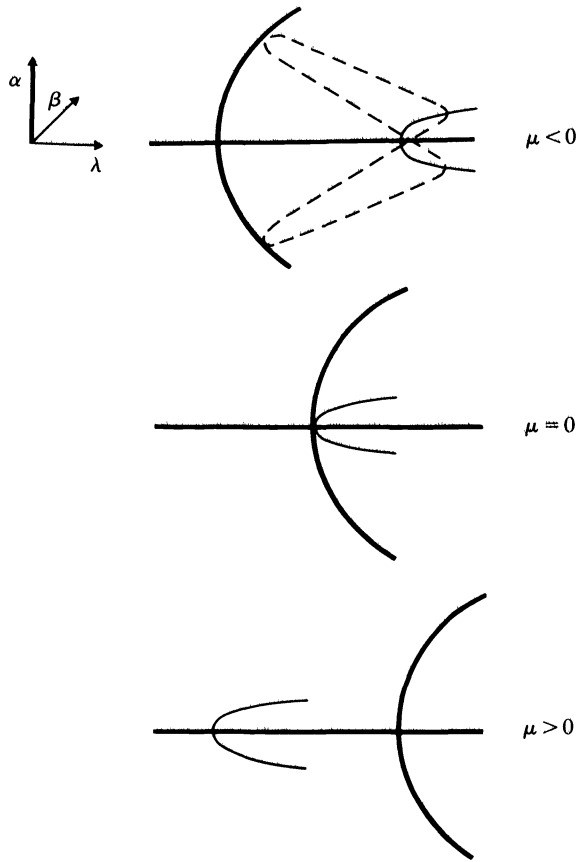


FIG. 4. CII:  $(a_1s - a_2r)(a_1q - a_2p) > 0$ , and  $k = 1$ . Example illustrated:  $b_2/a_2 > b_1/a_1$ ;  $q/a_2 > 0 > p/a_1$ ;  $a_2s < 0$ .

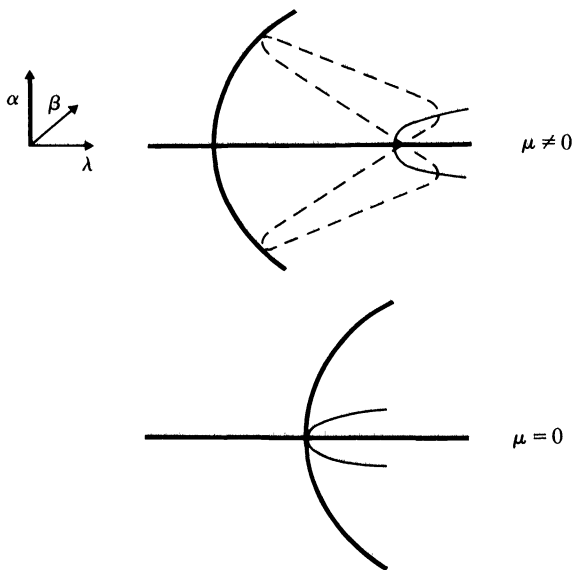


FIG. 5. CII:  $(a_1s - a_2r)(a_1q - a_2p) > 0$ ,  $k = 2$ , and  $(a_1b_2 - a_2b_1)(a_1q - a_2p) < 0$ . Example illustrated:  $b_1/a_1 > b_2/a_2$ ;  $q/a_2 > 0 > p/a_1$ ;  $a_2s < 0$ .

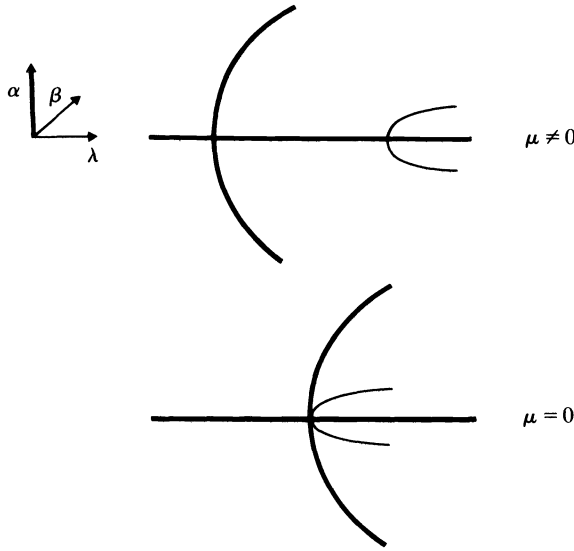


FIG. 6. CII:  $(a_1s - a_2r)(a_1q - a_2p) > 0$ ,  $k = 2$ , and  $(a_1b_2 - a_2b_1)(a_1q - a_2p) > 0$ . Example illustrated:  $b_1/a_1 > b_2/a_2$ ;  $0 > p/a_1 > q/a_2$ ;  $a_2s < 0$ .

Equation (5.1) has been studied by Kriegsmann and Reiss [12], in connection with the equations of magnetohydrodynamics, using formal perturbation methods, and by Budden and Norbury [4], using both formal and numerical methods. Under the condition  $g''(0) = 0$ ,  $g'''(0) \neq 0$ , these authors observe secondary bifurcation for values of  $a$  near  $\pi$ , where  $\lambda$  is the bifurcation parameter, and  $u(x, y) \equiv 0$  is the trivial solution.

To set (5.1) in the form of (1.1), transform  $x \rightarrow \pi x/a$ , and let  $D = (0, \pi)^2$  be the transformed domain. Define the linear operator  $\Delta_a$  from  $W^{2,2}(D) \cap \dot{W}^{1,2}(D)$  to  $L^2(D)$  by

$$(\Delta_a u)(x, y) = \frac{a^2}{\pi^2} \frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y),$$

where the derivatives are generalized ( $L^2$ ) derivatives. We write  $\Delta = \Delta_\pi$ . Let  $X = W^{2,2}(D) \cap \dot{W}^{1,2}(D)$  be the Banach space with graph norm of  $\Delta$ ,

$$\|u\|_X = \|u\|_{L^2(D)} + \|\Delta u\|_{L^2(D)}.$$

In fact,  $X$  is a Hilbert space with inner product

$$(u, v)_X = \langle u, v \rangle_{L^2(D)} + \langle \Delta u, \Delta v \rangle_{L^2(D)}.$$

Set  $Y = L^2(D)$  with inner product  $\langle \cdot, \cdot \rangle$ . The estimate [1]

$$\|u\|_{W^{2,2}(D)} \leq \text{const.} \|u\|_X, \quad u \in X$$

implies that  $\Delta_a: X \rightarrow Y$  is bounded for all  $a \neq 0$ . Moreover, the graph norms of  $\Delta_a$  for different  $a \neq 0$  are all equivalent to the  $W^{2,2}$  norm of  $W^{2,2}(D) \cap \dot{W}^{1,2}(D)$ .

Define  $\tilde{g}: X \rightarrow Y$  by

$$\tilde{g}(u)(x, y) = g(u(x, y)), \quad (x, y) \in D, \quad u \in X.$$

Since  $W^{2,2}(D)$  is embedded in  $C(D)$ ,  $\tilde{g}$  is of class  $C^4$  in a neighborhood the zero  $\theta$ , of  $X$ .



Now let  $F: \mathbb{R}^2 \times X \rightarrow Y$  be defined by

$$F(\lambda, a, u) = \Delta_a u + \lambda \tilde{g}(u).$$

Then for each  $a \neq 0$ ,  $F$  is of class  $C^4$ ,

$$F(\lambda, a, \theta) = 0 \quad \text{identically}$$

and  $F_u(\lambda, a, \theta): X \rightarrow Y$  is a Fredholm operator with Fredholm index zero

$$F_u(\lambda, a, \theta)u = (\Delta_a + \lambda)u, \quad u \in X, \quad a \neq 0.$$

The eigenvalues of  $\Delta_a$  are given by

$$\lambda_{m,n}(a) = \frac{m^2 a^2}{\pi^2} + n^2, \quad m, n = 1, 2, \dots$$

with corresponding eigenfunctions (normalized in  $Y$ ):

$$\phi_{m,n} = \frac{2}{\pi} \sin(mx) \sin(ny).$$

When  $a = \pi$ ,  $\lambda_{2,1} = \lambda_{1,2} = 5$ , and  $\lambda_{3,1} = \lambda_{1,3} = 10$  are double eigenvalues. We shall be concerned with bifurcation near  $u = \theta$ , when  $a$  is near  $\pi$ , and the two cases  $\lambda$  near 5 and  $\lambda$  near 10. These cases give rise to significantly different symmetry properties of  $F$  which affect the bifurcation analysis.

We first consider solutions  $(\lambda, a, u) \in \mathbb{R}^2 \times X$  near  $(5, \pi, \theta)$  of the equation

$$(5.2) \quad F(\lambda, a, u) = 0.$$

Define orthogonal subsets  $U_{\rho\sigma}$  ( $\rho, \sigma = \pm$ ) of  $L^2(D)$  by

$$U_{++} = \{\phi_{2m+1,2n+1} : m, n = 0, 1, 2, \dots\},$$

$$U_{+-} = \{\phi_{2m+1,2n} : m, n = 0, 1, 2, \dots\},$$

$$U_{-+} = \{\phi_{2m,2n+1} : m, n = 0, 1, 2, \dots\},$$

$$U_{--} = \{\phi_{2m,2n} : m, n = 1, 2, \dots\},$$

and set

$$X_{\rho\sigma} = Cl_X \text{ span } U_{\rho\sigma},$$

$$Y_{\rho\sigma} = Cl_Y \text{ span } U_{\rho\sigma},$$

( $\rho, \sigma = \pm$ ), where  $Cl_X, Cl_Y$  denotes the closure in  $X, Y$  respectively. Then

$$X = \bigoplus_{\rho, \sigma = \pm} X_{\rho\sigma}, \quad Y = \bigoplus_{\rho, \sigma = \pm} Y_{\rho\sigma}.$$

Next define linear operators  $T_1, T_2$  from  $Y$  to  $Y$  by

$$T_1 y = \rho y, \quad y \in Y_{\rho+} \oplus Y_{\rho-} \quad (\rho = \pm),$$

$$T_2 y = \sigma y, \quad y \in Y_{+\sigma} \oplus Y_{-\sigma} \quad (\sigma = \pm)$$

and let  $S_1, S_2$  be the restrictions to  $X$  of  $T_1, T_2$  respectively. The symmetry hypothesis (S2) of § 4 takes the form

$$F(\lambda, a, S_i u) = T_i F(\lambda, a, u), \quad (\lambda, a, u) \in \mathbb{R}^2 \times X \quad (i = 1, 2),$$

$$\phi_{1,2} \in X_{+-} \subset Y_{+-}, \quad \phi_{2,1} \in X_{-+} \subset Y_{-+},$$

$$\mathcal{R} \oplus \text{span} \{\phi_{1,2}, \phi_{2,1}\} = Y.$$

(Here,  $\mathcal{R} = \mathcal{R}(F_u(5, \pi, \theta))$ , and we write  $\mathcal{N} = \mathcal{N}(F_u(5, \pi, \theta))$ .) Hypothesis (H4) is satisfied,

$$F_{u\lambda}(5, \pi, \theta)\phi_{1,2} = \phi_{1,2} \notin \mathcal{R},$$

$$F_{u\lambda}(5, \pi, \theta)\phi_{2,1} = \phi_{2,1} \notin \mathcal{R}.$$

Therefore, for each  $a$  near  $\pi$ , we have two primary branches  $\mathcal{C}_a \subset \mathbb{R} \times (X_{++} \oplus X_{+-})$ ,  $\mathcal{C}'_a \subset \mathbb{R} \times (X_{++} \oplus X_{--})$  of solutions of (5.2).  $\mathcal{C}_a$  branches from  $\Gamma_a = \{(\lambda, \theta) : \lambda \in \mathbb{R}\}$  at  $\lambda_{1,2} = 1 + 4a^2/\pi^2$ , whereas  $\mathcal{C}'_a$  branches from  $\Gamma_a$  at  $\lambda_{2,1} = 4 + a^2/\pi^2$ .

The bifurcation equations must be of the form (4.14), (4.15), where  $\mu = a - \pi$ , and  $\lambda$  is transformed to  $\lambda - 5$ . Clearly,  $k = 1$  and  $a_1 = a_2; b_1 = -2/\pi, b_2 = -8/\pi^2$ . It remains to calculate the coefficients  $p, q, r, s$ .

Let  $B: X \times X \rightarrow Y$  be the symmetric bilinear operator given by  $B(u, v) = uv$ . Then  $B$  maps  $X_{\nu\mu} \times X_{\xi\eta}$  to  $Y_{\rho\sigma}$  whenever  $\nu\xi\rho = +$  and  $\mu\eta\sigma = +$ . In particular,  $\phi_{1,2}^2 \in Y_{++}$ ,  $\phi_{2,1}^2 \in Y_{++}$ , and  $\phi_{1,2} \cdot \phi_{2,1} \in Y_{--}$ . Let  $Z_+$  (respectively  $Z_-$ ) be the orthogonal complement of  $\text{span}\{\phi_{2,1}\}$  ( $\text{span}\{\phi_{1,2}\}$ ) in  $X_{+-}$  ( $X_{--}$ ). Setting  $Z = Z_+ \oplus Z_- \oplus X_{++} \oplus X_{--}$ , the symmetric bilinear mapping  $z_1: \mathcal{N} \times \mathcal{N} \rightarrow Z$ , defined in general by (3.10), is here given by

$$(5.3) \quad \Delta z_1(w, v) + 5z_1(w, v) + 5g''(0)wv = 0 \quad (w, v \in \mathcal{N}).$$

Then  $z_1(\phi_{1,2}, \phi_{1,2}) \in X_{++}$ ,  $z_1(\phi_{2,1}, \phi_{2,1}) \in X_{++}$  and  $z_1(\phi_{1,2}, \phi_{2,1}) \in X_{--}$ , so that

$$\begin{aligned} \phi_{1,2}z_1(\phi_{1,2}, \phi_{1,2}) &\in Y_{+-}, & \phi_{1,2}z_1(\phi_{2,1}, \phi_{2,1}) &\in Y_{+-}, \\ \phi_{2,1}z_1(\phi_{1,2}, \phi_{1,2}) &\in Y_{--}, & \phi_{2,1}z_1(\phi_{2,1}, \phi_{2,1}) &\in Y_{--}, \\ \phi_{1,2}z_1(\phi_{1,2}, \phi_{2,1}) &\in Y_{--}, & \phi_{2,1}z_1(\phi_{1,2}, \phi_{2,1}) &\in Y_{+-}. \end{aligned}$$

We can now write down the bifurcation equations

$$\begin{aligned} (\lambda - 4 - (a/\pi)^2)\alpha + p\alpha^3 + r\alpha\beta^2 + \alpha g_1(\lambda, a, \alpha, \beta) &= 0, \\ (\lambda - 1 - (2a/\pi)^2)\beta + q\alpha^2\beta + s\beta^3 + \beta g_2(\lambda, a, \alpha, \beta) &= 0, \end{aligned}$$

where

$$\begin{aligned} p &= \frac{5}{6}g'''(0)\langle \phi_{1,2}^3, \phi_{1,2} \rangle + \frac{5}{2}g''(0)\langle \phi_{1,2}z_1(\phi_{1,2}, \phi_{1,2}), \phi_{1,2} \rangle, \\ q &= \frac{5}{2}g'''(0)\langle \phi_{1,2}^2\phi_{2,1}, \phi_{2,1} \rangle \\ &\quad + \frac{5}{2}g''(0)\langle \phi_{2,1}z_1(\phi_{1,2}, \phi_{1,2}) + 2\phi_{1,2}z_1(\phi_{1,2}, \phi_{2,1}), \phi_{2,1} \rangle, \\ r &= \frac{5}{2}g'''(0)\langle \phi_{1,2}\phi_{2,1}^2, \phi_{1,2} \rangle \\ &\quad + \frac{5}{2}g''(0)\langle \phi_{1,2}z_1(\phi_{2,1}, \phi_{2,1}) + 2\phi_{2,1}z_1(\phi_{1,2}, \phi_{2,1}), \phi_{1,2} \rangle, \\ s &= \frac{5}{6}g'''(0)\langle \phi_{2,1}^3, \phi_{2,1} \rangle + \frac{5}{2}g''(0)\langle \phi_{2,1}z_1(\phi_{2,1}, \phi_{2,1}), \phi_{2,1} \rangle \end{aligned}$$

and  $(g_1, g_2): \mathbb{R}^4 \rightarrow \mathbb{R}^2$  is defined, and of class  $C^3$ , near  $(5, \pi, 0, 0)$ , and satisfies

$$|g_i(\lambda, a, \alpha, \beta)| \leq K|(\alpha, \beta)|^2\{ |(\alpha, \beta)| + |\lambda - 5| + |a - \pi| \} \quad (i = 1, 2),$$

where  $K > 0$  is independent of  $(\lambda, a, \alpha, \beta)$ .

Expressing  $z_1(w, v)(w, v \in \mathcal{N})$  as Fourier series, using (5.3) one obtains

$$\begin{aligned} p = s &\approx 0.1900g'''(0) - 0.3638(g''(0))^2, \\ q = r &\approx 0.2533g'''(0) + 27.4104(g''(0))^2. \end{aligned}$$

In order to appeal to the bifurcation analysis under (S2), of § 4, we have only to ensure

that  $g''(0), g'''(0)$  satisfy

$$(5.4) \quad |p| \neq |q|.$$

Clearly, (5.4) holds if either  $g'''(0) > 0$ , or if  $g''(0) = 0$  and  $g'''(0) \neq 0$ . If (5.4) holds, then

$$(a_1s - a_2r)(a_1q - a_2p) = -(p - q)^2 < 0$$

so that the bifurcation diagrams of case CI ( $k = 1$ ) apply.

We now turn to considering solutions of (5.2) near  $(10, \pi, \theta)$ , under the assumption that

$$(5.5) \quad g(-u) = -g(u), \quad u \in \mathbb{R}.$$

Define  $\mathcal{N} = \mathcal{N}(F_u(10, \pi, \theta)) = \text{span} \{\phi_{1,3}, \phi_{3,1}\}$ ,  $\mathcal{R} = \mathcal{R}(F_u(10, \pi, \theta))$ .

For  $j, k = 1, 2, 3$ , set

$$U_{j,k} = \text{span} \{\phi_{3m+j, 3n+k} : m, n = 0, 1, 2, \dots\}$$

and

$$X_{j,k} = Cl_X U_{j,k}, \quad Y_{j,k} = Cl_Y U_{j,k}.$$

Then  $\phi_{1,3} \in X_{1,3} \subset Y_{1,3}$ ,  $\phi_{3,1} \in X_{3,1} \subset Y_{3,1}$ .

LEMMA 5.1.  $\tilde{g}$  maps

$$\begin{aligned} \bigoplus_{k=1}^3 X_{3,k} & \text{ to } \bigoplus_{k=1}^3 Y_{3,k}, \\ \bigoplus_{j=1}^3 X_{j,3} & \text{ to } \bigoplus_{j=1}^3 Y_{j,3} \end{aligned}$$

and  $X_{3,3}$  to  $Y_{3,3}$ .

*Proof.* Since  $\tilde{g}$  is continuous and odd, and  $\bar{D} \subset \mathbb{R}^2$  is compact, we can approximate  $\tilde{g}: X \rightarrow Y$  by odd polynomials  $\tilde{g}_n: X \rightarrow Y$ ,  $\tilde{g}_n(u) \rightarrow \tilde{g}(u)$  as  $n \rightarrow \infty$  for each  $u \in X$ .

Set  $U_1 = \bigoplus_{k=1}^3 U_{3k}$ ,  $X_1 = Cl_X U_1$ ,  $Y_1 = Cl_Y U_1$ , and let  $u \in U_1$ . Then  $u^n \in Y_1$  for each odd integer  $n \geq 1$ , since  $\langle u^n, \phi_{p,q} \rangle$  contains only terms of the form

$$\int_0^\pi \int_0^\pi \sin(3m_1x) \cdots \sin(3m_nx) \sin(px) \sin(k_1y) \cdots \sin(k_ny) \sin(qy) \, dx \, dy,$$

each of which is zero unless  $p$  is a multiple of 3. Therefore, if  $f_n$  is an odd polynomial,  $f_n(u) \in Y_1$  whenever  $u \in U_1$ , since  $Y_1$  is linear.

Let  $u \in X_1$ , and let  $\{u_N\} \subset U_1$  be a sequence such that  $u_N \rightarrow u$  in  $X$  as  $N \rightarrow \infty$ . Then  $\tilde{g}(u_N) \rightarrow \tilde{g}(u)$  (since  $\tilde{g}$  is continuous), so let  $\{\tilde{g}_n\}$  be a sequence of polynomials such that  $\tilde{g}_n(w) \rightarrow \tilde{g}(w)$  as  $n \rightarrow \infty$  for all  $w \in X$ . Then  $\tilde{g}_n(u_N) \rightarrow \tilde{g}(u_N) \in Y_1$  as  $n \rightarrow \infty$  and  $\tilde{g}(u_N) \rightarrow \tilde{g}(u)$  in  $Y_1$  as  $N \rightarrow \infty$ . Therefore,  $\tilde{g}(u) \in Y_1$ . This proves the first statement of the lemma. The proof of the second statement is identical to that of the first. To show that  $\tilde{g}$  maps  $X_{3,3}$  to  $Y_{3,3}$ , it is sufficient to note that

$$u^n \in Y_{3,3} \quad \text{if } u \in U_{3,3} \text{ and } n \text{ is odd.}$$

The proof is then identical to that above, except that  $u \in X_{3,3}$  should be approximated by  $u_n \in U_{3,3}$ .

Now set

$$\begin{aligned} X^{(0)} &= X_{3,3}, & Y^{(0)} &= Y_{3,3}, \\ X^{(1)} &= \bigoplus_{k=1}^2 X_{3,k}, & Y^{(1)} &= \bigoplus_{k=1}^2 Y_{3,k}, \\ X^{(2)} &= \bigoplus_{j=1}^2 X_{j,3}, & Y^{(2)} &= \bigoplus_{j=1}^2 Y_{j,3}, \\ \tilde{X} &= \bigoplus_{j,k=1}^2 X_{j,k}, & \tilde{Y} &= \bigoplus_{j,k=1}^2 Y_{j,k}. \end{aligned}$$

Then  $X = X^{(0)} \oplus X^{(1)} \oplus X^{(2)} \oplus \tilde{X}$ ;  $Y = Y^{(0)} \oplus Y^{(1)} \oplus Y^{(2)} \oplus \tilde{Y}$ , and  $F(\mathbb{R}^2 \times (X^{(i)} \oplus X^{(0)})) \subset Y^{(i)} \oplus Y^{(0)}$ ,  $i = 1, 2$ . Moreover,  $\phi_{3,1} \in X^{(1)}$ ,  $\phi_{1,3} \in X^{(2)}$  and

$$\text{span} \{ \phi_{3,1}, \phi_{1,3} \} \oplus \mathcal{R} = Y.$$

These conditions provide a slight generalization of condition (J1), due to the inclusion of the closed subspaces  $\tilde{X}$ ,  $\tilde{Y}$  of  $X$ ,  $Y$  respectively, which play essentially no role in the symmetry property of  $F$  (for the purposes of a bifurcation analysis). It is easy to check directly that

$$\langle \phi_{1,3}^2 \phi_{3,1}, \phi_{1,3} \rangle = 0 = \langle \phi_{3,1}^2 \phi_{1,3}, \phi_{3,1} \rangle$$

so that (J3) is satisfied. We have assumed explicitly that (J2) holds. Consequently, the bifurcation equations have the form (4.14)–(4.15),

$$\begin{aligned} (\lambda - 9 - (a/\pi)^2)\alpha + p\alpha^3 + r\alpha\beta^2 + \alpha h_1(\lambda, a, \alpha, \beta) &= 0, \\ (\lambda - 1 - (3a/\pi)^2)\beta + q\alpha^2\beta + s\beta^3 + \beta h_2(\lambda, a, \alpha, \beta) &= 0. \end{aligned}$$

So,  $a_1 = 1 = a_2$ ,  $b_1 = -2/\pi$ ,  $b_2 = -18/\pi$ ,

$$p = 9(5g'''(0)/4\pi^2) = s, \quad q = 12(5g'''(0)/4\pi^2) = r.$$

In particular, under the assumption  $g'''(0) \neq 0$ , we have

$$|p| \neq |q|$$

which implies that all the conditions of the bifurcation analysis under (J1)–(J3) are satisfied. In fact, case CI ( $k = 1$ ) applies, since

$$(a_1s - a_2r)(a_1q - a_2p) = -(p - q)^2 < 0.$$

**Acknowledgment.** The author is grateful to the Science Research Council for their financial support.

REFERENCES

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.  
 [2] L. BAUER, H. B. KELLER AND E. L. REISS, *Multiple eigenvalues lead to secondary bifurcation*. SIAM Rev., 17 (1975), pp. 101–122.  
 [3] M. BUCHNER, J. MARSDEN AND S. SCHECHTER, *Differential topology and singularity theory in the solution of nonlinear equations*, preprint.  
 [4] P. J. BUDDEN AND J. NORBURY, *A non-linear elliptic eigenvalue problem*, preprint.  
 [5] A. H. CHILVER, *Coupled modes of elastic buckling*, J. Mech. Phys. Solids, 15 (1967), pp. 15–28.

- [6] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.
- [7] M. A. GOLUBITSKY AND D. G. SCHAEFFER, *Imperfect bifurcation in the presence of symmetry*, preprint.
- [8] ———, *Bifurcation analysis near a double eigenvalue of a model biochemical reaction*, MRC Rep. 1859, Univ. of Wisconsin, Madison, WI, 1978.
- [9] ———, *Boundary conditions and mode jumping in the buckling of a rectangular plate*, preprint.
- [10] P. HALL AND I. C. WALTON, *Benard convection in a finite box: Secondary and imperfect bifurcations*, preprint.
- [11] J. P. KEENER, *Secondary bifurcation in nonlinear diffusion reaction equations*, Studies in Appl. Math., 55 (1976), pp. 187–211.
- [12] G. A. KRIEGSMANN AND E. L. REISS, *New MHD equilibria by secondary bifurcation*, Phys. of Fluids, to appear.
- [13] S. E. LIST, *Generic bifurcation with application to the von Karman equations*, Ph.D. thesis, Brown Univ., Providence, RI, 1976.
- [14] R. J. MAGNUS, *On the local structure of the zero set of a Banach space valued mapping*, J. Functional Analysis, 22 (1976), pp. 57–73.
- [15] T. J. MAHAR AND B. J. MATKOWSKY, *A model biochemical reaction exhibiting secondary bifurcation*, SIAM J. Appl. math., 32 (1977), pp. 394–404.
- [16] J. MALLET-PARET, *Buckling of cylindrical shells with small curvature*, Quart. Appl. Math., 35 (1977), pp. 383–400.
- [17] D. H. SATTINGER, *Group representation theory and branch points of nonlinear functional equations*, this Journal, 8 (1977), pp. 179–201.
- [18] M. SHEARER, *Bifurcation of axisymmetric buckled states of a thin spherical shell*, Rep. 92, Fluid Mech. Res. Inst., Univ. of Essex, 1978.
- [19] ———, *Secondary bifurcation for one-parameter families of bifurcation problems*, Ibid., Rep. 97.
- [20] J. M. T. THOMPSON AND G. W. HUNT, *A General Theory of Elastic Stability*, Wiley, London, 1973.

## A SHORT NOTE ON A DIFFERENTIAL RECURSION FORMULA FOR APPELL'S HYPERGEOMETRIC FUNCTION $F_3$ \*

R. P. SINGAL†

**Abstract.** Mullen (SIAM J. Appl. Math., 14 (1966), pp. 1152-1163) gave differential recursion formulae for Appell's hypergeometric functions of two variables  $F_1$ ,  $F_2$ ,  $F_3$ , and  $F_4$ . The author pointed out (Dissertation, Punjabi University, Patiala, 1972) that Mullen's result for raising the double series denominator parameter for  $F_3$  is not correct. Here the correct result for  $F_3$  has been obtained.

Mullen [1] gave differential recursion formulae for Appell's hypergeometric functions  $F_1$ ,  $F_2$ ,  $F_3$  and  $F_4$ . The author [2] in his dissertation pointed out that Mullen's result for raising a double series denominator parameter for  $F_3(a, a'; b, b'; c; x, y)$  is not correct. In fact the correct result in Mullen's notation is

$$F_3(c+1) = \frac{c}{\Delta} [A - B(1-1/x)\theta - D(1-1/y)\phi + E(1-1/x-1/y)\theta\phi] F_3$$

where

$$A = \delta\delta' + (ab\delta + a'b'\delta')/F, \quad \delta = c - a - b, \quad \delta' = c - a' - b',$$

$$B = \delta' + (ab - a'b')/F, \quad D = \delta + (a'b' - ab)/F,$$

$$E = (\delta + \delta')/F, \quad F = c - a - a' - b - b',$$

$$E = (\delta + \delta')/F, \quad F = c - a - a' - b - b',$$

$$\Delta = c\delta\delta' + ab\delta' + a'b'\delta + [abc\delta + a'b'c\delta' + (a'b' - ab)^2]/F$$

$$\theta = x \frac{\partial}{\partial x}, \quad \phi = y \frac{\partial}{\partial y}.$$

*Proof.* The differential equations for  $F_3(c+1)$  are

$$[\theta(\theta + \phi + c) - x(\theta + a)(\theta + b)]F_3(c+1) = 0,$$

$$[\phi(\theta + \phi + c) - y(\phi + a')(\phi + b')]F_3(c+1) = 0$$

which can be rewritten as

$$(I) \quad (\theta + a)(\theta + b)F_3(c+1) = \frac{c}{x}\theta F_3,$$

$$(II) \quad (\phi + a')(\phi + b')F_3(c+1) = \frac{c}{y}\phi F_3,$$

with the help of the known operational result

$$(III) \quad (\theta + \phi + c)F_3(c+1) = cF_3.$$

Defining

$$(1-1/x)\theta F_3 = U, \quad (1-1/y)\phi F_3 = V, \quad (1-1/x-1/y)\theta\phi F_3 = W$$

\* Received by the editors July 26, 1978, and in revised form May 31, 1979.

† Department of Mathematics, G.N. College, Ferozepore-152001, India.

and taking

$$(\theta - \phi)(\text{III}) - (\text{I}) + (\text{II}),$$

we get

$$(\text{IV}) \quad (\delta\theta - \delta'\phi - ab + a'b')F_3(c+1) = c(U - V).$$

Again taking

$$(F\theta + F\phi - 2\theta\phi)(\text{III}) + (2\phi - F)(\text{I}) + (2\theta - F)(\text{II}),$$

we get

$$(\text{V}) \quad [(F\delta + 2a'b')\theta + (F\delta' + 2ab)\phi - F(ab + a'b')]F_3(c+1) \\ = cF(U + V) - 2cW.$$

Now eliminating  $\theta F_3(c+1)$  and  $\phi F_3(c+1)$  between (III), (IV) and (V) we get the desired result. I am thankful to the referee for his suggestions.

#### REFERENCES

- [1] JAMES A. MULLEN, *The differential recursion formulae for Appell's hypergeometric functions of two variables*, SIAM J. Appl. Math., 14 (1966), pp. 1152-1163.
- [2] R. P. SINGAL, *Dissertation*, Punjabi University, Patiala, India, 1972.

## AN APPROXIMATION THEOREM FOR A HAMMERSTEIN-TYPE EQUATION AND APPLICATIONS\*

VACLAV DOLEZAL†

**Abstract.** An approximation theorem for a Hammerstein-type equation in a Hilbert space is proved. Also, applications to the Ritz-Galerkin method and to construction of an approximating feedback system are discussed.

**1. Introduction.** In their paper [1] Brézis and Browder consider the Hammerstein-type equation  $(I + AB)x = y$  in a separable Banach space  $X$ . They give conditions for  $A$  and  $B$  under which the sequence of solutions  $x_n$  of  $(I + A_n B_n)x_n = P_n^* y$ ,  $n = 1, 2, \dots$  converges to the solution  $x$ . Here,  $A_n = P_n^* A P_n$ ,  $B_n = P_n B P_n^*$ .  $P_n^*$  is the conjugate of  $P_n$ , and  $(P_n)$  is a sequence of bounded projections each having a finite-dimensional range such that  $P_n z \rightarrow z$  as  $n \rightarrow \infty$  for every  $z \in X$ . In order to guarantee that  $x_n \rightarrow x$ , it is basically required that (a)  $A$  is continuous, monotone and bounded, and (b)  $B$  is continuous, angle-bounded, and maps bounded sets into weakly compact sets.

However, in many problems we encounter operators  $I + AB$  which do not satisfy (a), e.g., we can assume only that  $A + \alpha I$  is monotone for some  $\alpha > 0$ . Of course, in order to insure the existence and uniqueness of solutions as well as the convergence  $x_n \rightarrow x$ , we are compelled to strengthen the assumptions imposed on  $B$ .

The objective of the present paper is to do precisely this. To simplify our considerations, we assume that  $A$  and  $B$  are operators mapping a Hilbert space  $H$  into itself, and that the  $P_n$ 's are orthogonal projections. Our result is formulated in Theorem 1.

Moreover, as applications we consider the numerical aspects of Theorem 1, when  $A$  and  $B$  are linear and  $H$  is separable. In addition, we give a construction of feedback systems over a finite-dimensional space which approximate a given feedback system over an infinite-dimensional space.

**2. Results.** To simplify the formulation of results, let us introduce the following notation.

If  $H$  is a real Hilbert space, let  $\mathcal{M}(H)$  be the set of all operators  $N: H \rightarrow H$  such that

$$(1) \quad \mu_N = \inf_{\substack{x_1, x_2 \in H \\ x_1 \neq x_2}} \langle Nx_1 - Nx_2, x_1 - x_2 \rangle \|x_1 - x_2\|^{-2} > -\infty.$$

Similarly, let  $\text{Lip}(H)$  be the set of all operators  $N: H \rightarrow H$  such that

$$(2) \quad \|N\|^* = \sup_{\substack{x_1, x_2 \in H \\ x_1 \neq x_2}} \|Nx_1 - Nx_2\| \cdot \|x_1 - x_2\|^{-1} < \infty.$$

It is clear that we have:

- (i)  $N, M \in \mathcal{M}(H) \Rightarrow N + M \in \mathcal{M}(H)$  and  $\mu_{N+M} \geq \mu_N + \mu_M$ .
- (ii)  $\|\cdot\|^*$  is a seminorm, and  $N, M \in \text{Lip}(H) \Rightarrow NM \in \text{Lip}(H)$  with  $\|NM\|^* \leq \|N\|^* \|M\|^*$ .
- (iii) If  $N$  is linear, then  $N$  is bounded  $\Leftrightarrow N \in \text{Lip}(H)$ . In this case,  $\|N\|^* = \|N\|$ .
- (iv)  $\text{Lip}(H) \subset \mathcal{M}(H)$  and  $\|N\|^* \geq |\mu_N|$  for every  $N \in \text{Lip}(H)$ .

\* Received by the editors March 28, 1978, and in final revised form July 10, 1979.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794. This research was in part supported by the National Science Foundation under Grant MPS7505268.



LEMMA 1. Let  $N \in \mathcal{M}(H)$  and let  $\mu_N > 0$ ; if  $N$  is hemicontinuous, then  $N$  is invertible,  $N^{-1} \in \text{Lip}(H)$ ,  $\mu_{N^{-1}} \geq 0$  and

$$(3) \quad \|N^{-1}\|^* \leq \mu_N^{-1}.$$

If, in addition,  $N \in \text{Lip}(H)$ , then

$$(4) \quad \mu_{N^{-1}} \geq \mu_N \|N\|^{*-2}.$$

*Proof.* The assumption  $N \in \mathcal{M}(H)$  shows that

$$(5) \quad \langle Nx_1 - Nx_2, x_1 - x_2 \rangle \geq \mu_N \|x_1 - x_2\|^2$$

for all  $x_1, x_2 \in H$ . Hence,  $N$  is one-to-one.

Moreover, since  $\mu_N > 0$ ,  $N$  is monotone by (5), and since  $N$  is hemicontinuous, it is maximal monotone. Also, (5) shows that  $N$  is coercive, and consequently,  $NH = H$ , [2]. Thus,  $N$  is invertible.

Next, choosing  $y_i \in H$ ,  $i = 1, 2$  and putting  $x_i = N^{-1}y_i$  into (5), we get

$$(6) \quad \langle N^{-1}y_1 - N^{-1}y_2, y_1 - y_2 \rangle \geq \mu_N \|N^{-1}y_1 - N^{-1}y_2\|^2.$$

Hence,  $N^{-1} \in \mathcal{M}(H)$  with  $\mu_{N^{-1}} \geq 0$ . Furthermore, (6) yields by Schwarz inequality,  $\|N^{-1}y_1 - N^{-1}y_2\| \leq \mu_N^{-1} \|y_1 - y_2\|$ . Consequently,  $N^{-1} \in \text{Lip}(H)$  and  $\|N^{-1}\|^* \leq \mu_N^{-1}$ .

If, in addition,  $N \in \text{Lip}(H)$ , then  $\|Nx_1 - Nx_2\| \leq \|N\|^* \|x_1 - x_2\|$  for any  $x_1, x_2 \in H$ , so that  $\|N^{-1}y_1 - N^{-1}y_2\| \geq \|N\|^{*-1} \|y_1 - y_2\|$ . Introducing this inequality into (6) it follows that  $\mu_{N^{-1}} \geq \mu_N \|N\|^{*-2}$ . Hence the proof.

LEMMA 2. Let  $A \in \mathcal{M}(H)$  be hemicontinuous, and let  $B \in \text{Lip}(H)$  with  $\mu_B > 0$ . If  $\mu_A + \mu_B \|B\|^{*-2} > 0$ , then the operator  $N = I + AB$  is invertible,  $N^{-1} \in \text{Lip}(H)$  and

$$(7) \quad \|N^{-1}\|^* \leq \mu_B^{-1} (\mu_A + \mu_B \|B\|^{*-2})^{-1}.$$

*Proof.* The assumptions  $B \in \text{Lip}(H)$  and  $\mu_B > 0$  imply by Lemma 1 that  $B$  is invertible,  $B^{-1} \in \mathcal{M}(H)$ ,  $\mu_{B^{-1}} \geq \mu_B \|B\|^{*-2}$ ,  $B^{-1} \in \text{Lip}(H)$  and  $\|B^{-1}\|^* \leq \mu_B^{-1}$ . Thus, the operator  $B^{-1} + A$  is hemicontinuous,  $B^{-1} + A \in \mathcal{M}(H)$  and

$$(8) \quad \mu_{B^{-1}+A} \geq \mu_{B^{-1}} + \mu_A \geq \mu_A + \mu_B \|B\|^{*-2} > 0.$$

Hence, again by Lemma 1,  $B^{-1} + A$  is invertible,  $(B^{-1} + A)^{-1} \in \text{Lip}(H)$  and  $\|(B^{-1} + A)^{-1}\|^* \leq \mu_{B^{-1}+A}^{-1} \leq (\mu_A + \mu_B \|B\|^{*-2})^{-1}$  by (8).

On the other hand,  $N = (B^{-1} + A)B$ ; consequently,  $N$  is invertible, and  $N^{-1} = B^{-1}(B^{-1} + A)^{-1}$ . Hence, we have by the above,

$$(9) \quad \|N^{-1}\|^* \leq \|B^{-1}\|^* \|(B^{-1} + A)^{-1}\|^* \leq \mu_B^{-1} (\mu_A + \mu_B \|B\|^{*-2})^{-1}$$

which proves (7).

Note that the number  $\mu_A$  in Lemma 2 need not be nonnegative, i.e., it suffices that  $A + \alpha I$  is monotone for some  $\alpha > 0$ .

LEMMA 3. Let  $P_0 \neq 0$  be an orthogonal projection on  $H$ , let  $N: H \rightarrow H$  be an operator, and let  $[P_0N]: H_0 \rightarrow H_0$  be the restriction of  $P_0N$  to  $H_0 = P_0H$ .

(i) If  $N \in \mathcal{M}(H)$ , then  $[P_0N] \in \mathcal{M}(H_0)$  and

$$\mu_{[P_0N]} \geq \mu_N.$$

(ii) If  $N \in \text{Lip}(H)$ , then  $[P_0N] \in \text{Lip}(H_0)$  and

$$\|[P_0N]\|^* \leq \|N\|^*.$$

*Proof.* (i) We have

$$\begin{aligned} \mu_{[P_0N]} &= \inf_{\substack{x_1, x_2 \in H_0 \\ x_1 \neq x_2}} \langle [P_0N]x_1 - [P_0N]x_2, x_1 - x_2 \rangle \cdot \|x_1 - x_2\|^{-2} \\ &= \inf_{\substack{x_1, x_2 \in H_0 \\ x_1 \neq x_2}} \langle Nx_1 - Nx_2, x_1 - x_2 \rangle \|x_1 - x_2\|^{-2} \\ &\cong \inf_{\substack{x_1, x_2 \in H \\ x_1 \neq x_2}} \langle Nx_1 - Nx_2, x_1 - x_2 \rangle \|x_1 - x_2\|^{-2} \\ &= \mu_N > -\infty. \end{aligned}$$

The proof of (ii) follows a similar pattern.

Now we are ready to state the approximation theorem.

**THEOREM 1.** *Let  $A \in \mathcal{M}(H)$  be uniformly continuous, let  $B \in \text{Lip}(H)$  with  $\mu_B > 0$ , and let*

$$(10) \quad \mu_A + \mu_B \|B\|^{*-2} > 0.$$

*Furthermore, for  $n = 1, 2, \dots$ , let  $P_n: H \rightarrow H$  be an orthogonal projection such that  $P_n x \rightarrow x$  (strongly) as  $n \rightarrow \infty$  for any  $x \in H$ .*

*Let  $y \in H$  and let  $x \in H$  be the (unique) solution of  $(I + AB)x = y$ . Then, for each  $n = 1, 2, \dots$ , there exists a unique  $x_n \in H_n = P_n H$  such that*

$$(11) \quad (I + P_n A P_n B)x_n = P_n y,$$

*and we have  $x_n \rightarrow x$  (strongly) as  $n \rightarrow \infty$ .*

*Proof.* First, the existence and uniqueness of an  $x \in H$  satisfying the equation  $(I + AB)x = y$  is guaranteed by Lemma 2.

Next, observe that  $H_n = P_n H$  is a closed linear subspace of  $H$  and therefore it is a Hilbert space of its own right. By our hypothesis and Lemma 3 it follows that, for any  $n \geq 1$ ,  $[P_n A], [P_n B] \in \mathcal{M}(H_n)$ ,  $\mu_{[P_n A]} \geq \mu_A$ ,  $\mu_{[P_n B]} \geq \mu_B > 0$  and  $[P_n B] \in \text{Lip}(H_n)$ ,  $\|[P_n B]\|^{*} \leq \|B\|^{*}$ , where  $[P_n A]$  and  $[P_n B]$  is the restriction of  $P_n A$  and  $P_n B$  to  $H_n$ , respectively. Consequently, by (10),

$$(12) \quad \mu_{[P_n A]} + \mu_{[P_n B]} \|[P_n N]\|^{*-2} \geq \mu_A + \mu_B \|B\|^{*-2} > 0.$$

Also, it is clear that  $[P_n A]$  is hemicontinuous. Hence, by Lemma 2, the operator  $(I + [P_n A][P_n B]): H_n \rightarrow H_n$  is invertible. Since  $P_n y \in H_n$  for any  $y \in H$ , (11) possesses a unique solution  $x_n$  in  $H_n$ .

Next, denote  $\eta = \mu_A + \mu_B \|B\|^{*-2}$  and let

$$(13) \quad \begin{aligned} a &= \mu_B \quad \text{if } \mu_A \geq 0, \\ &= \|B\|^{*2} \eta \quad \text{if } \mu_A < 0. \end{aligned}$$

Thus, we always have  $a > 0$ .

Furthermore, observe that the family  $\{AP_n B: n = 1, 2, \dots\}$  is equicontinuous, i.e., for every  $\varepsilon_0 > 0$  there exists  $\delta_0 > 0$  such that, for any  $n \geq 1$  and  $x_1, x_2 \in H$  with  $\|x_1 - x_2\| < \delta_0$  we have

$$(14) \quad \|AP_n Bx_1 - AP_n Bx_2\| < \varepsilon_0.$$

Indeed, let  $\varepsilon_0 > 0$ ; then, by uniform continuity of  $A$ , there exists  $\delta' > 0$  such that  $\|Az_1 - Az_2\| < \varepsilon_0$  whenever  $z_1, z_2 \in H$  and  $\|z_1 - z_2\| < \delta'$ . Put  $\delta = \|B\|^{*-1} \delta' > 0$ . Choosing  $n \geq 1$  and  $x_1, x_2 \in H$  with  $\|x_1 - x_2\| < \delta$ , we have  $\|P_n Bx_1 - P_n Bx_2\| = \|P_n(Bx_1 - Bx_2)\| \leq$

$\|Bx_1 - Bx_2\| \leq \|B\|^* \|x_1 - x_2\| < \|B\|^* \delta = \delta'$ . Hence, (14) holds.

Choose now  $y \in H$ , and let  $x \in H$  be the solution of  $x + ABx = y$ . If  $x_n \in H_n$  satisfies (11), we have  $x_n = P_n x_n$ , and (11) can be written as

$$(15) \quad P_n(x_n + AP_n Bx_n) = P_n y.$$

Also,

$$(16) \quad P_n(x + ABx) = P_n y.$$

Consequently,

$$P_n(x_n - x + AP_n Bx_n - ABx) = 0,$$

i.e.,  $x_n - x + AP_n Bx_n - ABx \in H_n^\perp$ . Thus, for each  $c \in H_n$ ,

$$(17) \quad \langle c, x_n - x + AP_n Bx_n - ABx \rangle = 0.$$

However, (17) can be written as

$$(18) \quad \langle c, (x_n - P_n x) + ((AP_n B)x_n - (AP_n B)P_n x) \rangle = -\langle c, (P_n x - x) + (AP_n B P_n x - ABx) \rangle.$$

Now, let us put  $c = P_n(Bx_n - BP_n x) \in H_n$  into (18). We get

$$(19) \quad \begin{aligned} &\langle P_n(Bx_n - BP_n x), x_n - P_n x \rangle + \langle P_n Bx_n - P_n B P_n x, AP_n Bx_n - AP_n B P_n x \rangle \\ &= -\langle P_n(Bx_n - BP_n x), P_n x - x + (AP_n B P_n x - ABx) \rangle. \end{aligned}$$

Letting

$$(20) \quad h = \langle Bx_n - BP_n x, x_n - P_n x \rangle + \langle P_n Bx_n - P_n B P_n x, AP_n Bx_n - AP_n B P_n x \rangle,$$

we find that (19) reads

$$(21) \quad h = -\langle Bx_n - BP_n x, P_n(AP_n B P_n x - ABx) \rangle.$$

(We denoted the left-hand side of (19) by  $h$  and used the selfadjointness of  $P_n$ .)

On the other hand, by our hypotheses,

$$(22) \quad \mu_B \|x_n - P_n x\|^2 \leq \langle Bx_n - BP_n x, x_n - P_n x \rangle,$$

and

$$(23) \quad \mu_A \|P_n Bx_n - P_n B P_n x\|^2 \leq \langle P_n Bx_n - P_n B P_n x, AP_n Bx_n - AP_n B P_n x \rangle.$$

Now, if  $\mu_A \geq 0$ , then by (23), the second term on the right-hand side of (20) is nonnegative, and consequently  $\mu_B \|x_n - P_n x\|^2 \leq h$ . Thus, by our notation (13),

$$(24) \quad a \|x_n - P_n x\|^2 \leq h.$$

On the other hand, if  $\mu_A < 0$ , then  $\|P_n Bx_n - P_n B P_n x\| \leq \|Bx_n - BP_n x\| \leq \|B\|^* \|x_n - P_n x\|$ , so that

$$(25) \quad \mu_A \|P_n Bx_n - P_n B P_n x\|^2 \geq \mu_A \|B\|^{*2} \|x_n - P_n x\|^2.$$

Thus, by (22), (23) and (20),

$$(26) \quad (\mu_B + \mu_A \|B\|^{*2}) \|x_n - P_n x\|^2 \leq h.$$

However,  $\mu_B + \mu_A \|B\|^{*2} = \|B\|^{*2} (\mu_A + \mu_B \|B\|^{*-2}) = \|B\|^{*2} \eta = a$ . Hence, independently of the sign of  $\mu_A$ , we always have (24) with  $a > 0$ .

Using now (21), we can write

$$\begin{aligned}
 a\|x_n - P_nx\|^2 \leq h \leq |h| &\leq \|Bx_n - BP_nx\| \cdot \|P_n(AP_nBP_nx - ABx)\| \\
 &\leq \|B\|^* \|x_n - P_nx\| \cdot \|AP_nBP_nx - ABx\| \\
 &\leq \|B\|^* \|x_n - P_nx\| \{ \|(AP_nB)P_nx - (AP_nB)x\| + \|A(P_nBx) - A(Bx)\| \}.
 \end{aligned}$$

Hence,

$$(27) \quad \|x_n - P_nx\| \leq a^{-1} \|B\|^* \{ \|(AP_nB)P_nx - (AP_nB)x\| + \|A(P_nBx) - A(Bx)\| \}.$$

Observe that inequality (27) holds for any  $n \geq 1$ .

To conclude the proof, choose an  $\varepsilon > 0$ . Using the equicontinuity of  $\{AP_nB : n = 1, 2, \dots\}$  mentioned above it follows that there exists  $\delta_1 > 0$  such that

$$(28) \quad \|AP_nBu_1 - AP_nBu_2\| < \frac{\varepsilon}{4} a \|B\|^{*-1}$$

whenever  $\|u_1 - u_2\| < \delta_1$  and  $n$  is any integer. Moreover, by the uniform continuity of  $A$ , there exists  $\delta_2 > 0$  such that

$$(29) \quad \|Av_1 - Av_2\| < \frac{\varepsilon}{4} a \|B\|^{*-1}$$

whenever  $\|v_1 - v_2\| < \delta_2$ .

However, by our hypothesis,  $P_nw \rightarrow w$  for any  $w \in H$ . Thus, there exists integer  $M_1 > 0$  such that

$$(30) \quad \|P_nx - x\| < \min \left[ \delta_1, \frac{\varepsilon}{2} \right]$$

for all  $n \geq M_1$ . Similarly, there exists integer  $M_2 > 0$  such that

$$(31) \quad \|P_nBx - Bx\| < \delta_2$$

for all  $n \geq M_2$ .

Hence, putting  $M = \max [M_1, M_2]$ , we have for each  $n \geq M$ ,

$$\|AP_nBP_nx - AP_nBx\| < \frac{\varepsilon}{4} a \|B\|^{*-1}$$

and

$$\|AP_nBx - ABx\| < \frac{\varepsilon}{4} a \|B\|^{*-1}.$$

Consequently, by (27),  $\|x_n - P_nx\| < \varepsilon/2$ . Finally, since  $\|P_nx - x\| < \varepsilon/2$  for all  $n \geq M$ , it follows that  $\|x_n - x\| \leq \|x_n - P_nx\| + \|P_nx - x\| < \varepsilon$ . Thus,  $x_n \rightarrow x$  and the proof is complete.

Comparing our result with Theorem 4 in [1], we see that the former assumption on monotonicity of  $A$  was relaxed, but the assumption on  $B$  was strengthened.

**3. Applications.** Let us now consider two applications of Theorem 1.

(a) Solving the equation  $(I + AB)x = y$  becomes particularly simple, if  $H$  is separable (infinite dimensional), and if both operators  $A$  and  $B$  are linear. Note that in this case the assumptions  $A \in \mathcal{M}(H)$ , uniformly continuous, and  $B \in \text{Lip}(H)$  amount to the requirement that both  $A$  and  $B$  are bounded.

To outline the procedure, choose a fixed orthonormal basis  $\{e_i: i = 1, 2, \dots\}$  in  $H$ , and for each integer  $n \geq 1$  define  $P_n: H \rightarrow H$  by

$$(32) \quad P_n z = \sum_{i=1}^n \langle z, e_i \rangle e_i.$$

Clearly, every  $P_n$  is an orthogonal projection on  $H$ ,  $H_n = P_n H$  is the linear span of  $\{e_1, e_2, \dots, e_n\}$ , and  $P_n z \rightarrow z$  for every  $z \in H$ .

Moreover, the mapping  $T: H \rightarrow l_2$  defined by

$$(33) \quad Tz = (\langle z, e_1 \rangle, \langle z, e_2 \rangle, \dots)$$

is a norm-preserving isomorphism. Consequently, every linear bounded operator  $M: H \rightarrow H$  can be represented by a linear bounded operator  $\hat{M} = TMT^{-1}: l_2 \rightarrow l_2$ . However, interpreting elements in  $l_2$  as column vectors, it is easy to see that  $\hat{M}\xi$  is equal to the product  $\tilde{M} \cdot \xi$ , where the entries  $m_{ik}$  of the (infinite) matrix  $\tilde{M}$  are given by

$$(34) \quad m_{ik} = \langle Me_k, e_i \rangle, \quad i, k = 1, 2, \dots$$

Similarly, defining  $T_n: H_n \rightarrow R^n$  by

$$(35) \quad T_n z = (\langle z, e_1 \rangle, \langle z, e_2 \rangle, \dots, \langle z, e_n \rangle),$$

it follows that the restriction  $[P_n M]: H_n \rightarrow H_n$  of  $P_n M$  to  $H_n$  can be represented by an operator  $\hat{M}_n = T_n [P_n M] T_n^{-1}: R^n \rightarrow R^n$ .  $\hat{M}_n$  is described by a matrix  $\tilde{M}_n$  that is the upper left  $n \times n$  submatrix of  $\tilde{M}$ .

Using these facts, we confirm readily that the following is true.

Equation (11) has a unique solution  $x_n$  in  $H_n \Leftrightarrow$  the equation  $(I + \tilde{A}_n \tilde{B}_n) \cdot \xi_n = T_n y$  has a unique solution  $\xi_n$  in  $R^n$ . In this case,  $x_n = T_n^{-1} \xi_n$ . Thus, to find an approximation  $x_n$  to the exact solution  $x$  of  $(I + AB)x = y$ , it suffices to calculate matrices  $\tilde{A}_n, \tilde{B}_n$ , vector  $T_n y$  and solve the linear algebraic equation  $(I + \tilde{A}_n \tilde{B}_n) \cdot \xi_n = T_n y$ .

(b) The Hammerstein-type operator  $I + AB$  plays a central role in the theory of feedback systems. We are going to show that Theorem 1 permits us to construct feedback systems  $f_n$  which approximate a given system  $f$  (see Fig. 1). To this end, let us introduce the following definition.

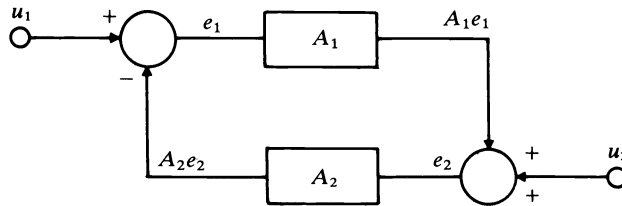


FIG. 1

Let  $H$  be a real Hilbert space, and let  $A_1, A_2: H \rightarrow H$ ; then the ordered pair  $f = [A_1, A_2]$  will be called a feedback system (further F.S.) over  $H$ .

(i) If  $(u_1, u_2) \in H^2 = H \times H$ , then a pair  $(e_1, e_2) \in H^2$  will be called a solution of  $f$  corresponding to  $(u_1, u_2)$ , if

$$(36) \quad e_1 = u_1 - A_2 e_2, \quad e_2 = u_2 + A_1 e_1.$$

(ii) The F.S.  $f$  is called normal, if for every  $(u_1, u_2) \in H^2$  there exists a unique solution  $(e_1, e_2) \in H^2$  of  $f$  corresponding to  $(u_1, u_2)$ .

A simple argument shows [3] that the following assertion is true.

LEMMA 4. Let  $f = [A_1, A_2]$ , and for each  $a \in H$  let the operator  $M_a: H \rightarrow H$  be defined by

$$(37) \quad M_a x = x + A_2(a + A_1 x).$$

Then  $f$  is normal  $\Leftrightarrow M_a$  is invertible for every  $a \in H$ . In this case, the solution  $(e_1, e_2)$  of  $f$  corresponding to  $(u_1, u_2) \in H^2$  is given by

$$(38) \quad (e_1, e_2) = (M_{u_2}^{-1} u_1, u_2 + A_1 M_{u_2}^{-1} u_1).$$

The approximation result mentioned above reads as follows.

THEOREM 2. Let  $f = [A_1, A_2]$  be a F.S. over  $H$ , and assume that  $A_1 \in \text{Lip}(H)$  with  $\mu_{A_1} > 0$ ,  $A_2 \in \mathcal{M}(H)$  is uniformly continuous, and that  $\mu_{A_2} + \mu_{A_1} \|A_1\|^{*-2} > 0$ . Furthermore, let  $P_n: H \rightarrow H, n = 1, 2, \dots$  be an orthogonal projection on  $H$  such that  $P_n x \rightarrow x$  (strongly) as  $n \rightarrow \infty$  for each  $x \in H$ . Then

- (a) The F.S.'s  $f$  and  $f_n = [P_n A_1, P_n A_2]$  over  $H_n = P_n H, n = 1, 2, \dots$  are normal.
- (b) If  $(u_1, u_2) \in H^2$ , and  $(e_1, e_2) \in H^2$  is the solution of  $f$  corresponding to  $(u_1, u_2)$ , and  $(e_1^n, e_2^n) \in H_n^2$  is the solution of  $f_n$  corresponding to  $(P_n u_1, P_n u_2) \in H_n^2, n = 1, 2, \dots$ , we have  $e_1^n \rightarrow e_1, e_2^n \rightarrow e_2$  (strongly) as  $n \rightarrow \infty$ .

*Proof.* Choose  $a \in H$  and define the operator  $B_a: H \rightarrow H$  by  $B_a x = a + A_1 x$ . Clearly,  $B_a \in \text{Lip}(H), \mu_{B_a} = \mu_{A_1} > 0$ , and by (37),  $M_a = I + A_2 B_a$ . Thus, by Theorem 1,  $M_a$  is invertible since  $y \in H$  is arbitrary. Hence,  $f$  is normal by virtue of Lemma 4.

Next, fix  $n \geq 1$ , choose  $b \in H_n$  and define  $M_b^{(n)}: H_n \rightarrow H_n$  by

$$(39) \quad M_b^{(n)} x = x + P_n A_2 (b + P_n A_1 x).$$

Since  $P_n b = b$ , we have  $M_b^{(n)} x = (I + P_n A_2 P_n B_b) x$  for each  $x \in H_n$ . Thus, referring to (11) in Theorem 1, for each  $y \in H_n$  there exists a unique  $x_n \in H_n$  such that  $M_b^{(n)} x_n = y$ , i.e., the operator  $M_b^{(n)}$  is invertible. Hence, by Lemma 4, the F.S.  $f_n$  over  $H_n$  is normal, and our claim (a) is proved.

To prove (b), choose  $(u_1, u_2) \in H^2$ . Then by (38) in Lemma 4,  $e_1 = M_{u_2}^{-1} u_1$ , so that

$$(40) \quad (I + A_2 B_{u_2}) e_1 = u_1.$$

Similarly,  $e_1^n = (M_{P_n u_2}^{(n)})^{-1} P_n u_1$ , i.e., by (39),

$$(41) \quad e_1^n + P_n A_2 (P_n u_2 + P_n A_1 e_1^n) = P_n u_1.$$

However, (41) can be written as

$$(42) \quad (I + P_n A_2 P_n B_{u_2}) e_1^n = P_n u_1.$$

Thus, invoking the second claim of Theorem 1 and (40), (42), it follows that  $e_1^n \rightarrow e_1$ .

Finally, by (36),  $e_2 = u_2 + A_1 e_1$  and  $e_2^n = P_n u_2 + P_n A_1 e_1^n$ . Since  $\|P_n\| = 1$ , we have

$$\begin{aligned} \|e_2^n - e_2\| &\leq \|P_n u_2 - u_2\| + \|P_n A_1 e_1^n - A_1 e_1\| \\ &\leq \|P_n u_2 - u_2\| + \|P_n A_1 e_1^n - P_n A_1 e_1\| + \|P_n A_1 e_1 - A_1 e_1\| \\ &\leq \|P_n u_2 - u_2\| + \|A_1\|^* \cdot \|e_1^n - e_1\| + \|P_n(A_1 e_1) - A_1 e_1\| \rightarrow 0. \end{aligned}$$

Hence the proof.

A comment on Theorem 2 is in order. For a majority of concrete F.S.'s the underlying space  $H$  is infinite-dimensional, but separable. Thus, if we define the orthogonal projection  $P_n$  by (32), Theorem 2 permits us to approximate such a F.S.  $f$  by a F.S.  $f_n$  over a finite-dimensional space  $H_n$ ; this can be readily modeled on a computer.

## REFERENCES

- [1] H. BRÉZIS AND F. E. BROWDER, *Nonlinear integral equations and systems of Hammerstein type*, *Advances in Math.*, 18 (1975), pp. 115–147.
- [2] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear monotone operators*, *Trans. Amer. Math. Soc.*, 149 (1970), pp. 75–88.
- [3] V. DOLEZAL, *Feedback systems described by monotone operators*, *SIAM J. Control Optimization*, 17 (1979), pp. 339–364.

## PERIODIC SOLUTIONS OF $x'' + g(t, x) = 0^*$

ROBERT R. STEVENS†

**Abstract.** Conditions are given which ensure that the nonlinear second order differential equation  $x'' + g(t, x) = 0$  has a nontrivial periodic solution with pre-assigned period. The results are obtained via the direct methods of the calculus of variations.

**1. Introduction.** In this paper we shall be concerned with the existence of periodic solutions of the differential equation

$$(I) \quad x'' + g(t, x) = 0 \quad (' = d/dt)$$

where the function  $g$  is assumed to be continuously differentiable for all real  $t, x$ .

This problem has a long history and many results have been established under various additional hypotheses concerning the function  $g$ . For example, Urabe [9] discusses the autonomous case ( $g$  independent of  $t$ ) and shows that if  $g$  satisfies a certain functional equation, then *all* solutions of (I) are periodic and have the same common period.

More recently, Jacobowitz [6] shows that (I) has an infinite number of periodic solutions of period  $2\pi$ , assuming the "superlinear" condition

$$g(t, x)/x \rightarrow \infty \quad \text{as } |x| \rightarrow \infty$$

for the function  $G$ . Using this same hypothesis, Hartman [5] establishes the existence of solutions of a wide class of separated two-point boundary value problems for (I). In both [5] and [6], the results are established using the Poincaré–Birkhoff (twist) fixed point theorem.

For a special case of (I), Nehari [7] proves the existence of infinitely many solutions of given period using an approach based on the calculus of variations and the hypothesis of superlinearity.

Many other similar results are presented in the book by Sansone and Conti [8] and are established using the fixed point theorems, mainly Brouwer's theorem.

Here we shall establish the existence of periodic solutions using the direct methods of the calculus of variations. The hypotheses which we use appear to be of a novel form; in particular, we shall make no use of the superlinearity condition.

For the principal result (Theorem 1), we shall assume that  $g(t, x)$  satisfies the following conditions:

- (1)  $g(t, x)$  is even and  $2\pi$ -periodic in  $t$ :  
 $g(t, x) = g(-t, x), \quad g(t + 2\pi, x) = g(t, x),$
- (2)  $xg(t, x) > 0$  for all  $t$  and all  $x \neq 0$ ,
- (3)  $g_x(t, x) > 0$  and  $g_x(t, 0) > 1$  for all  $t, x$  ( $g_x = \partial g / \partial x$ ),
- (4) there exist constants  $\beta, A, B$  with  $\beta < 2$  such  
 that  $G(t, x) \leq A|x|^\beta + B$  for all  $t, x$ .  
 (Here  $G(t, x) = \int_0^x g(t, s) ds$ .)

**THEOREM 1.** *If  $g(t, x)$  satisfies conditions (1), (2), (3), and (4) then there exists a nontrivial  $2\pi$  periodic solution of (I).*

\* Received by the editors August 9, 1978 and in revised form April 2, 1979.

† Department of Mathematics, University of Montana, Missoula, Montana 59801.



For the autonomous case ( $g(t, x) = g(x)$ ) we have the following result which we also state for the case of a single scalar equation.

**THEOREM 2.** *If  $\omega$  is a positive constant ( $\omega > 0$ ) and if (i)  $xg(x) > 0$  for all  $x \neq 0$ , (ii)  $dg/dx > 0$  for all  $x$  and  $dg/dx(0) > (\pi/\omega)^2$ , (iii) there exist constants  $\beta, A, B$  with  $\beta < 2$  such that  $\int_0^x g(s) ds \leq A|x|^\beta + B$  for all  $x$ , then the equation  $x'' + g(x) = 0$  has a nontrivial ( $\neq 0$ ) solution of period  $2\omega$ .*

**2. Proof of Theorems.** We assume throughout the discussion that the conditions (1), (2), (3), (4) hold. Let  $H$  denote the Hilbert space consisting of all real absolutely continuous functions  $\dot{x}(t)$  such that  $x'$  is square summable on  $[0, \pi]$ ,  $\int_0^\pi (x'(t))^2 dt < \infty$ , with inner product defined by

$$(x, y) = \int_0^\pi xy + x'y' dt.$$

We shall consider the isoperimetric problem of the minimum of the functional

$$J[x] = \int_0^\pi \frac{1}{2}(x')^2 - G(t, x) dt$$

on the constraint set

$$S = \left\{ x \in H : \int_0^\pi g(t, x) dt = 0 \right\}.$$

**LEMMA 1.** *If  $x \in H$  is a  $C^2$  (twice continuously differentiable) minimum point for the isoperimetric problem above, then  $x$  is a  $2\pi$  periodic solution of (I).*

*Proof.* The Euler equation here is

$$(5) \quad \lambda_0[x'' + g(t, x)] + \lambda_1 g_x(t, x) = 0,$$

where  $\lambda_0, \lambda_1$  ( $\lambda_0^2 + \lambda_1^2 = 1$ ) are the Lagrange multipliers. Natural boundary conditions (see [1], [2], [3]) for this nonfixed-endpoint problem are

$$(6) \quad \lambda_0 x'(0) = \lambda_0 x'(\pi) = 0.$$

Also  $\lambda_1 = 0$ . This follows by integrating (5) between 0 and  $\pi$  and using (6) and the constraint condition  $\int_0^\pi g(t, x) dt = 0$ . Hence  $\lambda_0 = 1$  and the function  $x$  satisfies

$$(7) \quad x'' + g(t, x) = 0, \quad x'(0) = x'(\pi) = 0.$$

The boundary conditions of (7) now imply that  $x$  is a  $2\pi$ -periodic solution of (I). This follows since the functions  $u(t) \equiv x(2\pi - t)$  and  $x(t)$  both satisfy (I) and  $u(\pi) = x(\pi)$ ,  $u'(\pi) = -x'(\pi) = 0 = x'(\pi)$ . This implies that  $x(t) \equiv x(2\pi - t)$ . Similarly,  $x(t)$  is even:  $x(t) \equiv x(-t)$ . Hence,  $x(t) \equiv x(-t) \equiv x(2\pi + t)$ .

Theorem 1 can now be proved by using Lemma 1 and showing that there exists a nonzero  $C^2$  solution of this isoperimetric problem. One of the main difficulties in doing this is to show that the functional  $J$  assumes negative values in the constraint set  $S$ ; i.e., that there exists  $y \in S$  satisfying  $J[y] < 0$ . That this actually happens is a consequence of the second part of condition (3). For (3) implies that

$$(8) \quad \lim_{x \rightarrow 0} \frac{G(t, x)}{x^2} = \frac{1}{2} g_x(t, 0) > \frac{1}{2}.$$

Thus there exists  $c > 0$  such that

$$(9) \quad G(t, x) > \frac{1}{2}x^2 \quad \text{for } 0 < |x| \leq c.$$

The functions

$$(10) \quad y_{a,b}(t) \equiv \begin{cases} -ac \cos t, & 0 \leq t \leq \pi/2, \\ -bc \cos t, & \pi/2 \leq t \leq \pi, \end{cases}$$

then satisfy  $|y_{a,b}(t)| \leq c$  for  $0 < a < 1, 0 < b < 1$ , and  $0 \leq t \leq \pi$ . Further, let  $F(a, b) \equiv \int_0^\pi g(t, y_{a,b}(t)) dt$ . Then for all  $0 < a < 1, 0 < b < 1$ :

- (i)  $y_{a,b}(t) < 0$  ( $> 0$ ) for  $t < \pi/2$  ( $t > \pi/2$ );
- (ii)  $y_{a,b} \in H$ ;
- (iii)  $J[y_{a,b}] < 0$ .

((iii) may be verified by a straightforward calculation using (9).) Also,  $F(a, b) < 0$  for  $(a, b)$  near  $(1, 0)$  and  $F(a, b) > 0$  for  $(a, b)$  near  $(0, 1)$ . (This follows from (2).) Thus, by continuity of  $F$ , there exists  $0 < a < 1, 0 < b < 1$  such that  $\int_0^\pi g(t, y_{a,b}(y)) dt = 0$  and  $J[y_{a,b}] < 0$ ; i.e., the functional  $J$  does assume negative value in the constraint set  $S$ .

Also, using (4), we have

LEMMA 2. Let  $y_0 \in S$  and  $J[y_0] = -k < 0$  and let  $T = \{y \in S | J[y] \leq -k\}$ . If  $y \in T$ , then

$$(11) \quad |y|_{\max} \leq C \quad (|y|_{\max} = \max |y(t)|, t \in [0, \pi]),$$

where  $C$  is a constant which depends only on  $\beta, A, B$ ; i.e., the set  $T$  is bounded relative to the supremum norm.

Proof. If  $y \in T$  then

$$\int_0^\pi \frac{1}{2} (y')^2 dt < \int_0^\pi G(t, y) dt \leq A\pi |y|_{\max}^\beta + B\pi.$$

Also  $y = 0$  for some  $t_0 \in [0, \pi]$ , since  $\int_0^\pi g(t, y) dt = 0$ . Hence, for all  $t \in [0, \pi]$ ,

$$|y(t)|^2 = \left| \int_{t_0}^t y'(s) ds \right|^2 \leq \pi \int_{t_0}^\pi |y'(s)|^2 ds < 2A\pi |y|_{\max}^\beta + 2B\pi$$

and  $|y|_{\max}^2 < 2A\pi |y|_{\max}^\beta + 2B\pi$ . The assertion of Lemma 2 now follows since  $\beta < 2$ .

We are now in a position to prove the main result, using what have become classical steps in the direct methods of the calculus of variations. (See [1], [2], [3].)

Proof of Theorem 1. Let  $\{y_n\}$  be a minimizing sequence for  $J$ :

$$y_n \in S \quad \text{and} \quad J[y_n] \rightarrow \mu,$$

where  $\mu = \inf J[y], y \in S$ . Note that we may assume that  $y_n \in T$  (Lemma 2). Since  $T$  is uniformly bounded relative to the supremum norm, it follows that  $J$  is bounded below on  $T$ , and  $\mu > -\infty$ . Then  $\{y_n\}$  is uniformly bounded in  $H$ ; i.e.,  $\|y_n\|_H^2 = \int_0^\pi y_n^2 + (y_n')^2 dt \leq \text{constant}$ . Hence, there exists a subsequence (we denote it also by  $\{y_n\}$ ) which converges weakly in  $H$  to  $x \in H$ . Hence  $y_n$  converges uniformly on  $[0, \pi]$  to  $x$ , and  $\int_0^\pi g(t, x) dt = \lim \int_0^\pi g(t, y_n) dt = 0$ . Also, since  $J$  is lower semi-continuous with respect to weak convergence in  $H$ ,

$$J[x] \leq \underline{\lim} J[y_n] \leq -k \quad \text{and} \quad x \in T.$$

Further, again using this lower semi-continuity of  $J, J[x] \leq \underline{\lim} J[y_n] = \mu$ . But also  $J[x] \geq \mu$ . Hence  $J[x] = \mu$ .

Actually  $x$  is a  $C^2$  function on  $[0, \pi]$ . As in [1, pp. 114–117], the functions

$$f_0(t) \equiv x'(t) + \int_0^t g(s, x(s)) ds - C_0$$

and

$$f_1(t) \equiv - \int_0^t g_x(s, x(s)) ds - C_1$$

( $C_0, C_1$  are constants chosen so that  $\int_0^\pi f_0 = \int_0^\pi f_1 = 0$ ) have Gram determinant equal to zero. Hence  $f_0$  and  $f_1$  are linearly dependent, in almost everywhere sense, on  $[0, \pi]$ :

$$l_0 f_0(t) + l_1 f_1(t) = 0 \quad (\text{a.e. in } [0, \pi]),$$

where  $l_0, l_1$  are constants  $l_0^2 + l_1^2 = 1$ . Clearly,  $l_0 \neq 0$ ; for otherwise  $\int_0^t g_x(s, x(s)) ds \equiv -C_1$  and  $g_x(t, x(t)) = 0$  for all  $t$ . This contradicts condition (3). Therefore  $f_0(t) = (-l_1/l_0)f_1(t)$  almost everywhere. That is,  $x'$  is equivalent to a continuously differentiable function. Hence,  $x$  is a nonzero  $C^2$  solution of the above isoperimetric problem and by Lemma 1, the proof is complete.

The proof of Theorem 2 is the same in every way as that for Theorem 1 except that (a) each occurrence of  $\pi$  is replaced by  $\omega$  and (b) each occurrence of  $\cos t$  in (10) is replaced by  $\cos(\pi/\omega)t$ .

#### REFERENCES

- [1] N. I. AKHIEZER, *The Calculus of Variations*, Blaisdell, New York, 1962.
- [2] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, IL, 1946.
- [3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, Wiley, New York, 1953.
- [4] G. M. EWING, *Calculus of Variations with Applications*, Norton, New York, 1969.
- [5] P. HARTMAN, *On boundary value problems for superlinear second order differential equations*, J. Differential Equations, 26 (1977), pp. 37-53.
- [6] H. JACOBOWITZ, *Periodic solutions of  $x'' + f(t, x) = 0$  via the Poincaré-Birkhoff theorem*, J. Differential Equations, 20 (1976), pp. 37-52.
- [7] Z. NEHARI, *Characteristic values associated with a class of nonlinear second order differential equations*, Acta Math., 105 (1961), pp. 141-175.
- [8] G. SANSONE AND R. CONTI, *Nonlinear Differential Equations*, Macmillan, New York, 1964.
- [9] M. URABE, *Potential forces which yield periodic motions of a fixed period*, J. Math. Mech., 10 (1961), pp. 569-578.

## GENERALIZED HANKEL MATRICES AND SYSTEM REALIZATION\*

MICHAEL A. ARBIB† AND ERNEST G. MANES‡

**Abstract.** We define the Hankel matrix of an adjoint system. Adjoint systems include linear and bilinear systems, automata, and group systems in both the time-varying and time-invariant cases. Our definition of the Hankel matrix unifies the familiar  $H_i^j = CA^{i+j}B$  of linear system theory (e.g. R. E. Kalman, P. L. Falb and M. A. Arbib, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969) with the bilinear Hankel matrix of A. Isidori (*Direct construction of minimal bilinear realizations from nonlinear input-output maps*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 626-631), T. J. Tarn and S. Nonoyama (*Realization of discrete-time internally bilinear systems*, Proc. IEEE Conf. Decision and Control, 76CH 1150-2CS (1976), pp. 125-133) and the Hankel matrix of M. Fliess (*Matrices de Hankel*, J. Math. Pure Appl., 53 (1974), pp. 197-224). The time-varying case is subsumed by regarding a time-varying system as a time-invariant system in a sequence category as in M. A. Arbib and E. G. Manes (*Time-varying systems*, SIAM J. Control, 13 (1975), pp. 1252-1270). For minimal realization theory and duality theory in the framework of this paper see B. D. O. Anderson, M. A. Arbib and E. G. Manes (*Foundations of system theory: Finitary and infinitary conditions*, Lecture Notes in Economics and Mathematical Systems, 115, Springer-Verlag, New York, 1976), M. A. Arbib and E. G. Manes (*Adjoint machines, state-behavior machines and duality*, J. Pure Appl. Algebra, 6 (1975), pp. 313-344) and S. J. Hegner (*Duality theory for discrete-time linear systems*, J. Comp. System Sci., 17 (1978), pp. 116-143). However, we lean much less heavily on category theory than in our earlier works on realization.

We introduce "adjoint correspondences" as the key algebraic ingredient in generalizing familiar linear system results to the nonlinear case. For example, the linear realizability criterion  $H_{i+1}^j = H_i^{j+1}$  does not make sense in the nonlinear setting; the precise condition needed is that " $H_{i+1}^j$  and  $H_i^{j+1}$  correspond under adjointness."

We provide a realizability theorem characterizing when a matrix  $H_i^j$  can be the Hankel matrix of a system, and offer partial realization and canonical realization theorems which associate systems with finite blocks of a Hankel matrix. We provide a general theory of "dimension in a category," and relate it to system realization via a simple recursion principle.

**1. Adjoint processes and systems.** In what follows,  $\mathcal{K}$  denotes an arbitrary category [3], [20]. Further axioms on  $\mathcal{K}$  will be added gradually—a summary appears before Lemma 2.13. In this section, we define adjoint processes and systems, and present a number of examples. Recall that a functor  $X: \mathcal{K} \rightarrow \mathcal{K}$  assigns to each object  $Q$  of  $\mathcal{K}$  another object  $QX$  of  $\mathcal{K}$  and assigns to each morphism  $f: Q \rightarrow R$  of  $\mathcal{K}$  another morphism of form  $fX: QX \rightarrow RX$  subject to the preservation of identities and composition, that is,  $id_{QX} = id_{QX}$  and, given  $f: Q \rightarrow R$ ,  $g: R \rightarrow S$ ,  $(gf)X = gXfX$ . As discussed below, basic examples include tensoring with a fixed vector space in the category of vector spaces and linear maps or assigning to a set  $Q$  the set of all functions from a fixed set to  $Q$  in the category of sets and functions. Let  $\mathcal{K}(Q, R)$  denote the set of morphisms from  $Q$  to  $R$  in  $\mathcal{K}$ .

**1.1. DEFINITION.** An adjoint process in  $\mathcal{K}$  is a pair  $(X, Z)$  of functors  $\mathcal{K} \rightarrow \mathcal{K}$  together with bijective correspondences  $\mathcal{K}(RX, S) \rightarrow \mathcal{K}(R, SZ)$  (one such for each pair  $(R, S)$  of objects) subject to the axiom that, given  $f: Q \rightarrow R$  and  $h: S \rightarrow T$ , if  $g: RX \rightarrow S$  and  $\psi: R \rightarrow SZ$  correspond then  $hg(fX): QX \rightarrow T$  and  $(hZ)\psi f: Q \rightarrow TZ$  correspond.

In the usual language of category theory, to say  $(X, Z)$  is an adjoint process amounts to saying that  $Z$  is right adjoint to  $X$  and, equivalently, that  $X$  is left adjoint to

---

\* Received by the editors March 10, 1978, and in revised form June 4, 1979. The research reported in this paper was supported in part by the National Science Foundation under grant DCR72-03722 A01.

† Department of Computer and Information Science, University of Massachusetts, Amherst, Massachusetts 01003.

‡ Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003.

Z. The literature of category theory offers a number of equivalent forms of these definitions. We have chosen the original one of [18, Def. 3.1].

For the duration of the paper we fix an adjoint process  $(X, Z)$  in  $\mathcal{K}$ . A convenient notation is the display

$$\frac{RX \xrightarrow{g} S}{R \xrightarrow{\psi} SZ}$$

to indicate that  $g$  and  $\psi$  correspond. We say  $g$  and  $\psi$  correspond under adjointness.

We may equally well use the notation

$$\frac{R \xrightarrow{\psi} SZ}{RX \xrightarrow{g} S}$$

and we will frequently use displays like

$$\frac{RX \xrightarrow{g} S}{\frac{R \xrightarrow{\psi} SZ}{RX \xrightarrow{k} S}}$$

to conclude that  $g = k$ .

The axiom is then succinctly displayed as

$$(1.2) \quad \frac{QX \xrightarrow{fX} RX \xrightarrow{g} S \xrightarrow{h} T}{Q \xrightarrow{f} R \xrightarrow{\psi} SZ \xrightarrow{hZ} TZ}$$

We call attention to the special cases that arise when  $f = id_R$  and when  $h = id_S$ .

1.3. LEMMA. For each  $B$ , define  $\varepsilon_B: BZX \rightarrow B$  as the correspondent of  $id_{BZ}: BZ \rightarrow BZ$ . Then if  $g: AX \rightarrow B$  and  $\psi: A \rightarrow BZ$  correspond under adjointness, we may recover  $g$  from  $\psi$  by

$$g = \varepsilon_B \cdot \psi X.$$

Proof. Applying (1.2) we conclude that

$$\frac{AX \xrightarrow{\psi X} BZX \xrightarrow{\varepsilon_B} B}{\frac{A \xrightarrow{\psi} BZ \xrightarrow{id_{BZ}} BZ}{AX \xrightarrow{g} B}}$$

□

1.4. Adjoint systems. An adjoint system is  $M = (Q, \delta, I, \tau, Y, \beta)$  where  $Q, I, Y$  are objects (the state object, input object and output object of  $M$ ) and  $\delta: QX \rightarrow Q, \tau: I \rightarrow Q$  and  $\beta: Q \rightarrow Y$  are morphisms (the dynamics, input map and output map of  $M$ ). (Note: “map” is here a synonym for “morphism.”) The codynamics of  $M$  is the map  $\Delta: Q \rightarrow QZ$  which corresponds to  $\delta$  under adjointness.

Given two dynamics  $\delta: QX \rightarrow Q$  and  $\theta: RX \rightarrow R$ , a *dynamorphism*  $h: (Q, \delta) \rightarrow (R, \theta)$  is a map  $h: Q \rightarrow R$  which “respects the dynamics”:

$$\begin{array}{ccc} QX & \xrightarrow{hX} & RX \\ \delta \downarrow & & \downarrow \theta \\ Q & \xrightarrow{h} & R \end{array}$$

The *time- $i$  reachability map*  $r_i: IX^i \rightarrow Q$  and the *time- $j$  observability map*  $\sigma_j: Q \rightarrow YZ^j$  are defined by

$$\begin{array}{c} r_0 = \tau \\ r_{i+1} = IX^{i+1} \xrightarrow{r_i X} QX \xrightarrow{\delta} Q \\ \sigma_0 = \beta \\ \sigma_{j+1} = Q \xrightarrow{\Delta} QZ \xrightarrow{\sigma_j Z} YZ^{j+1} \end{array}$$

The bisequence  $H_i^j$ , where  $H_i^j: IX^i \rightarrow YZ^j$  is defined by  $H_i^j = \sigma_j r_i$ , is the *Hankel matrix* of  $M$ .

Adjoint systems are closely related to the machines studied in [10], [11] and [13]. Realization theory for adjoint systems was developed in [1] and [4]. The Hankel matrix for adjoint systems is new, perhaps because the previous authors were motivated more by automata theory (where the Hankel matrix is not conventionally defined) than by system theory.

We conclude this section with a number of examples of adjoint systems and their Hankel matrices.

*Example 1.5. The decomposable case.* Here  $X = Z$  is the identity functor of  $\mathcal{K}$ . The realization theory in this special case was studied in [2]. When  $\mathcal{K}$  is the category of vector spaces (or of modules over a ring) an adjoint system is just a linear system

$$I \xrightarrow{B} Q \quad Q \xrightarrow{A} Q \quad Q \xrightarrow{C} Y.$$

The same system description holds in any category. The adjointness correspondence is just

$$\frac{Q \xrightarrow{s} R}{Q \xrightarrow{g} R}$$

so that the codynamics is again  $A$ . We have  $r_i = A^i B$  and  $\sigma_j = CA^j$  so that  $H_i^j = CA^{i+j} B$ .

*Example 1.6. Automata.* Let  $\mathcal{K}$  be the category of sets and functions. Let  $A$  be a fixed input alphabet. Define  $QX = Q \times A$ ,  $QZ = Q^A$ , the set of functions from  $A$  to  $Q$ . For  $f: Q \rightarrow R$ ,  $fX: Q \times A \rightarrow R \times A$  is defined by  $(q, a) \mapsto (f(q), a)$  whereas  $fZ: Q^A \rightarrow R^A$  sends  $g: A \rightarrow Q$  to  $fg: A \rightarrow R$ . The adjointness correspondence

$$\frac{Q \times A \xrightarrow{s} R}{Q \xrightarrow{\psi} R^A}$$

is the familiar  $(\psi q)(a) = g(q, a)$ . Let  $I$  have one element. Then  $\tau$  amounts to an element of  $Q$ , the initial state. The dynamics and output map have their usual forms  $\delta: Q \times A \rightarrow Q, \beta: Q \rightarrow Y$ . It is easily checked that  $r_i: A^i \rightarrow Q$  sends an  $i$ -tuple of input letters to the state reached from the initial state if the letters are inputted in sequence, whereas  $\sigma_j: Q \rightarrow Y^{(A^j)}$  sends  $q$  to the function  $A^j \rightarrow Y$  obtained by composing  $\beta$  with the time- $j$  reachability map that results if the initial state is  $q$ . Thus  $H_i^j: A^i \rightarrow Y^{(A^j)}$  is essentially a way of describing  $\beta \cdot r_{i+j}: A^{i+j} \rightarrow Y$  with emphasis on  $i$  as “present time.”

*Example 1.7. Internally bilinear machines* ([12], [16], [22]). Let  $\mathcal{K}$  be the category of real vector spaces and linear maps. Define  $QX = Q \otimes U$ , tensoring with a vector space  $U$ , while  $QZ = Q^U$ , the vector space of linear maps from  $U$  to  $Q$ . The adjointness correspondence

$$\begin{array}{ccc} Q \times U & \xrightarrow{\quad \xi \quad} & R \\ \hline Q & \xrightarrow{\quad \psi \quad} & R^U \end{array}$$

is then the familiar  $\psi q(u) = g(q \otimes u)$ . Let  $I$  be a vector space. Then  $\tau: I \rightarrow Q$  specifies the space  $\tau(I)$  of initial states “reachable in time 0,” the dynamics is then a bilinear map  $\delta: Q \otimes U \rightarrow Q$  while the output is a linear map  $\beta: Q \rightarrow Y$ .

It is easily checked that  $r_i: I \otimes U^{\otimes i} \rightarrow Q$  extends the map  $I \times U^i \rightarrow Q$  which sends a  $u$  in  $I$  and  $i$ -tuple of input vectors to the state reached from  $\tau(u)$  under that input sequence; whereas  $\sigma_j: Q \rightarrow Y^{U^j}$  sends  $q$  to the function  $U^j \rightarrow Y$  obtained by composing  $\beta$  with the time- $j$  reachability map that results if the initial state is  $q$ . The Hankel matrix  $H_i^j: I \otimes U^{\otimes i} \rightarrow Y^{U^j}$  can be viewed in a more symmetrical way as providing for each initial state label a matrix  $U^{\otimes i} \otimes U^{\otimes j} \rightarrow Y$ .

The previous three examples can be subsumed in one very general example, given below as example 1.11. But first we need to recall [3, § 1.2], [20, III.3, III.4] that if  $(Q_i: i \in I)$  is a family of objects of  $\mathcal{K}$  then their *product*  $pr_k: \prod Q_i \rightarrow Q_k$  satisfies the universal property that for all families of form  $f_i: Q \rightarrow Q_i$  ( $i \in I$ ) there exists unique  $f: Q \rightarrow \prod Q_i$  with  $pr_i f = f_i$  for all  $i$ . If it exists, the product is unique up to isomorphism.

(1.8)

The dual notion is the *coproduct*  $in_k: Q_k \rightarrow \coprod Q_i$ . As we see in (1.8), in both cases there is a bijective correspondence between arbitrary families  $(f_k: k \in I)$  and morphisms  $f$ . In the category of sets, coproducts are constructed as the disjoint union whereas in the category of modules over a ring (or a semiring), coproducts are constructed as weak direct sums. Both categories have products via the usual Cartesian product construction.

**PRESERVATION PRINCIPLE FOR ADJOINT PROCESSES 1.9.** *X preserves coproducts, that is, if  $in_k: Q_k \rightarrow \coprod Q_i$  is a coproduct, so is  $in_k X: Q_k X \rightarrow (\coprod Q_i) X$ . Similarly, Z preserves products.*

*Proof.* The result is standard in category theory [3, p. 134], [20, V.5]. To outline the proof, given a family  $f_k: Q_k X \rightarrow Q$ , let  $g_k: Q_k \rightarrow QZ$  correspond to  $f_k$  under adjointness, inducing the  $g: \prod Q_i \rightarrow QZ$  whose correspondent is the desired  $f$ . The second statement is dual.  $\square$

The following result should be viewed as generalizing the relationship between matrices and linear maps. (It is precisely that in the category of vector spaces, if  $I, J$  are finite and  $Q_i, R_j = \text{scalar field}$ .)

**BI-INDEX PRINCIPLE 1.10.** *If  $(Q_i: i \in I), (R_j: j \in J)$  and  $f_i^j: Q_i \rightarrow R_j$  then, so long as the coproduct and product exist, there exists a unique morphism  $f: \coprod Q_i \rightarrow \prod R_j$  such that  $pr_j f in_i = f_i^j$  for all  $i, j$ .*

*Proof.* Define  $f^j: \coprod Q_i \rightarrow R_j$  by  $f^j in_i = f_i^j$  and then define  $f$  by  $pr_j f = f^j$ . Uniqueness is left as an exercise.  $\square$

For the balance of the paper we assume our category  $\mathcal{K}$  to be such that every countable family of objects has a product and a coproduct.

**Example 1.11.** Let  $A$  be a fixed set (usually finite in applications). The following very general example of adjoint processes and systems works in any category  $\mathcal{K}$  obeying our standard assumptions, and subsumes Examples 5 ( $A$  has one element); 6 ( $\mathcal{K} = \text{Set}$ ) and 7 ( $\mathcal{K} = \text{Vect}$ , with  $A$  a basis for  $U$ ). Define

$$QX = Q \bullet A =_{\text{def}} \coprod_{a \in A} Q$$

the coproduct of  $|A|$  copies of  $Q$ . For  $f: Q \rightarrow R, fX$  is defined by the coproduct property

$$\begin{array}{ccc} Q & \xrightarrow{in_a} & QX \\ \downarrow f & & \downarrow fX \\ R & \xrightarrow{in_a} & RX \end{array} \quad (a \in A)$$

If we define  $Z$  by

$$QZ = Q^A =_{\text{def}} \prod_{a \in A} Q$$

the product of  $|A|$  copies of  $Q$ , with

$$\begin{array}{ccc} Q & \xleftarrow{pr_a} & QZ \\ \downarrow f & & \downarrow fZ \\ R & \xleftarrow{pr_a} & RZ \end{array}$$

it can easily be verified that  $(X, Z)$  is indeed an adjoint process, with the correspondence

$$\frac{QX \xrightarrow{g} R}{Q \xrightarrow{\psi} RZ}$$

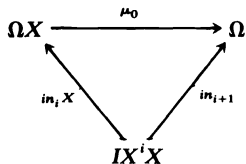
being simply given by  $g \cdot in_a = pr_a \cdot \psi: Q \rightarrow R$  for each  $a \in A$ .

Given a system  $(\tau: I \rightarrow Q, \delta: QX \rightarrow Q, \beta: Q \rightarrow Y)$ , we have that  $r_i: \prod_{v \in A^i} I \rightarrow Q, \sigma_j: Q \rightarrow \prod_{w \in A^j} Y$  and that, by the bi-index principle, the Hankel matrix  $H_i^j: \prod_{v \in A^i} I \rightarrow \prod_{w \in A^j} Y$  is equivalent to a  $|A^i| \times |A^j|$  “matrix” whose entries are the maps  $pr_w \cdot H_i^j \cdot in_v: I \rightarrow Y$ .

**2. Realizability and realizations.** Before stating the next theorem we define the *object of inputs*  $\Omega$  and the *observability space*  $\Gamma$ . The notation follows Kalman’s for the linear case [17, 10.3]. In [4] the notation used was  $IX^\circledast$  for  $\Omega$  and  $YX^\circledast$  for  $\Gamma$ .

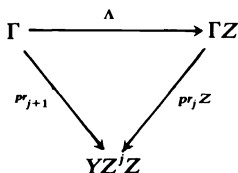


**2.1.** We set  $\Omega$  to be the coproduct  $\coprod (IX^i: i \geq 0)$  where  $IX^0 = I$  and  $IX^{i+1} = (IX^i)X$ .  $\Omega$  carries a dynamical structure  $\mu_0: \Omega X \rightarrow \Omega$  defined by



Here, we have used the preservation principle 1.9. (The story behind the cumbersome notation  $\mu_0$  instead of  $\mu$  is found in [21, § 4.2].)

**2.2.**  $\Gamma$  is defined as the product  $\prod (YZ^j: j \geq 0)$  with dynamical structure  $L: \Gamma X \rightarrow \Gamma$  the correspondent under adjointness of the map  $\Lambda$  defined by



These definitions coincide with Kalman's (save that he denotes both  $\mu_0$  and  $\Lambda$  by  $z$ ) when  $\mathcal{H}$  is the category of modules over a ring and when  $(X, Z)$  is the identity process.

**2.3. REALIZABILITY THEOREM.** Let  $H_i^j: IX^i \rightarrow YZ^j$  be an arbitrary bisequence of morphisms and let  $H: \Omega \rightarrow \Gamma$  be the unique morphism with  $\text{pr}_j H \text{in}_i = H_i^j$  as in 1.10. Then the following three conditions are equivalent (and we say  $H_i^j$  is a Hankel matrix with Hankel dynamorphism  $H$  if these conditions hold).

- (i)  $H_i^j$  is realizable, that is, is the Hankel matrix of some system.
- (ii) (The Hankel crossover condition): For all  $i, j$ :

$$\begin{array}{ccc}
 IX^{i+1} & \xrightarrow{H^{i+1}} & YZ^j \\
 \hline
 IX^i & \xrightarrow{H_i^{j+1}} & YZ^{j+1}
 \end{array}$$

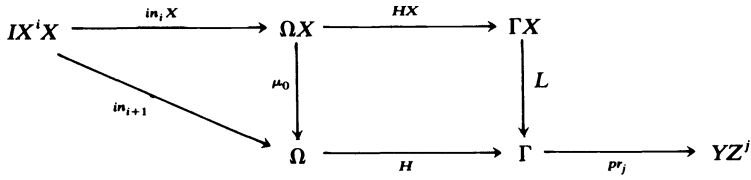
Equivalently, by 1.3, the condition states

$$\begin{array}{ccc}
 IX^i X & \xrightarrow{id} & IX^{i+1} \\
 H_i^{j+1} X \downarrow & & \downarrow H_{i+1}^j \\
 YZ^{j+1} X & \xrightarrow{\epsilon} & YZ^j
 \end{array}$$

- (iii)  $H: (\Omega, \mu_0) \rightarrow (\Gamma, L)$  is a dynamorphism, that is,  $L(HX) = H\mu_0: \Omega X \rightarrow \Gamma$ .
- Proof.* (i)  $\Rightarrow$  (ii) is immediate from

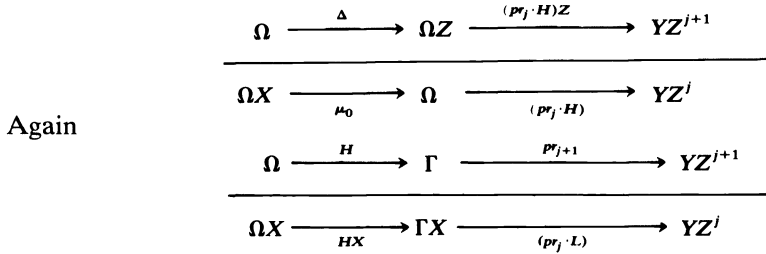
$$\begin{array}{ccccccc}
 IX^i X & \xrightarrow{\tau_i X} & QX & \xrightarrow{\delta} & Q & \xrightarrow{\sigma_j} & YZ^j \\
 \hline
 IX^i & \xrightarrow{\tau_i} & Q & \xrightarrow{\Delta} & QZ & \xrightarrow{\sigma_j Z} & YZ^j Z
 \end{array}$$

For (ii)  $\Rightarrow$  (iii), consult the diagram



By principles 1.9 and 1.10, it suffices to prove that the bottom and top paths from  $IX^i X$  to  $YZ^j$  are equal. But the bottom path is exactly  $H^i_{i+1}$ , whereas  $pr_j L$  corresponds under adjointness to  $pr_{j+1}: \Gamma \rightarrow YZ^{j+1}$  so that the top path corresponds to  $H^i_{i+1}$  and is thus also  $H^i_{i+1}$ .

To complete the proof we show (iii)  $\Rightarrow$  (i). We shall show that if  $H$  is a dynamorphism, then the “free realization”  $Q = \Omega$ ,  $\delta = \mu_0$ ,  $\tau = in_0$ ,  $\beta = pr_0 H$  has Hankel matrix  $(H^i_i)$ . One checks easily that  $r_i = in_i$ . To show that  $\sigma_{r_i} = pr_j H in_i$  it suffices to show that  $\sigma_j = pr_j H$ . This is true by definition for  $j = 0$ . The inductive step here is given by using the adjointness axiom with  $f = id_\Omega$ , and where  $\Delta$  is now the codynamics of  $\mu_0$ :



Again

But by the dynamorphism property,  $H \cdot \mu_0 = L \cdot HX$ , and so  $pr_{j+1} \cdot H = (pr_j \cdot H)Z \cdot \Delta = \sigma_{j+1}$ .  $\square$

The Hankel crossover condition provides evidence that adjointness arises naturally in system theory. In the decomposable case (example 1.5) we capture the familiar condition  $H^i_{i+1} = H^{i+1}_i$  of linear system theory.

In the general context of example 1.11,  $\Omega$  may be identified with  $I \bullet A^*$ , where  $A^*$  is the free monoid generated by  $A$  and  $\Gamma$  may be identified with  $Y^{A^*}$ .

In familiar system examples one can discuss the subspace of  $Q$  reached by time  $i$ . Such a subspace may be constructed by “taking the image” of the map  $f: \coprod (IX^k: 0 \leq k \leq i) \rightarrow Q$  defined by  $f in_k = r_k$ . To formalize taking the image we structure  $\mathcal{H}$  with an image factorization system.

**2.4.** An *image factorization system* for a category  $\mathcal{H}$  is a pair  $(\mathcal{E}, \mathcal{M})$  where  $\mathcal{E}, \mathcal{M}$  are subclasses of morphisms satisfying the following four axioms:

IFS1.  $\mathcal{E}$  and  $\mathcal{M}$  are each closed under composition.

IFS2. Every isomorphism is both in  $\mathcal{E}$  and in  $\mathcal{M}$ .

IFS3. Every element of  $\mathcal{E}$  is an epimorphism and every element of  $\mathcal{M}$  is a monomorphism. (A map  $f: R \rightarrow S$  is an *epimorphism* if whenever  $g, h: S \rightarrow T$  satisfy  $gf = hf$ , then  $g = h$ ; dually,  $f$  is a *monomorphism* if whenever  $a, b: Q \rightarrow R$  satisfy  $fa = fb$  then  $a = b$ .)

IFS4. Every morphism  $f: Q \rightarrow R$  admits an  $\mathcal{E}$ - $\mathcal{M}$  factorization  $(e, m)$ —that is,  $f = me$  with  $e \in \mathcal{E}$  and  $m \in \mathcal{M}$ —and such factorizations are unique up to isomorphism in the sense that if  $(e', m')$  is another one then there exists a unique isomorphism  $\psi$  with

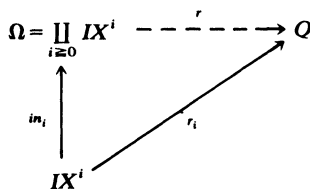
$\psi e = e'$  and  $m'\psi = m$ . We then may also denote any  $S$  with  $Q \rightarrow S \rightarrow R$  (an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $f$  as  $\text{Im}(f)$ ).

The category of sets and the category of modules over a ring both have  $\mathcal{E}$  = surjections and  $\mathcal{M}$  = injections as unique image factorization system. Thus in these categories  $\text{Im}(f)$  is the usual image  $f(Q)$  of  $f: Q \rightarrow R$ . The same construction works in the category of semigroups but in that category  $\mathcal{E}$  = epimorphisms determines (see (2.8) below) another system; the inclusion of the natural numbers into the integers is a nonsurjective epimorphism in that category. Image factorization systems in the category of linearly topologized vector spaces were investigated in a system-theoretic context in [14].

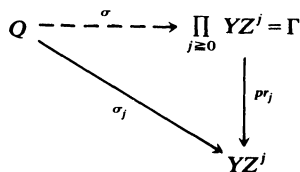
The notion of an image factorization system can be traced to [19]. The version presented here is due to [15]. References in the system literature to (2.7) below as the Zeiger fill-in lemma are historically inaccurate.

For the balance of this paper,  $(\mathcal{E}, \mathcal{M})$  is a fixed image factorization system in  $\mathcal{K}$ .

DEFINITION 2.5. Let  $M$  be an adjoint system. The reachability map  $r: \Omega \rightarrow Q$  of  $M$  is defined by



Dually, the observability map  $\sigma: Q \rightarrow \Gamma$  of  $M$  is defined by



We say  $M$  is *reachable* if  $r$  is in  $\mathcal{E}$ , *observable* if  $\sigma$  is in  $\mathcal{M}$ .  $M$  is *reachable in time  $i$*  if  $(r_k \mid 0 \leq k \leq i): \coprod (IX^k \mid 0 \leq k \leq i) \rightarrow Q$  is in  $\mathcal{E}$ . Correspondingly,  $M$  is *observable in time  $j$*  if  $(\sigma_k \mid 0 \leq k \leq j)$  is in  $\mathcal{M}$ .

$M$  is *reachable in bounded time* if  $M$  is reachable in time  $i$  for some  $i$ ; and  $M$  is *observable in bounded time* if  $M$  is observable in time  $j$  for some  $j$ .

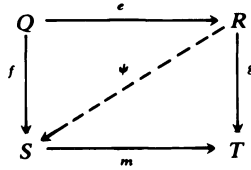
We conclude this section by collecting a number of standard results. Proofs of 2.6–2.11 appear in [21, § 3.4] although all are easy exercises.

Let  $\mathcal{K}^{op}$  denote the dual category of  $\mathcal{K}$ . For a discussion of duality for adjoint systems see the end of § 4 and [4]. Thus products in  $\mathcal{K}$  = coproducts in  $\mathcal{K}^{op}$ , monomorphisms in  $\mathcal{K}$  = epimorphisms in  $\mathcal{K}^{op}$  and

PROPOSITION 2.6.  $(\mathcal{M}, \mathcal{E})$  is an image factorization system in the opposite category  $\mathcal{K}^{op}$ .

Proposition 2.6 clearly plays a role in establishing the duality between results on reachability and corresponding results on observability—e.g. the fact noted after Proposition 2.10.

PROPOSITION 2.7. (Diagonal fill-in). *Given a commutative square  $ge = mf$*



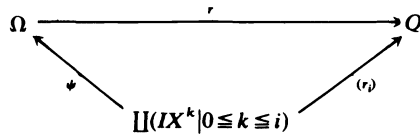
with  $e \in \mathcal{E}$ ,  $m \in \mathcal{M}$  there exists (necessarily unique)  $\psi$  with  $\psi e = f$  and  $m\psi = g$ . [Hint for proof: take the images of  $f$  and  $g$ .]

PROPOSITION 2.8. ( $\mathcal{E}$  determines  $\mathcal{M}$ ). *The converse of diagonal fill-in holds. That is, if  $m$  is an arbitrary morphism with the property that whenever  $ge = mf$  with  $e \in \mathcal{E}$  there exists  $\psi$  with  $\psi e = f$  then necessarily  $m \in \mathcal{M}$ . [Hint for proof: factor  $m = m'e$  and let  $f = id_S$ .] Dually,  $\mathcal{M}$  determines  $\mathcal{E}$ .*

PROPOSITION 2.9. *If  $f \in \mathcal{E}$  and  $f \in \mathcal{M}$  then  $f$  is an isomorphism.*

PROPOSITION 2.10. *If  $f: Q \rightarrow R$  and  $g: R \rightarrow S$  then  $gf \in \mathcal{E}$  implies  $g \in \mathcal{E}$  whereas  $gf \in \mathcal{M}$  implies  $f \in \mathcal{M}$ . [Hint for proof: use 2.8.]*

Reachability in bounded time implies reachable. To prove this, observe that

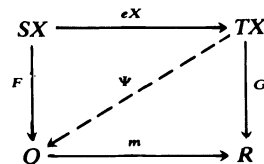
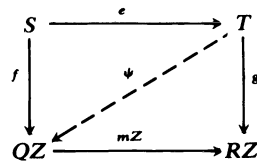


where  $\psi$  is defined by  $\psi in_k = in_k$ . Thus  $(r_i) = r \cdot \psi$  in  $\mathcal{E}$  implies  $r$  in  $\mathcal{E}$ . Dually, observability in bounded time implies observable.

PROPOSITION 2.11. *Given a family  $f_i: Q_i \rightarrow R_i$  with each  $f_i \in \mathcal{M}$  then the unique  $f: \coprod Q_i \rightarrow \coprod R_i$  defined by  $pr_i f = f_i pr_i$  is also in  $\mathcal{M}$ . Dually, given a family  $f_i: Q_i \rightarrow R_i$  with each  $f_i \in \mathcal{E}$ , the unique  $f: \coprod Q_i \rightarrow \coprod R_i$  with  $f in_i = in_i f_i$  is again in  $\mathcal{E}$ . [Hint for proof: use 2.8.]*

PROPOSITION 2.12.  *$X$  preserves  $\mathcal{E}$  if and only if  $Z$  preserves  $\mathcal{M}$ .*

*Proof.* Assuming  $X$  preserves  $\mathcal{E}$  we wish to show that  $mZ: QZ \rightarrow RZ \in \mathcal{M}$  given that  $m: Q \rightarrow R \in \mathcal{M}$ . This is immediate from 2.8 and the adjoint correspondences



(where capital and lower case letters correspond). The converse result is dual.  $\square$

We can now state all necessary standing assumptions and summarize them here for convenience. For the balance of the paper we will assume that:

$\mathcal{K}$  is a category with products and coproducts of countable families.

$(X, Z)$  is an adjoint process in  $\mathcal{K}$ .

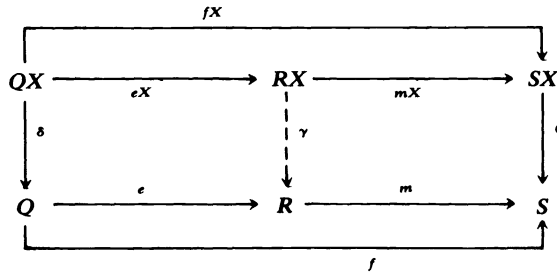
$(\mathcal{E}, \mathcal{M})$  is an image factorization system in  $\mathcal{K}$ .

$X$  preserves  $\mathcal{E}$  (and hence  $Z$  preserves  $\mathcal{M}$ ).

$I$  is a fixed input object.  $Y$  is a fixed output object.

It is often the case that  $\mathcal{E}$  = all epimorphisms or that  $\mathcal{E}$  = all morphisms which are the coequalizer of some pair [3, § 1.3], [20, p. 64]. This is the case for the category of sets and for the category of modules over a ring, the unique  $\mathcal{E}$  being the class of epimorphisms = the class of all coequalizers. In these two cases, it is well known that  $X$  must preserve  $\mathcal{E}$  [20, V.5].

DYNAMORPHIC IMAGE LEMMA 2.13. *Let  $f: (Q, \delta) \rightarrow (S, \theta)$  be a dynamorphism, that is, the perimeter of the diagram below commutes.*



Let  $f = me$  be an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $f$ . Then there exists a unique dynamics  $\gamma: RX \rightarrow R$  rendering the above diagram commutative.

*Proof.* Since  $eX \in \mathcal{E}$ , this is immediate from 2.7.  $\square$

DEFINITION 2.14. The canonical realization  $M_H$  of a Hankel matrix  $H^i_i$  is the system  $(Q_H, \delta_H, \tau_H, \beta_H)$  defined as follows. Let  $H: (\Omega, \mu_0) \rightarrow (\Gamma, L)$ , defined by  $pr_j \cdot H \cdot in_i = H^i_j$ , be the Hankel dynamorphism of Theorem 2.3. Let

$$\Omega \xrightarrow{r_H} Q_H \xrightarrow{\sigma_H} \Gamma$$

be an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $H$ . By the dynamorphic image lemma, there exists a unique dynamics  $\delta_H: Q_H X \rightarrow Q_H$  rendering  $r_H$  and  $\sigma_H$  dynamorphisms. Define  $\tau_H = r_H in_0$  and  $\beta_H = pr_0 \sigma_H$ .

It is proved in [4, Thms. 2.1, 3.15] that  $M_H$  is a realization of  $H^i_j$ , that the reachability and observability maps of  $M_H$  are  $r_H$  and  $\sigma_H$  (so that  $M_H$  is reachable and observable) and that any other reachable and observable realization is isomorphic to  $M_H$ .

The question of interest here, however, is under what conditions the canonical realization can be found from a finite fragment of the Hankel matrix  $(H^i_j | 0 \leq i \leq k, 0 \leq j \leq n)$ . We first present a partial realization which tells us when such a fragment lets us define an adjoint system whose behavior is consistent with that portion of the Hankel matrix. Then, in the remaining sections, we present general conditions on  $m$  and  $n$  under which this partial realization will be isomorphic to the canonical realization.

Let us fix the following notations:

$$\bar{n} = \{0, 1, \dots, n\}$$

$$IX^{\bar{k}} = \coprod_{i \in \bar{k}} IX^i; \quad YX^{\bar{n}} = \prod_{j \in \bar{n}} YZ^j$$

while

$$H_k^{\bar{n}}: IX^{\bar{k}} \rightarrow YZ^{\bar{n}}$$

is defined by  $pr_j \cdot H_k^{\bar{n}} \cdot in_i = H_i^j$  for  $i \in \bar{k}, j \in \bar{n}$ .

$$\begin{aligned} \text{Define } \tilde{\mu}: IX^{\bar{k}}X \rightarrow IX^{\bar{k}+1} & \text{ by } \tilde{\mu} \cdot in_i X = in_{i+1} \\ \tilde{\varepsilon}: YZ^{\bar{n}+1}X \rightarrow YZ^{\bar{n}} & \text{ by } pr_j \cdot \tilde{\varepsilon} = \varepsilon_{YZ^j} \cdot pr_{j+1}X. \end{aligned}$$

Then the Hankel crossover condition yields

$$(2.15) \quad \begin{array}{ccc} IX^{\bar{k}}X & \xrightarrow{\tilde{\mu}} & IX^{\bar{k}+1} \\ H_k^{\bar{n}+1}X \downarrow & & \downarrow H_{k+1}^{\bar{n}} \\ YZ^{\bar{n}+1}X & \xrightarrow{\tilde{\varepsilon}} & YZ^{\bar{n}} \end{array}$$

(just precede the square by  $in_i X$  and follow it by  $pr_j$  for  $0 \leq i \leq k, 0 \leq j \leq n$  to recapture the square of 2.3(ii)).

Let  $H_k^{\bar{n}+1}$  have  $\mathcal{E}$ - $\mathcal{M}$  factorization  $(\bar{e}, \bar{m})$  with image  $\bar{Q}$  while  $H_{k+1}^{\bar{n}}$  factors as  $(\hat{e}, \hat{m})$  with image  $\hat{Q}$ . Then, since  $X$  preserves  $\mathcal{E}$ , we may define  $\tilde{\delta}: \bar{Q}X \rightarrow \hat{Q}$  by diagonal fill-in:

$$(2.16) \quad \begin{array}{ccc} IX^{\bar{k}}X & \xrightarrow{\tilde{\mu}} & IX^{\bar{k}+1} \\ \bar{e}X \downarrow & & \downarrow \hat{e} \\ \bar{Q}X & \xrightarrow{\tilde{\delta}} & \hat{Q} \\ \bar{m}X \downarrow & & \downarrow \hat{m} \\ YZ^{\bar{n}+1}X & \xrightarrow{\tilde{\varepsilon}} & YZ^{\bar{n}} \end{array}$$

The important fact is that  $\tilde{\delta}$  is completely determined by the  $H_i^j$  for  $0 \leq i \leq k+1, 0 \leq j \leq n+1$ .

To obtain our partial realization theorem, we must establish conditions under which  $\tilde{\delta}$  may be viewed as a dynamics. To this end, define

$$\begin{aligned} in: IX^{\bar{k}} &\rightarrow IX^{\bar{k}+1} & \text{ by } in \cdot in_i &= in_i, \\ pr: YZ^{\bar{n}+1} &\rightarrow YZ^{\bar{n}} & \text{ by } pr_j \cdot pr &= pr_j. \end{aligned}$$

Then the bi-index principle, 1.10, yields

$$(2.17) \quad \begin{array}{ccc} IX^{\bar{k}} & \xrightarrow{H_k^{\bar{n}+1}} & YZ^{\bar{n}+1} \\ in \downarrow & \searrow H_k^{\bar{n}} & \downarrow pr \\ IX^{\bar{k}+1} & \xrightarrow{H_{k+1}^{\bar{n}}} & YZ^{\bar{n}} \end{array}$$

Forming the  $\mathcal{E}$ - $\mathcal{M}$  factorization  $(e, m)$  of  $H_k^{\bar{n}}$  with image  $R$ , we then obtain  $t$  and  $u$  by diagonal fill-in:

$$(2.18) \quad \begin{array}{ccccc} IX^{\bar{k}} & \xrightarrow{\bar{e}} & \bar{Q} & \xrightarrow{\bar{m}} & YZ^{\bar{n}+1} \\ \downarrow in & \searrow e & \downarrow t & & \downarrow pr \\ & & R & \xrightarrow{m} & \\ & & \downarrow u & & \\ IX^{k+1} & \xrightarrow{\hat{e}} & \hat{Q} & \xrightarrow{\hat{m}} & YZ^{\hat{n}} \end{array}$$

PARTIAL REALIZATION THEOREM 2.19. *If  $t, u$  are isomorphisms in 2.18, we may define the system  $M = (\bar{Q}, \delta, \tau, \beta)$  by*

$$(2.20) \quad \delta = t^{-1} \cdot u^{-1} \cdot \tilde{\delta}: \bar{Q}X \rightarrow \bar{Q} \quad (\text{using 2.16 and 2.18})$$

$$(2.21) \quad \tau = \bar{e} \cdot in_0: I \rightarrow IX^{\bar{k}} \rightarrow \bar{Q},$$

$$(2.22) \quad \beta = pr_0 \cdot \bar{m}: \bar{Q} \rightarrow YZ^{\bar{n}+1} \rightarrow Y.$$

*Then the Hankel matrix of  $M$  agrees with  $H_i^j$  for  $0 \leq i \leq n, 0 \leq j \leq k$ .*

*Proof.* Let  $r_i, \sigma_j$  be the  $i$ -step reachability and  $j$ -step observability maps, respectively, of  $M$ . We prove the theorem in two steps:

(i) We show  $r_i = IX \xrightarrow{in_i} IX^{\bar{k}} \xrightarrow{\bar{e}} \bar{Q}$  for  $0 \leq i \leq k$ .

(ii) We show  $\sigma_j = \bar{Q} \xrightarrow{\bar{m}} YZ^{\bar{n}+1} \xrightarrow{pr_j} YZ^j$  for  $0 \leq j \leq k$ .

It is then immediate that  $\sigma_j \cdot r_i = pr_j \cdot \bar{m} \cdot \bar{e} \cdot in_i = pr_j \cdot H_k^{n+1} \cdot in_i = H_i^j$ .

*Proof of (i).* For  $i = 0$ , this is 2.21. Now, for  $0 \leq i < k$ , we have by induction

$$\begin{aligned} r_{i+1} &= \delta \cdot r_i X \\ &= t^{-1} u^{-1} \cdot \tilde{\delta} \cdot r_i X \\ &= t^{-1} u^{-1} \cdot \tilde{\delta} \cdot \bar{e} X \cdot in_i X && \text{by induction hypothesis} \\ &= t^{-1} u^{-1} \cdot \hat{e} \cdot \tilde{\mu} \cdot in_i X && \text{by 2.16} \\ &= t^{-1} u^{-1} \cdot \hat{e} \cdot in_{i+1} && \text{by definition of } \tilde{\mu} \\ &= t^{-1} u^{-1} \cdot u \cdot t \cdot \bar{e} \cdot in_{i+1} && \text{by 2.18, and definition of } in \\ &= \bar{e} \cdot in_{i+1} && \text{as was to be shown.} \end{aligned}$$

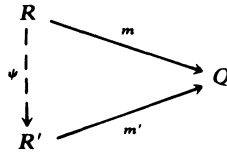
*Proof of (ii).* By definition  $\sigma_0 = \beta = pr_0 \cdot m$ . But then, for  $0 \leq j < k$ ,

$$\begin{aligned} \frac{\sigma_{j+1} = \sigma_j Z \cdot \Delta}{\sigma_j \cdot \delta} &&& \text{by 1.2} \\ &= pr_j \cdot \bar{m} \cdot \delta && \text{by induction hypothesis} \\ &= pr_j \cdot \hat{m} \cdot u \cdot t \cdot \delta && \text{by 2.18, and definition of } pr \\ &= pr_j \cdot \hat{m} \cdot \tilde{\delta} && \text{by definition of } \delta \\ &= pr_j \cdot \hat{e} \cdot \bar{m} X && \text{by 2.16} \\ &= \varepsilon_{YZ^j} \cdot pr_{j+1} X \cdot \bar{m} X && \text{by definition of } \hat{e}. \\ \hline &pr_{j+1} \cdot \bar{m} && \end{aligned}$$

□

**3. Generalizing the notion of finite dimensionality.** For a linear system  $M$ , the subspaces  $Q_i$  generated by the union of the images  $A^k B: I \rightarrow Q, 0 \leq k \leq i$ , constitute an ascending chain of subspaces of  $Q$ . If  $Q$  is finite-dimensional—or more generally, for modules over a ring rather than vector spaces, if  $Q$  is Noetherian—this chain is eventually stationary,  $Q_m = Q_{m+1} = \dots$ , and  $M$  is reachable in time  $m$ . In this section, we show how such dimensionality considerations may be extended to our category  $\mathcal{K}$ —with dimension reducing, essentially, to cardinality in the case of **Set**. The notions of  $\mathcal{E}$ -height and  $\mathcal{M}$ -height were introduced in [1]. Further properties of Noetherian objects appear in [7].

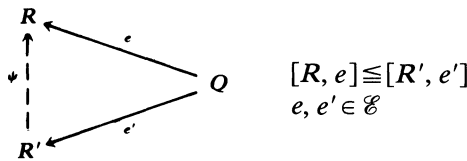
**DEFINITIONS 3.1.** Let  $Q$  be an object of  $\mathcal{K}$ . The set of all pairs  $(R, m)$  with  $m: R \rightarrow Q \in \mathcal{M}$  admits a reflexive and transitive order by defining  $(R, m) \leq (R', m')$  if there exists  $\psi$  with  $m'\psi = m$



(note that such  $\psi$  is necessarily unique and is itself in  $\mathcal{M}$ ). Thus  $(R, m) \sim (R', m')$  if  $(R, m) \leq (R', m')$  and  $(R', m') \leq (R, m)$  is an equivalence relation whose equivalence classes  $[R, m]$  are called the *subobjects* of  $Q$ .  $[R, m] \leq [R', m']$  if  $(R, m) \leq (R', m')$  is a well-defined partial order on the subobjects of  $Q$ . It is easily seen that  $[R, m] = [R', m']$  if and only if there exists an isomorphism  $\psi$  with  $m'\psi = m$ .

$Q$  is *Noetherian* if every strictly ascending chain of subobject of  $Q$  is finite. Let  $h \geq 0$  be an integer.  $Q$  has  $\mathcal{M}$ -height  $h$  if  $Q$  admits a strict chain of proper subobjects of length  $h$ , but none of length  $h + 1$ .  $Q$  has *finite  $\mathcal{M}$ -height* if  $Q$  has  $\mathcal{M}$ -height  $h$  for some  $h$ .

The dual concepts relative to  $\mathcal{K}$  are formulated by repeating the above definitions in  $\mathcal{K}^{op}$  (using 2.6). Thus, the ordering on *quotient objects* of  $Q$  is described by



(Note that we reverse the arrows, not the ordering.) We say  $Q$  is *Artinian* if  $Q$  is co-Noetherian, that is, if every strictly ascending chain of quotient objects of  $Q$  is finite. The definitions of “ $\mathcal{E}$ -height” and “*finite  $\mathcal{E}$ -height*” are clear.

**Examples 3.2.** In the category of sets, subobjects may be identified with subsets of  $Q$  and quotient objects may be identified with the canonical quotient projections induced by equivalence relations on  $Q$ . A set with  $h$  elements has  $\mathcal{M}$ -height  $h + 1$  and  $\mathcal{E}$ -height  $h$ , except that the empty set has  $\mathcal{E}$ -height 1. For sets, Noetherian = Artinian = finite. Notice that for both subsets and quotient sets, ascending chains mean increasing cardinality.

In the category of modules over a ring, the passage from a submodule  $S$  to its cokernel  $Q \setminus S$  establishes an anti-isomorphism of partially ordered sets between subobjects and quotient objects. For this reason, Artinian is equivalent to the descending chain condition on subobjects (the usual definition on module theory) and a module has finite height if and only if it is simultaneously Noetherian and Artinian. These two



properties do not hold in a general category where descending chain conditions are not equivalent to the ascending chain conditions defined above, and do not seem to be well motivated in a system context.

In Abelian groups, Noetherian = finitely-generated, whereas finite  $\mathcal{M}$ -height = finite. The group of additive integers is not Artinian. For vector spaces, on the other hand, Noetherian = Artinian = finite height and  $\mathcal{M}$ -height =  $\mathcal{E}$ -height = 1 + dimension.

We do not use the term “proper subobject” since no single usage seems consistent with all four finite-height conditions discussed above. One could exclude the proper subobject  $[Q, id_Q]$  although from the system point of view this subobject is not the trivial one; it is the *zero* subobject that is trivial from the point of view of building increasing chains. Recall that an object  $0$  is *initial* if there is a unique morphism  $0 \rightarrow Q$  to every  $Q$  and, dually, an object  $1$  is *terminal* if there is a unique morphism  $Q \rightarrow 1$  for every  $Q$ . For sets,  $0$  is the empty set,  $1$  is a one-element set and for modules  $0$  is both initial and terminal. While the unique  $0 \rightarrow Q$  is not always in  $\mathcal{M}$ , the image factorization of this map produces the least element of the partially ordered set of subobjects of  $Q$ . Dually, the image factorization of  $Q \rightarrow 1$  produces the least quotient object of  $Q$ . It seems hard to posit a natural procedure to decide when to omit the zero subobject from chains which preserves duality (i.e., the same procedure must be applied to quotient chains) and works right in the examples above.

Motivated by the sequence  $r_i: IX^i \rightarrow Q$  induced by an adjoint system, we consider an arbitrary sequence of morphisms of form  $f_i: P_i \rightarrow Q, i \in \mathcal{N}$  = the set  $\{0, 1, 2, \dots\}$  of natural numbers. The following four constructions are useful. We fix their notations for the remainder of the paper.

**3.3.** For each non-empty subset  $S$  of  $\mathcal{N}$  define

$$P_S = \coprod (P_i: i \in S),$$

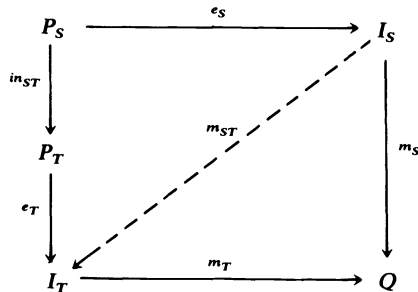
$$f_S = P_S \rightarrow Q, \text{ where } f_S in_i = f_i \quad (i \in S).$$

**3.4.** For  $S \subset T \subset \mathcal{N}, in_{ST}: P_S \rightarrow P_T$  is defined by  $in_{ST} in_i = in_i \quad (i \in S)$ .

**3.5.** Fix an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $f_S$ :

$$P_S \xrightarrow{\epsilon_S} I_S \xrightarrow{m_S} Q.$$

**3.6.** For  $S \subset T \subset \mathcal{N}, m_{ST}: I_S \rightarrow I_T$  is defined by diagonal fill-in (2.7):



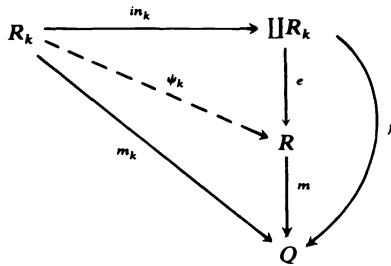
(to prove that the square commutes observe that both paths are  $f_i$  when preceded by  $in_i$ ).

We observe at once, using the results of § 2, that  $m_{ST} \in \mathcal{M}$ , that  $m_{TU}m_{ST} = m_{SU}$  for  $S \subset T \subset U$ , and that if  $f_S \in \mathcal{E}$  then  $f_T \in \mathcal{E}$  whenever  $S \subset T$ .

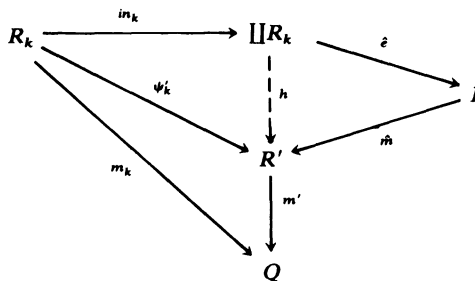
Motivated by system theory we should like to prove results such as “if  $Q$  is Noetherian and  $M$  is reachable then  $M$  is reachable in bounded time” and “if  $f_S$  is onto and if  $T$  is the subset of  $S$  obtained by deleting those  $k$  for which the union of the images of  $f_0, \dots, f_{k-1}$  is the same as the union of the images of  $f_0, \dots, f_k$  then  $f_T$  is still onto.” We observe that the reason these results are so easy to obtain in the category of sets is because the passage  $S \mapsto I_S$  is union-preserving;  $I_S$  is, after all, just the union of the images of the  $(f_s: s \in S)$ . Our approach below is to show that, in general, this passage is sufficiently supremum-preserving to lift the theory to a category. We present general results about dimension in a category in this section; and turn to their system-theoretic application in § 4.

LEMMA 3.7. *For any object  $Q$ , every nonempty countable family of subobjects of  $Q$  has a supremum.*

*Proof.* Given  $[R_k, m_k]$  define  $f$  by  $f \text{ in}_k = m_k$ , and consider



where  $(e, m)$  is an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $f$ . There exists  $\psi_k$  with  $m\psi_k = m_k$  (namely  $\psi_k = e \text{ in}_k$ ) which demonstrates that  $[R_k, m_k] \leq [R, m]$  for all  $k$ . We will show that  $[R, m]$  is the least upper bound. Suppose that  $[R_k, m_k] \leq [R', m']$ . Then there exist  $\psi'_k$

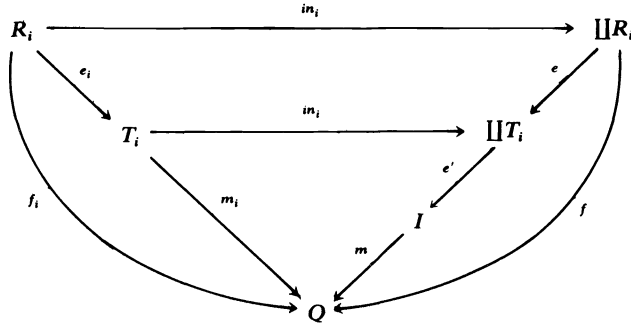


as shown and hence a unique  $h$  with  $h \text{ in}_k = \psi'_k$ . Clearly  $m'h = f$ . Hence, if  $(\hat{e}, \hat{m})$  is an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $h$ ,  $(\hat{e}, m'\hat{m})$  is an  $\mathcal{E}$ - $\mathcal{M}$  factorization of  $f$  so that  $[R, m] = [I, m'\hat{m}]$ . But then, via  $\hat{m}$ ,  $[R, m] \leq [R', m']$ .  $\square$

Given  $f: R \rightarrow Q$ , let  $[f]$  denote the subobject of  $Q$  obtained by taking the image factorization of  $f$ .

LEMMA 3.8. *Let  $f_i: R_i \rightarrow Q$  be a nonempty countable family of morphisms and let  $f: \coprod R_i \rightarrow Q$  be defined by  $f \text{ in}_i = f_i$ . Then  $[f] = \sup ([f_i])$ .*

*Proof.* Consider the diagram shown in which  $[f_i] = [T_i, m_i]$  and  $[I, m] = \sup ([f_i])$  according to the construction of Lemma 3.7.

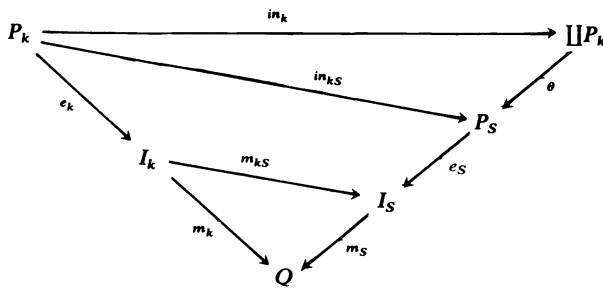


Define  $e$  by  $e in_i = in_i e_i$ . Then the diagram commutes—that is,  $f = m e' e$ —because both paths coincide with  $f_i$  when preceded by  $in_i$ . By 2.11,  $e \in \mathcal{E}$  so that  $[f] = [I, m]$  as desired.  $\square$

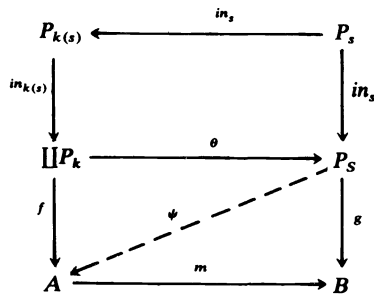
Before continuing, we introduce the abbreviation  $I_S$  for the more cumbersome  $[I_S, m_S]$ . Thus,  $I_S$  is a subobject of  $Q$ .

LEMMA 3.9. *The passage  $S \mapsto I_S$  preserves nonempty countable suprema.*

*Proof.* Let  $(S_k : k \in I)$  be a nonempty countable family of nonempty subsets of  $\mathcal{N}$  and set  $S = \cup S_k$ . We must show that  $I_S$  is the supremum of the  $I_k$  (where we use the subscript  $k$  for the more cumbersome  $S_k$  throughout). Consider the map  $\theta : \coprod P_k \rightarrow P_S$  defined by  $\theta in_k = in_{kS}$  ( $k \in I$ ). In view of the commutative diagram



it suffices to show that  $\theta \in \mathcal{E}$ , for then  $(e_S \theta, m_S)$  is precisely the construction of the supremum in Lemma 3.7. To prove that  $\theta \in \mathcal{E}$  we use the dual of 2.8. For  $s \in S$  choose  $k(s)$  with  $s \in S_{k(s)}$ . Then consider the diagram



where  $g, f$  and  $m$  are only required to satisfy  $g\theta = mf$  and  $m \in \mathcal{M}$ .

We must construct  $\psi$  with  $m\psi = g$ . Define  $\psi$  by  $\psi in_s = f in_{k(s)} in_s$  as shown. Since  $\theta in_{k(s)} in_s = in_{k(s)S} in_s = in_s$  (see 3.4) we have  $(m\psi)in_s = m f in_{k(s)} in_s = g \theta in_{k(s)} in_s = g in_s$  for all  $s \in S$ , so that  $m\psi = g$ .  $\square$

For the next definition and two propositions we consider an arbitrary nonempty-countable-supremum-preserving map  $I: R \rightarrow L$  where  $R$  is the partially ordered set of nonempty sets of  $\mathcal{N}$  and  $L$  is an arbitrary partially ordered set. For the general  $I$  we write  $I(S)$  instead of  $I_S$ .  $I_S$  is not the only application; see [7].

DEFINITION 3.10. Define  $\bar{n} = \{0, \dots, n\} \in R$ .  $I$  is stationary if  $I(\bar{n}) = I(\overline{n+1})$  for all  $n$ .  $A \in R$  is adequate if  $I(A) = I(S)$  whenever  $A \subset S$ . Equivalently,  $A$  is adequate if and only if  $I(S) = I(T)$  whenever  $A \subset S \subset T$ .

PROPOSITION 3.11. If  $A$  is one-step adequate in the sense that  $I(A) = I(A \cup \{k\})$  for all  $k$  then  $A$  is adequate.

Proof. If  $A \subset S$ ,  $S = \cup(A \cup \{k\}: k \in S)$ .  $\square$

PROPOSITION 3.12. For each nonempty-countable-supremum-preserving map  $I: R \rightarrow L$ , the set

$$A = \{0\} \cup \{k \in \mathcal{N} \mid I(\overline{k-1}) < I(\bar{k})\}$$

is adequate.

Proof. Note that if  $I$  is stationary, then  $A = \{0\}$ , and certainly this  $A$  is adequate. Otherwise, it suffices to prove  $I(A) = I(A \cup \{0, \dots, n\})$  for all  $n$ . Since  $\{0\} \subset A$ , this is certainly true for  $n = 0$ . Suppose now that  $I(A) = I(A \cup \{0, \dots, n-1\})$ . Then if  $n \in A$ , it is certainly true that  $I(A) = I(A \cup \{0, \dots, n\})$ . Otherwise,  $I(\overline{n-1}) = I(\bar{n})$ , and so

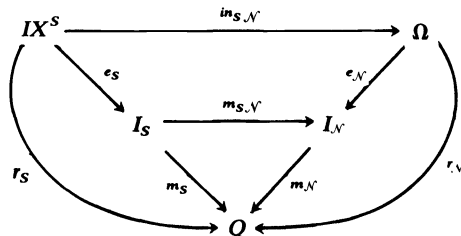
$$\begin{aligned} I(A \cup \{0, \dots, n\}) &= \sup(I(A), I(\bar{n})) \\ &= \sup(I(A), I(\overline{n-1})) \\ &= I(A \cup \overline{n-1}) = I(A). \end{aligned} \quad \square$$

COROLLARY 3.13. If  $Q$  has  $\mathcal{M}$ -height  $h$ ,  $f_i: P_i \rightarrow Q$  has an adequate set with  $h + 1$  or fewer elements.  $\square$

**4. Adequacy for systems and the simple recursion principle.** In this section, we study the implications of § 3 for an adjoint system  $M$ . We introduce all of the notions of § 3, with  $f_i: P_i \rightarrow Q = r_i: IX^i \rightarrow Q$ . We write  $IX^S$  instead of  $P_S$ .

PROPOSITION 4.1. If  $M$  is reachable and  $Q$  is Noetherian,  $M$  is reachable in bounded time. Dually, if  $M$  is observable and  $Q$  is Artinian,  $M$  is observable in bounded time.

Proof. By definition,  $r_S = r$  is the reachability of map  $M$  and  $M$  is reachable in bounded time if and only if  $r_S \in \mathcal{E}$  for some finite  $S$ . Since  $Q$  is Noetherian, there exists a finite one-step adequate set  $S$ . Then  $S$  is adequate by 3.11 and, in the diagram shown,  $m_{S,N}$  is an isomorphism.



If  $M$  is reachable,  $m_{N,N}$  is an isomorphism (2.10 and 2.9) so that  $r_S = m_{N,N} m_{S,N} e_S \in \mathcal{E}$  (by IFS1 and IFS2).

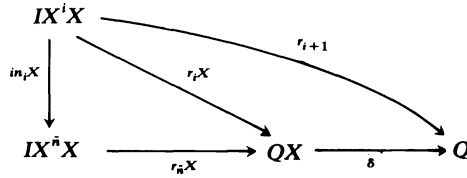
DEFINITION 4.2. Let  $V$  be a set. A sequence  $v_n$  in  $V$  is defined by simple recursion if there exists a function  $g: V \rightarrow V$  such that  $v_{n+1} = g(v_n)$ , that is,  $v_n = g^n(v_0)$ .

Example 4.3. The sequence  $\text{span}(B, AB, \dots, A^n B)$  of subspaces of the state space of a linear system is defined by simple recursion. Define  $g(S) = \text{span}(S \cup A(S))$  for each subspace  $S$ . The next result shows that this construction works for arbitrary adjoint systems.

Recall that  $\bar{n} = \{0, \dots, n\}$ .

SIMPLE RECURSION PRINCIPLE FOR ADJOINT SYSTEMS 4.4. Let  $I_{\bar{n}}$  be the subobject  $[r_{\bar{n}}]$  of the state object of an adjoint system “reachable in time  $n$ .” Then the ascending sequence  $I_{\bar{n}}$  is defined by simple recursion. Dually, the ascending sequence of observability quotient objects of the state object is also defined by simple recursion.

Proof. We define the endomorphism  $g$  on the subobjects of  $Q$  by  $g([R, m]) = \text{sup}([R, m], [\delta \cdot mX])$ . To verify that  $g(I_{\bar{n}}) = I_{\overline{n+1}}$ , recall that  $r_{i+1} = \delta \cdot r_i X$ , and that  $r_i = r_{\bar{n}} \cdot in_i$  for  $i \leq n$ , and that the vertical maps constitute a coproduct in the following diagram:



Now, by Lemma 3.8,  $[\delta \cdot r_{\bar{n}} X] = \text{sup}([r_1], \dots, [r_{n+1}]) = \text{sup}(I_1, \dots, I_{n+1})$ . But because  $X$  preserves  $\mathcal{E}$ ,

$$[\delta \cdot r_{\bar{n}} X] = [\delta \cdot m_{\bar{n}} X \cdot e_{\bar{n}} X] = [\delta \cdot m_{\bar{n}} X].$$

Moreover,  $I_{\bar{n}} = \text{sup}(I_0, \dots, I_n)$  by 3.9. Therefore

$$g(I_{\bar{n}}) = \text{sup}(\text{sup}(I_0, \dots, I_n), \text{sup}(I_1, \dots, I_{n+1})) = \text{sup}(I_0, \dots, I_{n+1}) = I_{\overline{n+1}}. \quad \square$$

An immediate consequence is a better proof of the general version [1, Thm. 4.6] of the “if you stick you’re stuck” result of [9]:

COROLLARY 4.5. If  $I_{\bar{n}} = I_{\overline{n+1}}$ , then  $\bar{n}$  is adequate.

Proof.  $I_{\overline{n+k+1}} = g(I_{\overline{n+k}}) = g(I_{\overline{n+k-1}})$  (induction hypothesis)  $= I_{\overline{n+k}}$ .  $\square$

While the proof of Corollary 4.5 bypasses Proposition 3.12, the latter is still a useful principle, as we shall show in [7].

COROLLARY 4.6. Let  $M$  be an adjoint system with state object  $Q$ . If  $Q$  has  $\mathcal{M}$ -height  $h$  then  $M$  is reachable in time  $h$ . Dually, if  $Q$  has  $\mathcal{E}$ -height  $h$ ,  $M$  is observable in time  $h$ .  $\square$

To tie this back to the realization theory of § 2, and especially the Partial Realization Theorem 2.19, we make the

Observation 4.7. It is clear from 2.18 and 2.10 that  $t$  is in  $\mathcal{E}$  and that  $u$  is in  $\mathcal{M}$ . It is then clear that if  $\mathcal{E}$ -height  $(\bar{Q}) = \mathcal{E}$ -height  $(R)$  and both are finite, then  $t$  is eventually an isomorphism; while if  $\mathcal{M}$ -height  $(R) = \mathcal{M}$ -height  $(\hat{Q})$  and both are finite, then  $u$  is eventually an isomorphism. Thus the condition “ $t$  and  $u$  are isomorphisms” in 2.19 may be replaced by “ $\mathcal{E}$ -height  $(\bar{Q}) = \mathcal{E}$ -height  $(R)$  and  $\mathcal{M}$ -height  $(R) = \mathcal{M}$ -height  $(\hat{Q})$  and both are finite.”

COROLLARY 4.8. For adjoint processes in **Vect**, we may obtain a partial realization as soon as

$$\dim(Q) = \dim(R) = \dim(\hat{Q}) = \text{finite}.$$

This yields both Tether’s [23] criterion for partial realization of linear systems, and Isidori’s [16] criterion for partial realization of bilinear systems (internal sense).

Given a matrix  $A: \mathbf{R}^m \rightarrow \mathbf{R}^n$ , one way to define its *rank* is simply as the *dimension* of the image (recall 2.4)  $\text{Im}(A) = A(\mathbf{R}^m)$ . We have seen that in **Vect** we have that  $\mathcal{E}\text{-height}(Q) = \mathcal{M}\text{-height}(Q) = \dim(Q)$  for any finite-dimensional vector space  $Q$ . This suggests the following:

**DEFINITION 4.9.** Let  $f: Q \rightarrow R$  be a morphism of  $\mathcal{K}$ . The *rank* of  $f$  is  $(h, l)$  if  $\text{Im}(f)$  has finite  $\mathcal{E}$ -height  $h$  and finite  $\mathcal{M}$ -height  $l$ , and is undefined if no such finite numbers exist.

Now recall that in 2.3 we associated with each Hankel matrix  $H_i^j$  its Hankel dynamorphism  $H: (\Omega, \mu_0) \rightarrow (\Gamma, L)$  satisfying  $pr_j H in_i = H_i^j$ , where  $\Omega = \coprod (IX^i \mid i \geq 0)$  is the object of inputs and  $\Gamma = \coprod (YZ^j \mid j \geq 0)$  is the observability space. In the usual linear case,  $H$  is precisely the infinite Hankel matrix whose blocks are the  $H_i^j = CA^{i+j}B$ .

**DEFINITION 4.10.** We say that the Hankel matrix  $H_i^j$  has rank  $(h, l)$  just in case its Hankel dynamorphism has rank  $(h, l)$ .

Combining the argument for Corollary 4.6 with the Partial Realization Theorem 2.19 and observation 4.7 we have

**THE HANKEL REALIZATION THEOREM 4.11.** Let  $H_i^j$  be a Hankel matrix with rank  $(h, l)$ . Then the canonical realization of  $H_i^j$  may be constructed by applying the construction of 2.19 with  $k = h$  and  $n = l$  in 2.18.

*Proof outline.* The crucial point is that the rank condition implies that items (i) and (ii) of the proof of 2.19— $r_i = \bar{e} \cdot in_i$  and  $\sigma_j = pr_j \cdot \bar{m}$ —hold for all  $i$  and  $j$  respectively. But this not only shows that the  $M$  of 2.19 has Hankel matrix  $H_i^j$ , but also that  $M$  has reachability map  $\bar{e}$  in  $\mathcal{E}$  and observability map  $\bar{m}$  in  $\mathcal{M}$ —so that  $M$  is canonical.  $\square$

*Observation 4.12.* As in 4.8, we note that when  $\mathcal{K} = \mathbf{Vect}$  the two height conditions in 4.9 collapse to the single condition “ $Q$  having finite dimension  $h$ ,” and we may then take  $k = n = h$  in forming the realization.

For the biadequacy criterion for Hankel realization, see [7, Prop. 8.6].

We close this section with a few brief remarks on duality as it relates to Hankel matrices. Recall from 1.4 that an adjoint system

$$M = (Q, \delta, I, \tau, Y, \beta)$$

is given by the input map  $\tau: I \rightarrow Q$ , dynamics  $\delta: QX \rightarrow Q$  and output map  $\beta: Q \rightarrow Y$ , and that we may associate with  $M$  its codynamics  $\Delta: Q \rightarrow QZ$  which corresponds to  $\delta$  under adjointness.

Let us use  $f: R \leftarrow Q$  for the  $\mathcal{K}$ -morphism  $f: Q \rightarrow R$  interpreted as  $\mathcal{K}^{op}$ -morphism. We may associate with  $M$  its *dual* [4, Definition 4.6]

$$M^{op} = (Q, \Delta, Y, \beta, I, \tau)$$

in the category  $\mathcal{K}^{op}$ , given by the input map  $\beta: Y \leftarrow Q$ , dynamics  $\Delta: QZ \leftarrow Q$  and output map  $\tau: Q \leftarrow I$ .

It is then easy to verify that the Hankel matrix  $K_i^j: YZ^j \leftarrow IX^i$  for  $M^{op}$  is just the  $H_i^j: IX^i \rightarrow YZ^j$  for  $M$  in  $\mathcal{K}$ , so we have that  $H: \Gamma \leftarrow \Omega$  corresponds to the  $K_i^j$  in  $\mathcal{K}^{op}$  in the fashion specified by 2.3.

Now we saw in 2.6 that  $(\mathcal{E}, \mathcal{M})$  is an image factorization system in the category  $\mathcal{K}$  iff  $(\mathcal{M}, \mathcal{E})$  is an image factorization system in the opposite category  $\mathcal{K}^{op}$ . Thus a realization  $M$  of a Hankel matrix in  $\mathcal{K}$  corresponds precisely to a realization  $M^{op}$  of the “reversed” Hankel matrix in  $\mathcal{K}^{op}$ . In [4, Example 4.12] we show how the usual duality results of linear system theory follow because the *finite-dimensional part* of  $\mathbf{Vect}^{op}$  can be modeled in **Vect**, with  $B^{op}$  for  $B: \mathbf{R}^m \rightarrow \mathbf{R}^n$  being modeled as its transpose  $B^T: \mathbf{R}^n \rightarrow \mathbf{R}^m$ .

Thus the linear system  $M = (A: Q \rightarrow Q, B: I \rightarrow Q, C: Q \rightarrow Y)$  has opposite system  $M^{op} = (A: Q \leftarrow Q, C: Y \leftarrow Q, B: Q \leftarrow I)$  modeled as the familiar dual  $M^T = (A^T: Q \rightarrow Q, C^T: Y \rightarrow Q, B^T: Q \rightarrow I)$ . Another perspective, based on the imposition of suitable topologies on linear vector spaces, is given in [14]. An interesting open problem is to explore conditions on a ring  $R$  for such results to extend from **Vect** to  $R\text{-Mod}$ , the category of modules over  $R$ .

**Acknowledgment.** This paper is the third reference promised in [6]. The authors gratefully acknowledge T. J. Tarn for conversations which initiated the investigations that led to this paper.

#### REFERENCES

- [1] B. D. O. ANDERSON, M. A. ARBIB AND E. G. MANES, *Foundations of system theory: Finitary and infinitary conditions*, Lecture Notes in Economics and Mathematical Systems, 115, Springer-Verlag, New York, 1976.
- [2] M. A. ARBIB AND E. G. MANES, *Foundations of system theory: decomposable systems*, *Automatica*, 10 (1974), pp. 285–302.
- [3] ———, *Arrows, Structures and Functors: The Categorical Imperative*, Academic Press, New York, 1975.
- [4] ———, *Adjoint machines, state-behavior machines and duality*, *J. Pure Appl. Algebra*, 6 (1975), pp. 313–344.
- [5] ———, *Time-varying systems*, this Journal, 13 (1975), pp. 1252–1270.
- [6] ———, *Recurrence and finiteness for systems in a general setting*, *Proc. IEEE Conf. Decision and Control* 76CH 1150-2CS, 1976, pp. 1180–1183.
- [7] ———, *Foundations of system theory: the Hankel matrix*, *J. Comput. System Sci.*, to appear.
- [8] ———, *On the evolution of generalized Hankel matrices*, to appear.
- [9] M. A. ARBIB AND H. P. ZEIGER, *On the relevance of abstract algebra to control theory*, *Automatica*, 5 (1969), pp. 589–606.
- [10] L. BUDACH AND H. -J. HOEHNKE, *Automaten und Funktoren*, VEB (1975).
- [11] H. EHRIG, K.-D. KIERMEIER, H.-J. KREOWSKI AND W. KÜHNEL, *Universal Theory of Automata*, Teubner, Leipzig, 1974.
- [12] M. FLIESS, *Matrices de Hankel*, *J. Math. Pure Appl.*, 53 (1974), pp. 197–224.
- [13] J. A. GOGUEN, *Minimal realizations of machines in closed categories*, *Bull. Amer. Math. Soc.*, 78 (1972), pp. 777–783.
- [14] S. J. HEGNER, *Duality theory for discrete-time linear systems*, *J. Comp. System Sci.*, 17 (1978), pp. 116–143.
- [15] J. R. ISBELL, *Some remarks concerning categories and subspaces*, *Canad. J. Math.*, 9 (1957), pp. 563–577.
- [16] A. ISIDORI, *Direct construction of minimal bilinear realizations from nonlinear input-output maps*, *IEEE Trans. Automatic Control*, AC-18 (1973), pp. 623–631.
- [17] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [18] D. M. KAN, *Adjoint functors*, *Trans. Amer. Math. Soc.*, 87 (1958), pp. 294–329.
- [19] S. MAC LANE, *Groups, categories and duality*, *Proc. Nat. Acad. Sci. U.S.A.*, 34 (1948), pp. 263–267.
- [20] ———, *Categories for the Working Mathematician*, Springer-Verlag, New York, 1971.
- [21] E. MANES, *Algebraic Theories*, Springer-Verlag, New York, 1976.
- [22] T. J. TARN AND S. NONOYAMA, *Realization of discrete-time internally bilinear systems*, *Proc. IEEE Conf. Decision and Control* 76CH 1150-2CS, (1976), pp. 125–133.
- [23] A. J. TETHER, *Construction of linear state-variable models from finite input-output data*, *IEEE Trans. Automatic Control*, AC-15 (1970), pp. 427–436.

## A QUADRATIC TRANSFORMATION OF A BASIC HYPERGEOMETRIC SERIES\*

ARUN VERMA†

**Abstract.** A quadratic transformation for a basic hypergeometric series is obtained.

**1. Introduction.** Hypergeometric series have been studied and used for almost two hundred years. One of the useful facts about some hypergeometric series is that they have a quadratic transformation. In fact, the general Legendre function is just an algebraic function times a hypergeometric function that satisfies a condition for the existence of a quadratic transformation. Hypergeometric series have been generalized to basic hypergeometric series. However basic hypergeometric series have not been studied as extensively. In particular, there only seems to be one quadratic transformation of a basic hypergeometric series in point. This transformation was found by Carlitz [3]. Before stating Carlitz's identity we need some notation.

$$(1.1) \quad \begin{aligned} [a; q]_n &= (1-a)(1-aq) \cdots (1-aq^{n-1}), & n = 1, 2, \dots, \\ &= 1, & n = 0. \end{aligned}$$

If  $|q| < 1$  then

$$(1.2) \quad [a; q]_\infty = \prod_{k=0}^{\infty} (1-aq^k).$$

Set

$$(1.3) \quad {}_{r+1}\Phi_r \left[ \begin{matrix} a_1, \dots, a_{r+1}; q, t \\ b_1, \dots, b_r \end{matrix} \right] = \sum_{k=0}^{\infty} \frac{[a_1; q]_k \cdots [a_{r+1}; q]_k}{[b_1; q]_k \cdots [b_r; q]_k} \frac{t^k}{[q; q]_k}.$$

Carlitz [3] proved

$$(1.4) \quad \begin{aligned} {}_3\Phi_2 \left[ \begin{matrix} a, b, c; q, \frac{aqx}{bc} \\ \frac{aq}{b}, \frac{aq}{c} \end{matrix} \right] &= \frac{[ax; q]_\infty}{[x; q]_\infty} \sum_{n=0}^{\infty} \frac{\left[ \frac{aq}{bc}; q \right]_n [a; q]_{2n} q^n}{[q; q]_n \left[ \frac{aq}{b}; q \right]_n \left[ \frac{aq}{c}; q \right]_n [ax; q]_n \left[ \frac{q}{x}; q \right]_n} \\ &= \frac{[ax; q]_\infty}{[x; q]_\infty} {}_5\Phi_4 \left[ \begin{matrix} \frac{aq}{bc}, a^{1/2}, -a^{1/2}, (aq)^{1/2}, -(aq)^{1/2}; q, q \\ \frac{aq}{b}, \frac{aq}{c}, ax, \frac{q}{x} \end{matrix} \right], \end{aligned}$$

when  $a = q^{-k}$ ,  $k = 0, 1, \dots$ . He does not state the restriction that  $a = q^{-k}$  which terminates both series; but without it his proof does not work, for one of his series diverges.

\* Received by the editors September 19, 1978, and in revised form July 10, 1979.

† Department of Mathematics, Roorkee University, Roorkee, India.



**2. Another quadratic transformation.** Andrews [1] found a  $q$ -extension of a terminating form of an identity of Watson. It is

$$(2.1) \quad {}_4\Phi_3 \left[ \begin{matrix} a, b, c^{1/2}, -c^{1/2}; q, q \\ (abq)^{1/2}, -(abq)^{1/2}, c \end{matrix} \right] = \frac{a^{n/2} [aq; q^2]_{\infty} [bq; q^2]_{\infty} \left[ \frac{cq}{a}; q^2 \right]_{\infty} \left[ \frac{cq}{b}; q^2 \right]_{\infty}}{[q; q^2]_{\infty} [abq; q^2]_{\infty} [cq; q^2]_{\infty} \left[ \frac{cq}{ab}; q^2 \right]_{\infty}}, \quad b = q^{-n}.$$

This can be rewritten as

$$(2.2) \quad {}_4\Phi_3 \left[ \begin{matrix} q^{-2n}, a^2 q^{2n}, \frac{aq^{1/2}}{b}, -\frac{a}{b} q^{1/2}; q, q \\ aq^{1/2}, -aq^{1/2}, \frac{a^2 q}{b^2} \end{matrix} \right] = \left( \frac{aq^{1/2}}{b} \right)^{2n} \frac{[q; q^2]_n [b^2; q^2]_n}{[a^2 q; q^2]_n \left[ \frac{a^2}{b^2} q^2; q^2 \right]_n}$$

when  $b = q^{-2n}$ . When  $b = q^{-2n-1}$  then the series in (2.1) vanishes. Use these results as follows.

$$(2.3) \quad {}_2\Phi_1 \left[ \begin{matrix} a^2, b^2; q^2, y^2 \\ \frac{a^2}{b^2} q^2 \end{matrix} \right] = \sum_{n=0}^{\infty} \frac{[a^2; q]_{2n}}{[q; q]_{2n}} y^{2n} \frac{[q; q^2]_n [b^2; q^2]_n}{[a^2 q; q^2]_n \left[ \frac{a^2}{b^2} q^2; q^2 \right]_n}$$

$$= \sum_{n, k \geq 0} \frac{[a^2; q]_n}{[q; q]_n} y^n \left( \frac{b}{aq^{1/2}} \right)^n \frac{[q^{-n}; q]_k [a^2 q^n; q]_k \left[ \frac{a^2}{b^2} q; q^2 \right]_k q^k}{[a^2 q; q^2]_k \left[ \frac{a^2}{b^2} q; q \right]_k [q; q]_k}.$$

Here (2.2) is used when  $n$  is even and when  $n$  is odd the sum vanishes. To simplify this series we will need the  $q$ -binomial theorem.

$$(2.4) \quad \sum_{n=0}^{\infty} \frac{[a; q]_n}{[q; q]_n} x^n = \frac{[ax; q]_{\infty}}{[x; q]_{\infty}}.$$

This result is true if  $|x| < 1$  or for all  $x$  when  $a = q^{-j}$  for some  $j, j = 0, 1, 2, \dots$ . To simplify (2.3) we will assume  $a = q^{-j}$ . Then

$$(2.5) \quad {}_2\Phi_1 \left[ \begin{matrix} a^2, b^2; q^2, y^2 \\ \frac{a^2}{b^2} q^2 \end{matrix} \right] = \sum_{k=0}^{\infty} \frac{\left[ \frac{a^2}{b^2} q; q^2 \right]_k [a^2; q^2]_k}{\left[ \frac{a^2}{b^2} q; q \right]_k [q; q]_k} q^k \sum_{n=0}^{\infty} \frac{[q^{-n-k}; q]_k [a^2 q^{2k}; q]_n}{[q; q]_{n+k}} \left( \frac{by}{aq^{1/2}} \right)^{n+k}$$

$$= \sum_{k=0}^{\infty} \frac{\left[ \frac{a^2}{b^2} q; q^2 \right]_k [a^2; q^2]_k}{\left[ \frac{a^2}{b^2} q; q \right]_k [q; q]_k} \left( -\frac{by}{a} \right)^k q^{-(k^2/2)} \sum_{n=0}^{\infty} \frac{[a^2 q^{2k}; q]_n}{[q; q]_n} \left( \frac{by}{aq^{k+1/2}} \right)^n$$

$$= \sum_{k=0}^{\infty} \frac{\left[ \frac{a^2}{b^2} q; q^2 \right]_k [a^2; q^2]_k}{\left[ \frac{a^2}{b^2} q; q \right]_k [q; q]_k} \left( -\frac{by}{a} \right)^k q^{-(k^2/2)} \frac{[abyq^{k-(1/2)}; q]_{\infty}}{[a^{-1}byq^{-k-(1/2)}; q]_{\infty}}$$

$$\begin{aligned}
 &= \frac{[abyq^{-1/2}; q]_\infty}{[a^{-1}byq^{-(1/2)}; q]_\infty} \sum_{k=0}^{\infty} \frac{[a^2; q^2]_k \left[ \frac{a^2}{b^2} q; q^2 \right]_k q^k}{\left[ \frac{a^2}{b^2} q; q \right]_k [q; q]_k [abyq^{-(1/2)}; q]_k} \\
 &= \frac{[abyq^{-1/2}; q]_\infty}{[a^{-1}byq^{-(1/2)}; q]_\infty} {}_4\Phi_3 \left[ \begin{matrix} a, -a, \frac{a}{b} q^{1/2}, -\frac{a}{b} q^{1/2} \\ \frac{a^2}{b^2} q, abyq^{-(1/2)}, \frac{a}{by} q^{3/2} \end{matrix}; q, q \right], \quad a = q^{-j}.
 \end{aligned}$$

To see what this identity generalizes set  $a = q^\alpha, b = q^\beta$  and let  $q \rightarrow 1^-$ . The result is

$$(2.6) \quad {}_2F_1 \left[ \begin{matrix} \alpha, \beta; y^2 \\ 1 + \alpha - \beta \end{matrix} \right] = (1 - y)^{-2\alpha} {}_2F_1 \left[ \begin{matrix} \alpha, \alpha - \beta + \frac{1}{2}; -\frac{4y}{(1 - y)^2} \\ 2\alpha - 2\beta + 1 \end{matrix} \right].$$

Formula (2.6) holds even when  $\alpha$  is not a negative integer.

REFERENCES

[1] G. E. ANDREWS, *On q-analogues of the Watson and Whipple summations*, this Journal, 7 (1976), pp. 332-336.  
 [2] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London, 1935, reprinted Hafner, New York, 1964.  
 [3] L. CARLITZ, *Some formulas of F. H. Jackson*, Monatsh. Math., 73 (1969), pp. 193-198.

# THE LAPLACE TRANSFORM OF A PRODUCT OF BESSEL FUNCTIONS\*

B. C. CARLSON†

**Abstract.** The Laplace transform of  $J_\mu(at)J_\nu(bt)t^\lambda$  is an  $R$ -function if  $\lambda = \mu - \nu$  or  $\lambda = \mu - \nu + 1$ . It is a finite sum of  $R$ -functions if  $\mu = \nu$  and  $\lambda - 2\mu$  is a nonnegative integer, or if both  $\lambda + \mu - \nu$  and  $\lambda + \nu - \mu$  are nonnegative integers. The last result is proved by the techniques of double Dirichlet averages. If  $2\lambda$ ,  $2\mu$ , and  $2\nu$  are integers, rules are given for determining by inspection whether the  $R$ -functions are rational, algebraic, elementary transcendental, or elliptic. When  $\mu$  and  $\nu$  are nonnegative integers and  $\lambda$  is an integer, the Laplace transform is a complete elliptic integral of the first or second kind if  $\lambda \geq |\mu - \nu|$  and is conjectured to be a complete elliptic integral of the third kind if  $\lambda < |\mu - \nu|$ .

## I. Introduction. Integrals of the type

$$(1.1) \quad I(\mu, \nu; \lambda) = \int_0^\infty e^{-pt} J_\mu(at) J_\nu(bt) t^\lambda dt$$

occur in electromagnetism, gravitational potential theory, heat conduction, hydrodynamics, and elasticity, typically in connection with axially symmetric systems. Five examples are cited with references in [10, p. 530]. The fourth is the potential of a uniform mass distribution on a circular disk (see also [12], [19]), which is proportional to  $I(0, 1; -1)$ ; it is currently of interest in calculating the effect of Saturn's rings on a passing spacecraft [14]. The gravitational field of a right circular cylinder [12], [20] is used in interpreting gravity data, and the magnetic field of a right circular cylinder [21] or a circular disk [22] is used in interpreting geomagnetic anomalies. The potential energy of interaction between two coaxial circular disks [4] is proportional to  $I(1, 1; -2)$ . Other examples are met in signal statistics [17] and in calculating the magnetic field of axially symmetric coils [13], [24]. Integrals in which one Bessel function is spherical occur in elasticity [23, pp. 459-468], [27]. Both may be spherical in hydrodynamics [18] and quantum-mechanical collision theory [9].

Watson [25, pp. 389-391] summarizes results of Gegenbauer and others for the cases  $\mu = \nu$ ,  $\lambda = \mu - \nu$ , and  $\lambda = \mu + \nu$ . Eason, Noble, and Sneddon [10] transform the integral, deduce recurrence relations, and analyze and tabulate numerically eleven cases with integral values of  $\lambda$ ,  $\mu$ ,  $\nu$ . Luke [15, pp. 314-320] cites further numerical tables and gives additional results, notably representations by series. Benton [3] considers the case  $\lambda = \mu - \nu$ . Erdélyi [11, vol. 1, pp. 182-184, 196] lists integrals with various restrictions on the parameters, and Okui [17] expresses  $I(\mu, \nu; \lambda)$  in terms of complete elliptic integrals for many numerical values of  $\mu$ ,  $\nu$ ,  $\lambda$ . Cases in which both Bessel functions are spherical are evaluated by Detrich and Conn [9] in terms of elementary functions.

Since  $J_{-m} = (-1)^m J_m$  if  $m$  is an integer, we assume throughout that  $\mu, \nu \neq -1, -2, -3, \dots$ . Since the case  $p = 0$  (the possibly discontinuous Weber-Schafheitlin integral) is discussed thoroughly by Watson [25, pp. 398-410], we assume further that  $p \neq 0$ . Then the integral converges at the upper limit of integration if  $\text{Re } p > |\text{Im } a| + |\text{Im } b|$  and at the lower limit if  $\text{Re}(\lambda + \mu + \nu) > -1$ . Expansion of the Bessel

\* Received by the editors January 11, 1979, and in revised form July 16, 1979.

† Ames Laboratory, United States Department of Energy and Departments of Mathematics and Physics, Iowa State University, Ames, Iowa 50011. This research was supported in part by the Department of Energy under Contract W-7405-Eng-82.

functions in  ${}_0F_1$ -series leads to [11, 4.16(13)],

$$(1.2) \quad I(\mu, \nu; \lambda) = C(\mu, \nu) \Gamma(\lambda + \mu + \nu + 1) p^{-\lambda - \mu - \nu - 1} \\ \cdot F_4\left(\frac{\lambda + \mu + \nu + 1}{2}, \frac{\lambda + \mu + \nu + 2}{2}; \mu + 1, \nu + 1; \frac{-a^2}{p^2}, \frac{-b^2}{p^2}\right),$$

where

$$(1.3) \quad C(\mu, \nu) = \frac{(a/2)^\mu (b/2)^\nu}{\Gamma(\mu + 1) \Gamma(\nu + 1)}.$$

The series representation [11, vol. 1, p. 384] of Appell's function  $F_4$  converges if  $|p| > |a| + |b|$ . Hence (1.2) is directly useful only for sufficiently large  $|p|$ . Expansion of the Bessel functions in  ${}_1F_1$ -series leads to [11, 4.14(24)],

$$(1.4) \quad I(\mu, \nu; \lambda) = C(\mu, \nu) \Gamma(\lambda + \mu + \nu + 1) (p + ia + ib)^{-\lambda - \mu - \nu - 1} \\ \cdot F_2\left(\lambda + \mu + \nu + 1, \mu + \frac{1}{2}, \nu + \frac{1}{2}; 2\mu + 1, 2\nu + 1; \frac{2ia}{p + ia + ib}, \frac{2ib}{p + ia + ib}\right).$$

Unfortunately the arguments are complex if  $p, a, b$  are real. Equations (1.2) and (1.4) are related by a known quadratic transformation of a restricted  $F_4$  [2, (3.1)], [6, (4.4)].

In this paper we note several cases in which restrictions on  $\lambda, \mu, \nu$  allow  $F_4$  or  $F_2$  to be expressed in terms of  $R$ -functions of two or three variables. The  $R$ -function of two variables is a variant of the hypergeometric function  ${}_2F_1$  [8, (5.9–11)]:

$$(1.5) \quad R_{-\alpha}(\beta_1, \beta_2; z_1, z_2) = z_2^{-\alpha} {}_2F_1(\alpha, \beta_1; \beta_1 + \beta_2; 1 - z_1/z_2) \\ = z_2^{-\alpha} \sum_{m=0}^{\infty} \frac{(\alpha)_m (\beta_1)_m}{(\beta_1 + \beta_2)_m m!} (1 - z_1/z_2)^m,$$

where  $(\alpha)_m$  is Pochhammer's symbol, the  $R$ -function is symmetric in the indices 1 and 2, and the series converges if  $|1 - z_1/z_2| < 1$ . The  $R$ -function of three variables is a variant of Appell's  $F_1$  [8, Ex. 6.3–5]:

$$(1.6) \quad R_{-\alpha}(\beta_1, \beta_2, \beta_3; z_1, z_2, z_3) \\ = z_3^{-\alpha} F_1(\alpha, \beta_1, \beta_2; \beta_1 + \beta_2 + \beta_3; 1 - z_1/z_3, 1 - z_2/z_3) \\ = z_3^{-\alpha} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n} (\beta_1)_m (\beta_2)_n}{(\beta_1 + \beta_2 + \beta_3)_{m+n} m! n!} (1 - z_1/z_3)^m (1 - z_2/z_3)^n,$$

where the  $R$ -function is symmetric in the indices 1, 2, 3, and the series converges if  $|1 - z_1/z_3| < 1$  and  $|1 - z_2/z_3| < 1$ .

The use of  $R$  has three advantages in this context. First, if  $2\lambda, 2\mu, 2\nu$  are given integers, it is easy to tell by inspection whether the  $R$ -function is rational, algebraic, elementary transcendental, or elliptic (see § 2). Second, if the  $R$ -function is an elliptic integral, a table and other aids for reduction to standard integrals are available [8, § 9.3]. Third, if  $p, a, b$  are positive or if  $p$  is positive and  $a, b$  are purely imaginary, the  $R$ -function can be expanded in a convergent power series without searching for a suitable linear transformation of  ${}_2F_1$  or  $F_1$ . For example, if  $p, a, b$  are positive we may use the symmetry of  $R$  to choose  $z_3 = A^2$  in (2.3), but if the Bessel functions are replaced by modified Bessel functions ( $a$  and  $b$  purely imaginary), we choose  $z_3 =$

$A^2 - b^2$ . In each case  $z_3$  is then the largest of three positive arguments, and the series in (1.6) converges.

**2. List of results.** We assume throughout that  $\mu, \nu \neq -1, -2, -3, \dots$  and  $\operatorname{Re} p > |\operatorname{Im} a| + |\operatorname{Im} b|$ , which implies that  $p \pm ia \pm ib$  lies in the open right half-plane for all four choices of the signs. We define

$$(2.1) \quad r_{\pm}^2 = p^2 + (a \pm b)^2, \quad A = \frac{1}{2}(r_+ + r_-), \quad G^2 = r_+ r_-,$$

where  $A$  and  $G$  denote arithmetic and geometric means. Since  $r_+^2 = (p + ia + ib) \cdot (p - ia - ib)$  we may choose  $r_+$  (and similarly  $r_-$  and  $G$ ) in the open right half-plane. Some useful identities are

$$(2.2) \quad \begin{aligned} 2A^2 - G^2 &= \frac{1}{2}(r_+^2 + r_-^2) = p^2 + a^2 + b^2, & A(r_+ - r_-) &= \frac{1}{2}(r_+^2 - r_-^2) = 2ab, \\ A^2 - G^2 &= \left(\frac{r_+ - r_-}{2}\right)^2 = \left(\frac{ab}{A}\right)^2, & (A^2 - a^2)(A^2 - b^2) &= p^2 A^2, \\ A^2 - a^2 - p^2 &= b^2 \left(1 - \frac{a^2}{A^2}\right), & A^2 - b^2 - p^2 &= a^2 \left(1 - \frac{b^2}{A^2}\right). \end{aligned}$$

We define  $I(\mu, \nu; \lambda)$  and  $C(\mu, \nu)$  by (1.1) and (1.3). (To replace  $J_{\mu}(at)$  in (1.1) by a modified Bessel function  $I_{\mu}(at)$ , leave  $C(\mu, \nu)$  unchanged and replace  $a$  by  $ia$  everywhere else; similarly for  $J_{\nu}(bt)$ .) In § 3 it is shown that

$$(2.3) \quad \begin{aligned} I(\mu, \nu; \mu - \nu) &= C(\mu, \nu)\Gamma(2\mu + 1) \\ &\cdot R_{-\mu-1/2}(\nu + \frac{1}{2}, \mu - \nu + \frac{1}{2}, \nu - \mu; A^2, G^2, A^2 - b^2), \quad \operatorname{Re} \mu > -\frac{1}{2}, \\ I(\mu, \nu; \mu - \nu + 1) &= C(\mu, \nu)\Gamma(2\mu + 2)p \\ (2.4) \quad &\cdot R_{-\mu-3/2}(\nu - \frac{1}{2}, \mu - \nu + \frac{3}{2}, \nu - \mu; A^2, G^2, A^2 - b^2), \end{aligned}$$

$\operatorname{Re} \mu > -1.$

Similar formulas for  $I(\mu, \nu; \nu - \mu)$  and  $I(\mu, \nu; \nu - \mu + 1)$  are obtained by interchanging  $\mu$  with  $\nu$  and  $a$  with  $b$  (which does not change  $A$  or  $G$ ). These four formulas include ten of the eleven integrals tabulated in [10].

In  $n$  is a nonnegative integer, we find also

$$(2.5) \quad \begin{aligned} I(\mu, \mu; n) &= C(\mu, \mu)\Gamma(2\mu + 1) \sum_{m=0}^{[n/2]} \frac{(-1)^m}{m!} (-n)_{2m} (\mu + \frac{1}{2})_{n-m} (2p)^{n-2m} \\ &\cdot R_{-\mu-(1/2)-n+m}(\mu + \frac{1}{2}, \mu + \frac{1}{2}; r_+^2, r_-^2), \quad \operatorname{Re} \mu > -(n+1)/2, \\ I(\mu, \mu; 2\mu + n) &= C(\mu, \mu)\Gamma(4\mu + 1) \sum_{m=0}^{[n/2]} \frac{(-1)^m}{m!} (-n)_{2m} (2\mu + \frac{1}{2})_{n-m} \\ (2.6) \quad &\cdot (2p)^{n-2m} R_{-2\mu-(1/2)-n+m}(\mu + \frac{1}{2}, \mu + \frac{1}{2}, -\mu; r_+^2, r_-^2, p^2), \end{aligned}$$

$\operatorname{Re} \mu > -(n+1)/4.$

Here  $[n/2]$  is the largest integer not exceeding  $n/2$ . In (2.5) the  $R$ -functions have equal parameters and hence are Legendre functions [8, pp. 158–159]. If  $\mu$  is a nonnegative integer, the  $R$ -functions in (2.6) can be expressed in terms of Legendre functions by successive applications of [8, (5.9–8)].

A product of Legendre functions occurs in the previously known case  $\lambda = -\frac{1}{2}$  [11, 4.16(12)]. We give it here in several forms that may prove convenient in different

circumstances:

$$(2.7) \quad \begin{aligned} I(\mu, \nu; -\frac{1}{2}) &= C(\mu, \nu)\Gamma(\mu + \nu + \frac{1}{2})A^{\mu+\nu+1/2}f(\mu, \nu; a, b)f(\nu, \mu; b, a) \\ &= C(\mu, \nu)\Gamma(\mu + \nu + \frac{1}{2})p^{\mu+\nu+1/2}g(\mu, \nu; a, b)g(\nu, \mu; b, a), \end{aligned}$$

where  $\text{Re}(\mu + \nu + \frac{1}{2}) > 0$ ;

$$(2.8) \quad \begin{aligned} f(\mu, \nu; a, b) &= R_{-(2\mu+2\nu+1)/4}\left(\frac{2\mu-2\nu+1}{4}, \frac{2\mu+2\nu+3}{4}; A^2-a^2, A^2\right) \\ &= R_{-\mu-\nu-1/2}[\mu + \frac{1}{2}, \mu + \frac{1}{2}; (A^2-a^2)^{1/2} + ia, (A^2-a^2)^{1/2} - ia] \\ &= R_{-\mu-\nu-1/2}\left[\mu - \nu + \frac{1}{2}, \nu + \frac{1}{2}; \frac{A + (A^2-a^2)^{1/2}}{2}, A\right]; \end{aligned}$$

$$(2.9) \quad \begin{aligned} g(\mu, \nu; a, b) &= R_{-(2\mu+2\nu+1)/4}\left(\frac{2\mu-2\nu+1}{4}, \frac{2\mu+2\nu+3}{4}; p^2, A^2-b^2\right) \\ &= R_{-\mu-\nu-1/2}[\mu + \frac{1}{2}, \mu + \frac{1}{2}; p + (p^2-A^2+b^2)^{1/2}, p - (p^2-A^2+b^2)^{1/2}] \\ &= R_{-\mu-\nu-1/2}\left[\mu - \nu + \frac{1}{2}, \nu + \frac{1}{2}; \frac{p + (A^2-b^2)^{1/2}}{2}, (A^2-b^2)^{1/2}\right]. \end{aligned}$$

If  $\lambda + \mu - \nu$  and  $\lambda + \nu - \mu$  are nonnegative integers, it is shown in § 4 that  $I(\mu, \nu; \lambda)$  is a finite sum of  $R$ -functions. Define

$$(2.10) \quad \begin{aligned} m &= \lambda + \nu - \mu, & n &= \lambda + \mu - \nu, & (m, n &= 0, 1, 2, \dots); \\ x &= p + ia + ib, & y &= p + ia - ib, & z &= p - ia + ib, & w &= p - ia - ib. \end{aligned}$$

Some useful identities are

$$(2.11) \quad \begin{aligned} xw &= r_+^2, & yz &= r_-^2, & (xy)^{1/2} + (wz)^{1/2} &= 2(A^2-a^2)^{1/2}, \\ (xz)^{1/2} + (wy)^{1/2} &= 2(A^2-b^2)^{1/2}. \end{aligned}$$

If  $\text{Re}(\lambda + \mu + \nu) > -1$  we find

$$(2.12) \quad \begin{aligned} I(\mu, \nu; \lambda) &= C(\mu, \nu) \frac{\Gamma(\lambda + \mu + \nu + 1)}{2(2\mu + 1)_m(2\nu + 1)_n} \sum_{r=0}^m \sum_{s=0}^n \binom{m}{r} \binom{n}{s} \\ &\cdot (\mu + \frac{1}{2})_r (\mu + \frac{1}{2})_{m-r} (\nu + \frac{1}{2})_s (\nu + \frac{1}{2})_{n-s} (x^{r+s-\lambda} y^{r+n-s-\lambda} + w^{r+s-\lambda} z^{r+n-s-\lambda}) \\ &\cdot R_{-\mu-(1/2)-r}(\nu + \frac{1}{2} + s, \nu + \frac{1}{2} + n - s; r_+^2, r_-^2), \end{aligned}$$

where  $\binom{m}{r}$  is a binomial coefficient. The equation still holds if  $\sum_{r=0}^m \sum_{s=0}^n$  is replaced for quicker computation by

$$\sum_{r=0}^m \sum_{s=0}^{[n/2]} (2 - \delta_{n,2s}) \quad \text{or} \quad \sum_{r=0}^{[m/2]} \sum_{s=0}^n (2 - \delta_{m,2r}).$$

The singularity of  $(\mu + \frac{1}{2})_r (\mu + \frac{1}{2})_{m-r} / (2\mu + 1)_m$  when  $\mu = -\frac{1}{2}, -\frac{3}{2}, \dots$  is removable, and similarly for  $\nu$ . If  $\mu, \nu, a, b$  are real, then  $w = \bar{x}$  and  $z = \bar{y}$  and the summand is real.

When  $2\mu, 2\nu, 2\lambda$  are integers, the nature of the  $R$ -functions in the preceding formulas can readily be identified. Let  $m, n, r, s$  be integers. Setting  $s = 0$  is equivalent

[8, (6.3-3)] to omitting  $s$  and  $z$  in the following rules, wherein  $x, y, z$  are independent variables unrelated to the notation of (2.10):

1)  $R_m(n, r; x, y)$ , where  $n + r > 0$ , is a polynomial in  $x$  and  $y$  if  $m \geq 0$  and a rational function if  $m < 0$  and  $m + n + r \leq 0$  or if  $m < 0$  and exactly one of  $n$  and  $r$  is positive. Otherwise it involves a logarithm.

2)  $R_m(n + \frac{1}{2}, r, s; x, y, z)$  is a polynomial in  $x, y, z$  if  $m \geq 0$ ; a rational function if  $m < 0, r \leq 0, \text{ and } s \leq 0$ ; and a logarithm or arctangent otherwise.

3)  $R_m(n + \frac{1}{2}, r + \frac{1}{2}, s; x, y, z)$ , where  $n + r + s \geq 0$ , is a polynomial in  $x, y, z$  if  $m \geq 0$ ; a logarithm or arctangent if  $m < 0, m + n + r + s \geq 0, \text{ and } s > 0$ ; and an algebraic function otherwise.

4)  $R_{m-1/2}(n, r; x, y)$ , where  $n + r > 0$ , is an algebraic function of  $x$  and  $y$ .

5)  $R_{m-1/2}(n + \frac{1}{2}, r, s; x, y, z)$  is an algebraic function of  $x, y, z$  if  $m + n + r + s \leq 0$  or if  $r \leq 0$  and  $s \leq 0$ . Otherwise it is a logarithm or arctangent.

6)  $R_{m-1/2}(n + \frac{1}{2}, r + \frac{1}{2}, s; x, y, z)$ , where  $n + r + s \geq 0$ , is a complete elliptic integral of the first kind if  $m = n = r = s = 0$ ; the third kind if  $s > 0$ ; and the second kind otherwise.

Suppose for example that  $\lambda, \mu, \nu$  are integers ( $\mu$  and  $\nu$  being nonnegative). If  $\lambda \geq |\mu - \nu|$ , (2.12) and the last rule show that  $I(\mu, \nu; \lambda)$  is a complete elliptic integral of the first or second kind. The conjecture that it is a complete elliptic integral of the third kind if  $\lambda < |\mu - \nu|$  is supported by (2.3), (2.4), and the recurrence relations in [10].

For another example suppose that one of  $\lambda, \mu, \nu$  is an integer and the other two are half-odd integers. Then  $\lambda + \mu - \nu$  and  $\lambda + \nu - \mu$  are integers, and (2.12) shows that  $I(\mu, \nu; \lambda)$  is an elementary function if  $\lambda \geq |\mu - \nu|$ .

Near the end of § 3 it is shown that

$$(2.13) \quad I(\mu, \nu; \mu + \nu) = C(\mu, \nu) \Gamma(2\mu + 2\nu + 1) p^{-2\mu - 2\nu - 1} \cdot F_2\left(\mu + \nu + 1, \mu + \nu + \frac{1}{2}, \mu + \nu + \frac{1}{2}; \mu + 1, \nu + 1; \frac{p^2 + b^2 - A^2}{p^2}, \frac{p^2 + a^2 - A^2}{p^2}\right),$$

$\text{Re}(\mu + \nu) > -\frac{1}{2},$

$$(2.14) \quad I(\mu, \nu; \mu + \nu + 1) = C(\mu, \nu) \Gamma(2\mu + 2\nu + 2) p^{-2\mu - 2\nu - 2} \cdot F_2\left(\mu + \nu + 1, \mu + \nu + \frac{3}{2}, \mu + \nu + \frac{3}{2}; \mu + 1, \nu + 1; \frac{p^2 + b^2 - A^2}{p^2}, \frac{p^2 + a^2 - A^2}{p^2}\right),$$

$\text{Re}(\mu + \nu) > -1.$

We note also some special cases of (2.4):

$$(2.15) \quad I\left(\mu, \mu + \frac{1}{2}; \frac{1}{2}\right) = \left(\frac{2}{\pi}\right)^{1/2} a^\mu b^{\mu+1/2} \frac{(A^2 - a^2)^{1/2}}{G^2 A^{2\mu+1}}, \quad \text{Re } \mu > -1;$$

$$(2.16) \quad \int_0^\infty e^{-pt} J_0(at) \cos(bt) dt = \frac{(A^2 - b^2)^{1/2}}{G^2};$$

$$(2.17) \quad \int_0^\infty e^{-pt} J_0(at) \sin(bt) dt = \frac{(A^2 - a^2 - p^2)^{1/2}}{G^2}.$$

Equation (2.17) is given by Benton [3] in a different form.

**3. Proofs by reduction of  $F_4$ .** From [1, p. 102] and [8, Ex. 6.3-5] (alternatively see [6, (4.2)] and [7, (1.8)]), we find

$$(3.1) \quad F_4[\alpha, \beta; \gamma, \beta; X(1-Y), Y(1-X)] \\ = R_{-\alpha}[\beta + \gamma - \alpha - 1, 1 + \alpha - \gamma, \gamma - \beta; (1-X)(1-Y), 1-X-Y, 1-Y].$$

Define

$$(3.2) \quad x = X(1-Y), \quad y = Y(1-X), \\ z_{\pm}^2 = 1 - (x^{1/2} \pm y^{1/2})^2, \quad \xi = \frac{1}{2}(z_+ + z_-), \quad \eta^2 = z_+ z_-.$$

If  $1 \pm x^{1/2} \pm y^{1/2}$  is in the open right half-plane for all four choices of signs, we can choose  $z_+, z_-, \eta$  in the open right half-plane. The first two equations of (3.2) can be solved for

$$(3.3) \quad X = x + \xi^2 - \eta^2 = 1 - y - \xi^2, \quad Y = y + \xi^2 - \eta^2 = 1 - x - \xi^2, \\ (1-X)(1-Y) = \xi^2, \quad 1-X-Y = \eta^2, \quad XY = \xi^2 - \eta^2.$$

Substitution in (3.1) yields

$$(3.4) \quad F_4(\alpha, \beta; \gamma, \beta; x, y) = R_{-\alpha}(\beta + \gamma - \alpha - 1, 1 + \alpha - \gamma, \gamma - \beta; \xi^2, \eta^2, \xi^2 + x).$$

Equations (1.2) and (3.4), together with the symmetry of  $F_4$  in its first two parameters, imply (2.3) and (2.4). The case  $\mu = \nu$  of (2.3) can be transformed by [8, (6.10-1)] into the case  $n = 0$  of (2.5). The general case of (2.5) is then obtained by  $n$  differentiations with respect to  $p$  using [8, Ex. 5.9-18].

From [6, (2.5)] and [8, Ex. 6.3-5] we find

$$(3.5) \quad F_4(\alpha, 2\gamma - 1; \gamma, \gamma; x, y) = R_{-\alpha}(\gamma - \frac{1}{2}, \gamma - \frac{1}{2}, 1 - \gamma; z_+^2, z_-^2, 1).$$

Equations (1.2) and (3.5) imply the case  $n = 0$  of (2.6), and the general case is then obtained by  $n$  differentiations with respect to  $p$  using [8, Ex. 5.9-18].

From [1, p. 81] and [8, (5.9-12)] (alternatively from [6, (4.1)]), we find

$$(3.6) \quad F_4(\alpha, \beta; \gamma, 1 + \alpha + \beta - \gamma; x, y) \\ = R_{-\alpha}(\beta, \gamma - \beta; \xi^2 + y, 1) R_{-\alpha}(\beta, 1 + \alpha - \gamma; \xi^2 + x, 1).$$

Equations (1.2) and (3.6) lead to (2.7) and the first form of  $g$  given in (2.9). The second form of  $g$  is deduced from the first by [8, (6.9-7)], and the third form from the second by [8, (6.10-1)]. The expressions for  $f$  in (2.8) follow from those for  $g$  by (2.2) and the homogeneity of  $R$ .

From [1, p. 102], or from [6, (4.3)] and [5, (3.1)], we find

$$(3.7) \quad F_4(\gamma + \delta - 1, \beta; \gamma, \delta; x, y) = F_2(\gamma + \delta - 1, \beta, \beta; \gamma, \delta; 1 - y - \xi^2, 1 - x - \xi^2).$$

Equations (1.2) and (3.7), together with the symmetry of  $F_4$  in its first two parameters, imply (2.13) and (2.14).

The six rules about the nature of the  $R$ -function are deduced from the results and methods of [8, §§ 8.4, 8.5, 9.3]. Equation (2.15) follows from (2.4) and [8, (6.8-15)], and (2.16) and (2.17) are the cases  $\mu = -\frac{1}{2}$  and  $\mu = 0$ , respectively.



**4. Proof by reduction of  $F_2$ .** By [5, (3.1), (2.9)] we can rewrite (1.4) in the form

$$(4.1) \quad I(\mu, \nu; \lambda) = C(\mu, \nu)\Gamma(\lambda + \mu + \nu + 1)\mathcal{R}_{-\lambda - \mu - \nu - 1}(\mu + \frac{1}{2}, \mu + \frac{1}{2}; Z; \nu + \frac{1}{2}, \nu + \frac{1}{2}),$$

$$Z = \begin{bmatrix} x & y \\ z & w \end{bmatrix} = \begin{bmatrix} p + ia + ib & p + ia - ib \\ p - ia + ib & p - ia - ib \end{bmatrix}.$$

To prove (2.12) we now assume that  $\lambda + \nu - \mu = m$  and  $\lambda + \mu - \nu = n$ , where  $m$  and  $n$  are nonnegative integers. Then the sum of the homogeneity parameter  $(-\lambda - \mu - \nu - 1)$  and the row parameters  $(\mu + \frac{1}{2}, \mu + \frac{1}{2})$  is  $-m$ . We aim to increase by  $m$  the sum of the row parameters by using a relation between associated  $\mathcal{R}$ -functions, and similarly for the column parameters. In the notation of [5],

$$(4.2) \quad \mathcal{R}_t(\alpha, \alpha'; Z; \beta, \beta') = \int_0^1 \int_0^1 (u \cdot Z \cdot v)^t d\mu_{(\alpha, \alpha')}(u) d\mu_{(\beta, \beta')}(v).$$

Inserting unity in the integrand in the form

$$1 = (u + 1 - u)^m = \sum_{r=0}^m \binom{m}{r} u^r (1 - u)^{m-r}$$

and using [8, (5.6-7)], we find

$$(4.3) \quad \mathcal{R}_t(\alpha, \alpha'; Z; \beta, \beta') = \frac{1}{(\alpha + \alpha')_m} \sum_{r=0}^m \binom{m}{r} (\alpha)_r (\alpha')_{m-r} \cdot \mathcal{R}_t(\alpha + r, \alpha' + m - r; Z; \beta, \beta').$$

From (4.3) and a corresponding relation that raises the column parameters, we get

$$(4.4) \quad I(\mu, \nu; \lambda) = C(\mu, \nu) \frac{\Gamma(\lambda + \mu + \nu + 1)}{(2\mu + 1)_m (2\nu + 1)_n} \sum_{r=0}^m \sum_{s=0}^n \binom{m}{r} \binom{n}{s} \cdot (\mu + \frac{1}{2})_r (\mu + \frac{1}{2})_{m-r} (\nu + \frac{1}{2})_s (\nu + \frac{1}{2})_{n-s} \cdot \mathcal{R}(\mu + \frac{1}{2} + r, \mu + \frac{1}{2} + m - r; Z; \nu + \frac{1}{2} + s, \nu + \frac{1}{2} + n - s).$$

The last function is a bare  $\mathcal{R}$ -function [5, (4.5)], in which the sum of the row parameters, the sum of the column parameters, and the negative of the degree of homogeneity are all equal. From [5, (5.3)] and [8, (5.9-11)] we find

$$(4.5) \quad \mathcal{R}(\alpha, \alpha'; Z; \beta, \beta') = z^{\alpha - \beta} w^{\alpha - \beta'} R_{-\alpha}(\beta, \beta'; xw, yz), \quad Z = \begin{bmatrix} x & y \\ z & w \end{bmatrix}.$$

The double sum in (4.4) will now be rearranged so that it is plainly real when  $\mu, \nu, p, a, b$  are real. We denote the summand by  $u_{r,s}(Z)$  and write the double sum in the form

$$(4.6) \quad \sum_{r=0}^m \sum_{s=0}^n u_{r,s}(Z) = \frac{1}{2} \sum_{r=0}^m \sum_{s=0}^n v_{r,s},$$

$$v_{r,s} = u_{r,s}(Z) + u_{m-r, n-s}(Z).$$

The row and column symmetries of  $\mathcal{R}$  imply

$$(4.7) \quad v_{r,s} = u_{r,s}(Z) + u_{r,s}(Z'), \quad Z = \begin{bmatrix} x & y \\ z & w \end{bmatrix}, \quad Z' = \begin{bmatrix} w & z \\ y & x \end{bmatrix}.$$

If  $\mu, \nu, p, a, b$  are real,  $Z$  and  $Z'$  are conjugate complex and  $v_{r,s}$  is real. From (4.4) and (4.5) we find

$$(4.8) \quad u_{r,s}(Z) = \binom{m}{r} \binom{n}{s} (\mu + \frac{1}{2})_r (\mu + \frac{1}{2})_{m-r} (\nu + \frac{1}{2})_s (\nu + \frac{1}{2})_{n-s} W^{r+s-\lambda} Z^{r+n-s-\lambda} \\ \cdot R_{-\mu-(1/2)-r}(\nu + \frac{1}{2} + s, \nu + \frac{1}{2} + n - s; xW, yZ).$$

Adding  $u_{r,s}(Z')$  to get  $v_{r,s}$ , we obtain (2.12). Since  $v_{r,s} = v_{m-r,n-s}$  by (4.6), we can write (with Kronecker deltas)

$$(4.9) \quad \sum_{r=0}^m \sum_{s=0}^n v_{r,s} = \sum_{r=0}^m \sum_{s=0}^{[n/2]} (2 - \delta_{n,2s}) v_{r,s} = \sum_{r=0}^{[m/2]} \sum_{s=0}^n (2 - \delta_{m,2r}) v_{r,s}$$

to reduce the number of terms for purposes of computation.

## REFERENCES

- [1] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London 1935.
- [2] ———, *The generating function of Jacobi polynomials*, J. London Math. Soc., 13 (1938), pp. 8–12.
- [3] T. C. BENTON, *Concerning  $\int_0^\infty e^{-at} J_\mu(bt) J_\nu(ct) t^{\mu-\nu} dt$* , this Journal, 5 (1975), pp. 761–765.
- [4] C. J. BOUWKAMP, *Solution of Problem 74-20*, SIAM Rev., 18 (1976), pp. 123–126.
- [5] B. C. CARLSON, *Appell functions and multiple averages*, this Journal, 2 (1971), pp. 420–430.
- [6] ———, *Appell's function  $F_4$  as a double average*, this Journal, 6 (1975), pp. 960–965.
- [7] ———, *Quadratic transformations of Appell functions*, this Journal, 7 (1976), pp. 291–304.
- [8] ———, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.
- [9] J. DETRICH AND R. W. CONN, *Analytic evaluation of an important integral in collision theory*, J. Mathematical Phys., 18 (1977), pp. 2348–2351.
- [10] G. EASON, B. NOBLE AND I. N. SNEDDON, *On certain integrals of Lipschitz–Hankel type involving products of Bessel functions*, Philos. Trans. Roy. Soc. London Ser. A, 247 (1955), pp. 529–551.
- [11] A. ERDÉLYI, ed., *Tables of Integral Transforms*, 2 vols., McGraw-Hill, New York, 1954.
- [12] A. GRAY, *Notes on electric and magnetic field constants and their expression in terms of Bessel functions and elliptic integrals*, Phil. Mag. (6), 38 (1919), pp. 201–214.
- [13] T. H. HAVELOCK, *On certain Bessel integrals and the coefficients of mutual induction of coaxial coils*, Phil. Mag. (6), 15 (1908), pp. 332–345.
- [14] F. T. KROGH, E. W. NG AND W. V. SNYDER, *The gravitational attraction due to a disk*, Sect. 366, Computing Memorandum 444, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 1979.
- [15] Y. L. LUKE, *Integrals of Bessel Functions*, McGraw-Hill, New York, 1962.
- [16] F. OBERHETTINGER AND L. BADI, *Tables of Laplace Transforms*, Springer-Verlag, New York, 1973.
- [17] S. OKUI, *Complete elliptic integrals resulting from infinite integrals of Bessel functions*, J. Res. Nat. Bur. Standards Sect. B, 78 (1974), pp. 113–135; 79 (1975), pp. 137–170.
- [18] D. D. PENROD AND C. FARELL, *On the evaluation of certain integrals containing products of Bessel functions*, Indust. Math., 21 (1971), pp. 85–89.
- [19] S. K. SINGH, *Gravitational attraction of a circular disk*, Geophysics, 42 (1977), pp. 111–113.
- [20] ———, *Gravitational attraction of a vertical right circular cylinder*, Geophysical J. Roy. Astron. Soc., 50 (1977), pp. 243–246.
- [21] S. K. SINGH AND F. J. SABINA, *Magnetic anomaly due to a vertical right circular cylinder with arbitrary polarization*, Geophysics, 43 (1978), pp. 173–178.
- [22] S. K. SINGH, R. CASTRO, E. AND M. GUZMAN S., *Magnetic anomaly of a circular lamina*, Geophysics, 44 (1979), pp. 102–107.
- [23] I. N. SNEDDON, *Fourier Transforms*, McGraw-Hill, New York, 1951.
- [24] C. SNOW, *Magnetic fields of cylindrical coils and annular coils*, National Bureau of Standards Applied Mathematics Series 38, U.S. Government Printing Office, Washington, DC, 1953.
- [25] G. N. WATSON, *Bessel Functions*, 2nd ed., Cambridge University Press, London, 1944.
- [26] A. D. WHEELON, *Tables of Summable Series and Integrals Involving Bessel Functions*, Holden-Day, San Francisco, 1968.
- [27] J. C. BELL, *Stresses from arbitrary loads on a circular crack*, International Journal of Fracture, 15 (1979), pp. 85–104.

## DEGREE OF $L_p$ APPROXIMATION BY MONOTONE SPLINES\*

C. K. CHUI,† P. W. SMITH† AND J. D. WARD†

**Abstract.** Using elementary techniques, we obtain Jackson type estimates for the approximation of monotone nondecreasing functions by monotone nondecreasing splines with equally spaced knots in  $L_p[0, 1]$ ,  $1 \leq p \leq \infty$ . Our method, which works for all  $p$ , is different from that of De Vore.

**1. Introduction.** In [4], De Vore proved that the Jackson type estimates for approximating monotone nondecreasing functions by monotone nondecreasing splines with equally spaced knots in  $L_\infty[0, 1]$  are of the same order as the Jackson type estimates in unconstrained approximation. The proofs in [4] are very clever but at the same time are quite complicated, and it is not clear to these authors whether the techniques employed there can be extended to settle the  $L_p[0, 1]$  problem for  $1 \leq p < \infty$ . The object of this paper is to introduce different and perhaps more transparent methods to give the  $L_p[0, 1]$  results for  $1 \leq p \leq \infty$ .

Let  $k$  and  $N$  be positive integers and let  $\mathcal{S}(k, N)$  denote the space of all splines of order  $k$  with knots  $\{i/N\}_{i=0}^N$ . If  $A$  is a collection of functions defined on  $[0, 1]$ , then  $A^*$  will denote the subcollection of functions in  $A$  which are nondecreasing on  $[0, 1]$ . Hence,  $\mathcal{S}^*(k, N)$  is the set of those splines  $s$  in  $\mathcal{S}(k, N)$  with  $s \uparrow$ . For any nonnegative integer  $j$  and  $1 \leq p < \infty$ , let  $L_p^j[0, 1]$  be the space of functions which are  $j$ -fold integrals of  $L_p[0, 1]$  functions. For convenience, we also let  $L_\infty^j[0, 1] = C^j[0, 1]$  denote the space of all  $j$  times continuously differentiable functions on  $[0, 1]$ . If  $f$  is a nondecreasing function on  $[0, 1]$ , we will study the  $L_p[0, 1]$  distance,  $1 \leq p \leq \infty$ , of  $f$  from  $\mathcal{S}^*(k, N)$ , denoted by  $E_{N,p}^*(f, k) = \inf \{\|f - s\|_p : s \in \mathcal{S}^*(k, N)\}$ , where  $\|\cdot\|_p = \|\cdot\|_{L_p[0,1]}$  is the usual  $L_p[0, 1]$  norm on  $[0, 1]$ . We also let  $\omega(f, h)_p \equiv \omega(f, h; [0, 1])_p$  denote the  $L_p[0, 1]$  modulus of continuity of  $f$  on  $[0, 1]$ . That is,

$$\omega(f, h)_p \equiv \omega(f, h; [0, 1])_p = \sup_{0 \leq t \leq h} \left( \int_0^{1-t} |f(x+t) - f(x)|^p dx \right)^{1/p}$$

if  $1 \leq p < \infty$ , and

$$\omega(f, h)_\infty \equiv \omega(f, h; [0, 1])_\infty = \sup_{0 \leq x \leq 1-h} |f(x+h) - f(x)|.$$

We will establish the following result.

**THEOREM 1.1.** *Let  $1 \leq p \leq \infty$  and  $k$  be a positive integer. There is a constant  $C > 0$ , depending only on  $k$  and  $p$ , such that if  $f \in L_p^{j*}[0, 1]$  where  $0 \leq j \leq k - 1$ , then*

$$(1.1) \quad E_{N,p}^*(f, k) \leq CN^{-j} \omega\left(f^{(j)}, \frac{1}{N}\right)_p,$$

where  $N = 1, 2, \dots$ .

As mentioned above, the case where  $p = \infty$  is a result due to De Vore [4]. Our proof for  $p = \infty$  is different from that of De Vore's and employs "local techniques" which work for all  $p$ ,  $1 \leq p \leq \infty$ . If  $k$  equals 1 or 2, Theorem 1 follows easily. So we always assume  $k \geq 3$ . In the next section, we will first study unconstrained piecewise polynomial approximation, proving the classical Jackson type estimates as stated in

\* Received by the editors February 8, 1979, and in revised form August 14, 1979. This research was supported by the United States Army Research Office under Grants DAHC 04-75-G-0186 and DAAG 29-78-G-0097.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

Proposition 2.1. In Section 3, the analogous estimates for monotone approximation by piecewise polynomials are obtained. All these are preliminary results needed for our proof of Theorem 1.1, which is given in § 4. We remark again that the techniques employed in this paper are elementary and self-contained.

**2. Unconstrained approximation by piecewise polynomials.** In this section we will study best approximation in  $L_p[0, 1]$ ,  $1 \leq p \leq \infty$ , by (continuous) piecewise polynomials with equally spaced knots. For  $f \in L_p[0, 1]$ , let

$$\|f\|_p = \inf \{ \|f - c\|_p : c \text{ a real constant} \}$$

denote the distance in  $L_p[0, 1]$  of  $f$  from the one-dimensional space of all real constants. We will first prove that  $\|\cdot\|_p$  and  $\omega(\cdot, 1)_p$  are equivalent. A proof of this result is also given in [3, Thm. 3.1] by using the  $K$ -functional. Our proof is very elementary.

LEMMA 2.1. *Let  $1 \leq p \leq \infty$ . There exist positive constants  $c_1$  and  $c_2$  such that  $c_1\|f\|_p \leq \omega(f, 1)_p \leq c_2\|f\|_p$  for all  $f \in L_p[0, 1]$ .*

*Proof.* The second inequality is trivial with  $c_2 = 2$  for all  $p$ ,  $1 \leq p \leq \infty$ . We proceed with the proof of the existence of  $c_1 > 0$  by contradiction. Suppose  $c_1$  does not exist. Then there exists a sequence of functions  $f_n \in L_p[0, 1]$ , satisfying  $\|f_n\|_p = 1$ ,  $\int_0^1 f_n = 0$ , and  $\omega(f_n, 1)_p \rightarrow 0$  as  $n \rightarrow \infty$ . If  $Q_1 f = \int_0^1 f$  denotes the projection from  $L_p[0, 1]$  to  $\mathbb{R}$ , then  $\|f_n\|_p = \|f_n - \int_0^1 f_n\|_p = \|(I - Q_1)f_n\|_p = \|(I - Q_1)(f_n - c)\|_p \leq \|I - Q_1\| \|f_n - c\|_p$  for all real constants  $c$ , so that  $\|f_n\|_p \leq \|I - Q_1\| \cdot \|f_n\|_p = \|I - Q_1\|$ . Thus,  $\{f_n\}$  is a bounded sequence in  $L_p[0, 1]$ . Also, since  $\omega(f_n, 1)_p \rightarrow 0$ , we have

$$(2.1) \quad \|f_n(x+t) - f_n(x)\|_{L_p[0,1-t]} \rightarrow 0$$

for every  $t \in [0, 1]$ . Let

$$g_n(x) = 2 \int_0^{1/2} f_n(x+t) dt.$$

Then  $\{g_n\}$  is a bounded sequence in  $L_p[0, 1]$ , and because of (2.1), is also equicontinuous on  $[0, \frac{1}{2}]$ . By the Ascoli theorem, it has a subsequence, which we will also denote by  $\{g_n\}$ , that converges uniformly on  $[0, \frac{1}{2}]$  to some function  $g$ . But from (2.1), using the generalized Minkowski inequality, we also have  $\|g_n - f_n\|_{L_p[0,1/2]} \rightarrow 0$ , so that  $\|f_n - g\|_{L_p[0,1/2]} \rightarrow 0$ . By (2.1) again,  $g(x+t) - g(x) = 0$  a.e. for almost all  $t \in [0, 1/2]$ , and this means that  $g$  is a constant a.e. This constant is in fact equal to zero, since  $\int_0^1 f_n = 0$  for all  $n$ . Hence,  $g_n \rightarrow 0$  uniformly on  $[0, \frac{1}{2}]$ , so that  $\|f_n\|_{L_p[0, \frac{1}{2}]} \rightarrow 0$ . By the change of variable  $x \rightarrow 1-x$ , we also have  $\|f_n\|_{L_p[\frac{1}{2}, 1]} \rightarrow 0$ . Therefore, we conclude that  $\|f_n\|_p \rightarrow 0$ . This contradicts the fact that  $\|f_n\|_p \geq \|f_n\|_p = 1$  for all  $n$ , completing the proof of the lemma.

If  $f$  is a function defined on  $[0, 1]$ ,  $P_k f$  will denote the polynomial of degree  $k$  that interpolates  $f$  at the points  $0, 1/k, 2/k, \dots, 1$ . We will establish the following inequality, which also follows from [3, Thm. 3.1]. Again, since our proof here is very elementary and it does not depend on the  $K$ -functional as in [3], we include it here for completeness.

LEMMA 2.2. *Let  $j$  be a nonnegative integer and  $1 \leq p \leq \infty$ . There is a positive constant  $c_3$  such that*

$$(2.2) \quad \|(f - P_j f)^{(i)}\|_p \leq c_3 \omega(f^{(i)}, 1)_p,$$

for all  $f \in L_p^j[0, 1]$  and  $i = 0, \dots, j$ .

*Proof.* Fix  $i \leq j$  and consider  $(f - P_j f)^{(i)}$ . If  $i \neq j$ , then  $(f - P_j f)^{(i)}$  has at least  $j - i + 1$  zeros, so that

$$\begin{aligned} (f - P_j f)^{(i)}(x) &= \int_{x_1}^x (f - P_j f)^{(i+1)}(\tau_1) d\tau_1 \\ &= \dots = \int_{x_1}^x \dots \int_{x_{j-i}}^{\tau_{j-i-1}} (f - P_j f)^{(i)}(\tau_{j-i}) d\tau_1 \dots d\tau_{j-i} \end{aligned}$$

for some appropriate  $x_1, \dots, x_{j-i} \in [0, 1]$ . Hence, we have

$$\|(f - P_j f)^{(i)}\|_p \leq \|(f - P_j f)^{(j)}\|_p$$

for all  $i = 0, \dots, j$ , and it is therefore sufficient to obtain (2.2) for  $i = j$ . To do this, recall that by the Peano kernel theorem,

$$(P_j f)^{(j)} = \int_0^1 M_j f^{(j)}$$

where  $M_j$  is a nonnegative  $B$ -spline of order  $j$  with knots at  $\{i/j\}, i = 0, \dots, j$ , which integrates to 1. Hence, the mapping  $Q_2 g = \int_0^1 M_j g$  is a linear projection from  $L_p[0, 1]$  to the space of constants. It follows that

$$\begin{aligned} \|(f - P_j f)^{(j)}\|_p &= \|(I - Q_2)f^{(j)}\|_p \\ &= \|(I - Q_2)(f^{(j)} - c)\|_p \leq \|I - Q_2\| \|f^{(j)} - c\|_p \end{aligned}$$

for any constant  $c$ , so that

$$\begin{aligned} \|(f - P_j f)^{(j)}\|_p &\leq \|I - Q_2\| \|f^{(j)}\|_p \\ &\leq c_1^{-1} \|I - Q_2\| \omega(f^{(j)}, 1)_p \end{aligned}$$

by applying the first inequality in Lemma 2.1. This completes the proof of Lemma 2.2 with  $c_3 = c_1^{-1} \|I - Q_2\|$ .

Before we can change the interval  $[0, 1]$  to an arbitrary subinterval so that we can study piecewise polynomial approximation in  $L_p[0, 1]$ , we need the following lemma to estimate the sum of the errors. Again, we need some notation which will be used throughout the rest of the paper:

$$\omega(f, h; [a, b])_p = \sup_{0 \leq t \leq h} \left( \int_a^{b-t} |f(x+t) - f(x)|^p dx \right)^{1/p}$$

if  $1 \leq p < \infty$ , and

$$\omega(f, h; [a, b])_\infty = \sup_{a \leq x \leq b-h} |f(x+h) - f(x)|.$$

Here,  $0 < h \leq b - a$ . Let  $0 = x_0 < x_1 < \dots < x_N = 1$  be a partition of  $[0, 1]$  and let

$$\delta_i = x_i - x_{i-1} \quad \text{and} \quad \delta = \max_{1 \leq i \leq N} \delta_i.$$

We have the following

LEMMA 2.3. *Let  $1 \leq p < \infty$ , and  $0 = x_0 < \dots < x_N = 1$ ,  $\delta > 0$  be as above. There exists a positive constant  $c_4$ , depending only on  $p$ , such that*

$$(2.3) \quad \sum_{i=1}^N \omega(f, \delta_i; [x_{i-1}, x_i])_p \leq c_4 \omega(f, \delta)_p^p.$$

*Proof.* Let  $g \in L_p^1[0, 1]$  and  $0 < t \leq \delta_i$ . Then we have

$$\begin{aligned}
 \int_{x_{j-1}}^{x_j-t} |g(x+t) - g(x)|^p dx &= \int_{x_{j-1}}^{x_j-t} \left| \int_x^{x+t} g' \right|^p dx \\
 (2.4) \qquad \qquad \qquad &\leq \int_{x_{j-1}}^{x_j-t} t^{p/q} \left( \int_x^{x+t} |g'|^p \right) dx \\
 &\leq \delta^p \int_{x_{j-1}}^{x_j} |g'|^p.
 \end{aligned}$$

By Jensen's inequality, it follows that

$$\begin{aligned}
 \int_{x_{j-1}}^{x_j-t} |f(x+t) - f(x)|^p dx &\leq 3^{p-1} \left( \int_{x_{j-1}}^{x_j-t} |f(x+t) - g(x+t)|^p dx + \int_{x_{j-1}}^{x_j-t} |g(x+t) - g(x)|^p dx \right. \\
 &\qquad \qquad \qquad \left. + \int_{x_{j-1}}^{x_j-t} |f(x) - g(x)|^p dx \right) \\
 &\leq 2 \cdot 3^{p-1} \left( \int_{x_{j-1}}^{x_j} |f - g|^p + \int_{x_{j-1}}^{x_j-t} |g(x+t) - g(x)|^p dx \right).
 \end{aligned}$$

Hence, by using (2.4), we have

$$\int_{x_{j-1}}^{x_j-t} |f(x+t) - f(x)|^p dx \leq 2 \cdot 3^{p-1} (\|f - g\|_{L_p[x_{j-1}, x_j]}^p + \delta^p \|g'\|_{L_p[x_{j-1}, x_j]}^p),$$

and this gives

$$\begin{aligned}
 \sum_{i=1}^N \omega(f, \delta_i; [x_{i-1}, x_i])_p^p &\leq 2 \cdot 3^p (\|f - g\|_p^p + \delta^p \|g'\|_p^p) \\
 &\leq 2 \cdot 3^p (\|f - g\|_p + \delta \|g'\|_p)^p.
 \end{aligned}$$

The above inequality holds for all  $g \in L_p^1[0, 1]$ . Thus, we have

$$(2.5) \qquad \sum_{i=1}^N \omega(f, \delta; [x_{i-1}, x_i])_p^p \leq 2 \cdot 3^p (K_{1,p}(\delta)f)^p$$

where

$$K_{1,p}(\delta)f = \inf_{g \in L_p^1[0, 1]} (\|f - g\|_p + \delta \|g'\|_p)$$

is a  $K$ -functional of Peetre. It is well known (cf. [4], [8]) that  $K_{1,p}(\delta)f$  is equivalent to  $\omega(f, \delta)_p$ . Hence, (2.5) implies (2.3) for some positive constant  $c_4$ , depending only on  $p$ . This completes the proof of the lemma.

We are now ready to derive the Jackson type estimates for (unconstrained) approximation by (continuous) piecewise polynomials. Let  $\pi(k, N)$  be the space of all continuous functions  $f$  on  $[0, 1]$  such that the restriction  $f|_{[(i-1)/N, i/N]}$  of  $f$  on  $[(i-1)/N, i/N]$  is a polynomial of degree  $\leq k$ ,  $i = 1, \dots, N$ ; and for  $f \in L_p[0, 1]$ ,  $1 \leq p \leq \infty$ , let

$$D_{N,p}(f, k) = \inf \{ \|f - g\|_p : g \in \pi(k, N) \}.$$

The following well known result is now easily proved. We will give a proof of this result in order to facilitate the proof of Proposition 3.1 in the next section.

PROPOSITION 2.1. *Let  $1 \leq p \leq \infty$ . There exists a positive constant  $c_5$ , depending only on  $k$  and  $p$ , such that*

$$(2.6) \quad D_{N,p}(f, k) \leq c_5 N^{-j} \omega(f^{(j)}, 1/N)_p$$

for all  $f \in L_p^j[0, 1]$ ,  $0 \leq j \leq k$ .

*Proof.* Let  $f \in L_p^j[0, 1]$ ,  $0 \leq j \leq k$ , and set  $f_i(t) = f((t+i-1)/N)$ . Also, let  $g_i = P_j f_i$  be the polynomial of degree  $\leq j$ , interpolating  $f_i$  at the points  $0, 1/j, 2/j, \dots, 1$ . By Lemma 2.2, we have

$$(2.7) \quad \|f_i - g_i\|_p \leq c_3 \omega(f_i^{(j)}, 1)_p.$$

Let  $h \in \pi(j, N)$  be such that  $g_i(t) = h((t+i-1)/N)$ ,  $i = 1, \dots, N$ . That  $h$  is continuous follows since  $g_{i-1}(1) = g_i(0) = f((i-1)/N)$ ,  $i = 1, \dots, N$ . For  $p = \infty$ , (2.7) immediately gives

$$D_{N,\infty}(f, k) \leq D_{N,\infty}(f, j) \leq \|f - h\|_\infty \leq c_3 N^{-j} \omega(f^{(j)}, 1/N)_\infty$$

which is (2.6) with  $c_5 = c_3$ . For  $1 \leq p < \infty$ , we can apply Lemma 2.3 with  $x_i = i/N$  and  $\delta = 1/N$  to obtain

$$\begin{aligned} \|f - h\|_p^p &= \frac{1}{N} \sum_{i=1}^N \|f_i - g_i\|_p^p \\ &\leq c_3^p N^{-1} \sum_{i=1}^N \omega(f_i^{(j)}, 1)_p^p \\ &\leq c_3^p N^{-1} N^{-jp+1} \sum_{i=1}^N \omega\left(f^{(j)}, \frac{1}{N}; \left[\frac{i-1}{N}, \frac{i}{N}\right]\right)_p^p \\ &\leq c_3^p c_4 N^{-jp} \omega(f^{(j)}, 1/N)_p^p. \end{aligned}$$

Hence, we have

$$D_{N,p}(f, k) \leq D_{N,p}(f, j) \leq \|f - h\|_p \leq c_5 N^{-j} \omega(f^{(j)}, 1/N)_p,$$

with  $c_5 = c_3 c_4^{1/p}$ , completing the proof of the proposition.

**3. Monotone approximation by piecewise polynomials.** As in the above section,  $\pi(k, N)$  will denote the collection of all continuous functions on  $[0, 1]$  whose restrictions on each subinterval  $[(i-1)/N, i/N]$ ,  $i = 1, \dots, N$ , are polynomials of degree  $\leq k$ . Hence,  $\pi^*(k, N)$  is the subcollection of functions in  $\pi(k, N)$  which are non-decreasing on  $[0, 1]$ . Let

$$D_{N,p}^*(f, k) = \inf \{ \|f - g\|_p : g \in \pi^*(k, N) \}.$$

In this section, we will establish the following result.

PROPOSITION 3.1. *Let  $1 \leq p \leq \infty$  and  $k$  be a positive integer. There exists a positive constant  $c_6$ , depending only on  $p$  and  $k$ , such that*

$$(3.1) \quad D_{N,p}^*(f, k) \leq c_6 N^{-j} \omega(f^{(j)}, 1/N)_p$$

for all  $f \in L_p^{j*}[0, 1]$ ,  $0 \leq j \leq k$ .

We remind the reader that  $L_p^{j*}[0, 1]$  denotes the subcollection of functions in  $L_p^j[0, 1]$  which are nondecreasing on  $[0, 1]$ . If  $f \in L_p^{j*}[0, 1]$ , we wish to modify  $P_j f$ , the  $j$ th degree polynomial interpolating  $f$  at  $0, 1/j, \dots, 1$ , to give a nondecreasing polynomial approximant of  $f$ . To do this, we need the following two lemmas.

LEMMA 3.1. Let  $1 \leq p \leq \infty$  and  $j \geq 1$ . There exists a positive constant  $c_7$  such that

$$(3.2) \quad -\min \left( \min_{x \in [0,1]} (P_j f)'(x), 0 \right) \leq c_7 \omega(f^{(j)}, 1)_p$$

for all  $f \in L_p^{j*}[0, 1]$ .

*Proof.* Since  $f' \geq 0$  a.e., we have

$$-\min \left( (P_j f)'(x), 0 \right) \leq |(f - P_j f)'(x)|$$

for almost all  $x \in [0, 1]$ . Hence, (3.2) follows from the proof of Lemma 2.1 with  $i = 1$  and  $c_7 = c_3$ .

LEMMA 3.2. Let  $1 \leq p \leq \infty$  and  $j \geq 1$ . There exists a positive constant  $c_8$ , with the property that for every  $f \in L_p^{j*}[0, 1]$ , there is a non-decreasing polynomial  $g$  with degree  $\leq j$ , such that  $g(0) = f(0)$ ,  $g(1) = f(1)$  and

$$(3.3) \quad \|(f - g)^{(i)}\|_p \leq c_8 \omega(f^{(i)}, 1)_p,$$

for  $i = 0, \dots, j$ .

*Proof.* Let  $h = P_j f$  be the polynomial with degree  $\leq j$  which interpolates  $f$  at  $1/j$ ,  $i = 0, \dots, j$ , and set

$$g(x) = f(0) + \left[ \int_0^x h'(t) dt + d_h x \right] \frac{f(1) - f(0)}{f(1) - f(0) + d_h}$$

where  $d_h = -\min(\min_{x \in [0,1]} h'(x), 0)$ . Since  $h$  interpolates  $f$  at 0 and 1, it is clear that  $g(0) = f(0)$  and  $g(1) = f(1)$ . It is also clear that  $g' \geq 0$ , so that  $g$  is nondecreasing on  $[0, 1]$ . Without loss of generality, we assume that  $f(0) = 0$ . By using Lemmas 2.2 and 3.1, we have

$$\begin{aligned} \|(f - g)^{(i)}\|_p &\leq \|(f - h)^{(i)}\|_p + \|h^{(i)}\|_p (d_h / (f(1) + d_h)) + d_h (f(1)) / (f(1) + d_h) \\ &\leq (c_3 + c_7) \omega(f^{(i)}, 1)_p + \|h^{(i)}\|_p \frac{d_h}{h(1) + d_h}. \end{aligned}$$

Consider

$$B(f) \equiv \|h^{(i)}\|_p \frac{d_h}{h(1) + d_h},$$

where  $h = P_j f$ . Since  $B(\alpha f) = \alpha B(f)$  and  $\omega((\alpha f)^{(i)}, 1)_p = \alpha \omega(f^{(i)}, 1)_p$  for any positive constant  $\alpha$ , it is sufficient to prove that  $\{B(f) : f \in L_p^j[0, 1], \omega(f^{(i)}, 1)_p \leq 1\}$  is bounded. Also, since all  $L_p$  norms are equivalent on the space of all polynomials with degree  $\leq k$ , we conclude, using the Markov's inequality, that it is sufficient to prove that

$$\sup \left\{ \|h\|_\infty \frac{d_h}{h(1) + d_h} : f \in L_p^j[0, 1], \omega(f^{(i)}, 1)_p \leq 1 \right\} < \infty.$$

Let  $x_0 \in [0, 1]$  be chosen such that  $|h(x_0)| = \|h\|_\infty$ . If  $h(x_0) = \|h\|_\infty$ , then we have, for some  $a \in (x_0, 1)$ ,

$$h(1) - h(x_0) = h'(a)(1 - x_0) \geq -d_h(1 - x_0)$$



so that

$$\begin{aligned} \|h\|_\infty &= h(x_0) \\ &\leq h(1) + d_h(1 - x_0) \leq h(1) + d_h, \end{aligned}$$

and

$$\begin{aligned} \|h\|_\infty \frac{d_h}{h(1) + d_h} &\leq d_h \\ &\leq c_7 \omega(f^{(j)}, 1)_p \leq c_7 \end{aligned}$$

by Lemma 3.1. If, on the other hand,  $h(x_0) = -\|h\|_\infty$ , then we have, for some  $b \in (0, x_0)$ ,

$$h(x_0) = h'(b)x_0 \geq -d_h x_0,$$

so that

$$\begin{aligned} \|h\|_\infty \frac{d_h}{h(1) + d_h} &= \frac{-(h(x_0))d_h}{h(1) + d_h} \\ &\leq \frac{d_h^2 x_0}{h(1) + d_h} \leq d_h \leq c_7 \omega(f^{(j)}, 1)_p \leq c_7, \end{aligned}$$

again by Lemma 3.1. This completes the proof of the lemma.

By using Lemmas 2.3 and 3.2, and following the proof of Proposition 2.1, we have Proposition 3.1.

**4. Monotone approximation by splines.** In this section we will study monotone approximation by splines and prove Theorem 1.1. Let  $\mathbf{t} = \{t_i\}_{i=-\infty}^{k-1}$  where  $t_i = i$ , and  $\mathbf{s} = \{s_j\}_{j=0}^\infty$  where  $s_0 = \dots = s_{k-1} = 0$  and  $s_j = j - k + 1$  for  $j \geq k$ , be two knot sequences. For  $-\infty < i \leq -1$ , let  $N_i = N_{i,k,\mathbf{t}}$ , and for  $0 \leq i < \infty$ , let  $N_i = N_{i,k,\mathbf{s}}$  be the normalized  $B$ -splines of order  $k$  and with knots at  $\mathbf{t}$  and  $\mathbf{s}$  as indicated by the third subscripts (cf. [1]). Also, let  $X$  be the spline space spanned by these normalized  $B$ -splines  $N_i$ ,  $-\infty < i < \infty$ , and  $Y$  the subspace spanned by  $N_i$  where  $-\infty < i \leq -1$  and  $k - 1 \leq i < \infty$ . Denote by  $T$  a “smoothing operator” mapping  $X$  to  $Y$ , defined by

$$T\left(\sum_{i=-\infty}^\infty a_i N_i\right) = \sum_{i=-\infty}^{-1} a_i N_i + \sum_{i=k-1}^\infty a_i N_i,$$

and set  $E = I - T$ , where  $I$  is the identity map. Note that for any  $s \in X$ ,  $Es$  has finite support and is an element of  $L_p^{k-1}[(-\infty, 0) \cup (0, \infty)]$ ,  $1 \leq p \leq \infty$ . In addition, we have the following estimate.

LEMMA 4.1. *There is a positive constant  $c_9$ , depending only on  $k$ , such that for any  $s \in X$ ,*

$$(4.1) \quad \|(Es)^{(j)}\|_{L_p(\mathbb{R})} \leq c_9 \sum_{i=0}^{k-2} |s^{(i)}(0^+) - s^{(i)}(0^-)|$$

for  $0 \leq j \leq k - 1$  and  $1 \leq p \leq \infty$ .

*Proof.* By  $(Es)^{(j)}$  we mean the function defined pointwise by taking  $j$  derivatives. Since  $Y$  is the kernel of the linear operator  $E$ , we may assume that  $s^{(i)}(0^-) = 0$ ,  $0 \leq i \leq k - 1$ , by simply adding a suitable polynomial of degree  $\leq k - 1$  to  $s$ . Thus, if

$s = \sum \alpha_i N_i$ , then

$$\begin{aligned} \|(Es)^{(j)}\|_{L_p(\mathbb{R})} &= \left\| \sum_{i=0}^{k-2} \alpha_i N_i^{(j)} \right\|_{L_p(\mathbb{R})} \\ &\leq \sum_{i=0}^{k-2} |\alpha_i| \|N_i^{(j)}\|_{L_p(\mathbb{R})} \\ &\leq \left( \sum_{i=0}^{k-2} |\alpha_i| \right) \max_{0 \leq i \leq k-2} \|N_i^{(j)}\|_{L_p(\mathbb{R})} \\ &\leq c_9 \sum_{i=0}^{k-2} |s^{(i)}(0^+) - s^{(i)}(0^-)|, \end{aligned}$$

for all  $j$ ,  $0 \leq j \leq k - 1$ , and some constant  $c_9$  depending only on  $k$ . The latter inequality follows from considering the  $B$ -spline representation of derivatives of splines or from the work of de Boor-Fix [2]. This completes the proof of the lemma.

We also have the following

LEMMA 4.2. *Let  $k$  be a positive integer and  $1 \leq p \leq \infty$ . There exist positive numbers  $\epsilon$  and  $\delta$ , depending only on  $k$  and  $p$ , such that if  $q$  is a polynomial of degree  $\leq k$  with  $q(0) = 0$  and  $\|q\|_p \geq 1$ , there is an interval  $I \subset [0, 1]$  of length  $\epsilon$ , such that*

$$\min \{|q'(x)| : x \in I\} \geq \delta.$$

This lemma follows easily from a compactness argument by using the fact that the  $L_p[0, 1]$  norm is homogeneous and the collection of all polynomials  $q$  of degree  $\leq k$  with  $q(0) = 0$  and  $\|q\|_p = 1$  is a compact set.

Let  $C(\mathbb{R})$  be the space of all continuous functions on the real line  $\mathbb{R}$ . Hence,  $C^*(\mathbb{R})$  denotes the collection of all nondecreasing functions which are continuous on  $\mathbb{R}$ , and  $Y^*$  the nondecreasing spline functions in  $Y$ . We have the following result on monotone approximation of piecewise polynomials by monotone splines.

LEMMA 4.3. *There is a positive constant  $c_{10}$ , depending only on  $k$ , such that for every  $f \in C^*(\mathbb{R})$  whose restrictions to  $(-\infty, 0)$  and  $(0, \infty)$  are  $(k - 1)$ -st degree polynomials, there is an  $s \in Y^*$  such that  $s = f$  on  $(-\infty, -4k^2)$  and  $(4k^2, \infty)$ , and*

$$(4.2) \quad \|s - f\|_{L_p(\mathbb{R})} \leq c_{10} \sum_{i=1}^{k-2} |f^{(i)}(0^+) - f^{(i)}(0^-)|.$$

*Proof.* Let  $d = 4k^2$  and  $F$  be the collection of functions  $f \in C^*(\mathbb{R})$  whose restrictions to  $(-\infty, 0)$  and  $(0, \infty)$  are  $(k - 1)$ st degree polynomials such that  $f(0) = 0$  and  $\sum_{i=1}^{k-2} |f^{(i)}(0^+) - f^{(i)}(0^-)| \leq 1$ . 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  $\|s - f\|_{L_p[-d, d]} \leq c_{10}$  for some constant  $c_{10}$  depending only on  $k$ . We divide the proof into two cases: (i)  $\|f\|_{L_p[-d, d]} \leq \alpha$  and (ii)  $\|f\|_{L_p[-d, d]} > \alpha$ , where  $\alpha > 0$  is to be determined later. The referee has kindly informed us that cases i) and ii) are somewhat analogous to the cases handled by type 1 and type 4 intervals respectively in [3].

(i) Suppose that  $f \in F$  and  $\|f\|_{L_p[-d, d]} \leq \alpha$ . Write  $f = \sum_{-\infty}^{\infty} \alpha_i N_i$  and  $g \equiv (Tf)' = \sum_{-\infty}^{\infty} \beta_i N_{i, k-1, z}$ , where  $z = \{z_i = i\}$ . Note that  $f$  has at most a  $k$ -fold knot at 0. Since  $Tf = f$  except on  $[0, k - 1]$ , we conclude (cf. [1]) that

$$\Delta^{k-1} \beta_i = 0 \quad \text{for } i \notin [1 - k, 0],$$

where  $\Delta^{k-1}$  denotes, as usual, the  $(k - 1)$ st order forward difference operator. Hence, it follows that there are two  $(k - 2)$ nd degree polynomials  $q_1$  and  $q_2$  such that  $q_1(i) = \beta_i$  for

$i \geq 0$  and  $q_2(i) = \beta_i$  for  $i < 0$ . Since  $g(x) = f'(x) \geq 0$  for  $x \in [0, k - 1]$ , we have, in every segment  $(\beta_{l+1}, \dots, \beta_{l+k-1})$  of  $\{\beta_i\}$  with length  $k - 1$ , where  $l \geq 0$  or  $l + k - 1 < 0$ , there must exist an index  $j_l$ , where  $l + 1 \leq j_l \leq l + k - 1$ , so that  $\beta_{j_l} \geq 0$ . Furthermore, since  $q'_1$  has at most  $k - 3$  sign changes, there must be a segment  $(\beta_{l^*+1}, \dots, \beta_{l^*+2k-2})$  of nonnegative coefficients with length  $2(k - 1)$  for some  $l^*$ ,  $0 \leq l^* \leq 4(k - 1)(k - 2) < 4k^2 - 2(k - 1)$ . Similarly, there is an  $l_*$ ,  $-4k^2 \leq l_* \leq -2(k - 1)$ , so that all coefficients of the segment  $(\beta_{l_*}, \dots, \beta_{l_*+2k-3})$  are nonnegative. Let

$$h = \sum_{i=l_*+k-1}^{l^*+k-1} \beta_i N_{i,k,z}.$$

Then  $0 \leq h \leq g$  on  $(-\infty, l_* + 2k - 2) \cup (l^* + k, \infty)$  and  $h = g = (Tf)'$  on  $(l_* + 2k - 2, l^* + k - 1)$ . Since  $f$  is nondecreasing and  $Tf = f$  except on  $[0, k - 1]$ , we have

$$\int_{-\infty}^{\infty} h \geq \int_{l_*+2(k-1)}^{l^*+k-1} (Tf)' \geq 0.$$

Since the integral of  $N_{0,k-1,z}$  on  $\mathbb{R}$  is  $k - 1$ , we let

$$\gamma = \frac{1}{k - 1} \int_{-\infty}^{\infty} h$$

and set

$$s(x) = f(-d) + \int_{-d}^x (g - h + \gamma N_{0,k-1,z}).$$

Clearly,  $s \in Y^*$  and  $s = f$  on  $(-\infty, -d) \cup (d, \infty)$ , and by applying Lemma 4.1, we conclude that  $\|h - \gamma N_{0,k-1,z}\|_{L_p[-d,d]}$  and  $\|s - f\|_{L_p[-d,d]}$ , are bounded by some constant which depends only on  $k$  and  $\alpha$ .

(ii) Suppose now  $f \in F$  and  $\|f\|_{L_p[-d,d]} > \alpha$ . Since  $f(0) = 0$ , it is intuitively clear that  $f'$  is large on quite substantial sets if  $\alpha$  is very large. This allows us to modify  $Tf$  in order to guarantee that it is nondecreasing and satisfies (4.2). Indeed, by using Lemma 4.2, we can choose  $\alpha$  so large that if  $\|f\|_{L_p[-d,d]} \geq \alpha$ , then there is an integer  $l$  such that  $[l, l + k - 1] \subset [-d, d]$  and  $f'(x) \geq \beta(k - 1)$  on  $[l, l + k - 1]$  where

$$\beta = c_9 \int_{-\infty}^{\infty} \sum_{i=-k+2}^{k-2} N_{i,k-1,z}.$$

Here,  $c_9$  is the positive constant introduced in the estimate (4.1). Note that  $\alpha$  can be chosen independent of  $f$  in  $F$ . Now, set

$$\begin{aligned} s(x) &= (Tf)(x) + \int_{-\infty}^x \left[ c_9 \sum_{i=-k+2}^{k-2} N_{i,k-1,z} - \beta(2k - 4)N_{l,k-1,z} \right] \\ &\equiv (Tf)(x) + M(x) \end{aligned}$$

where  $M(x)$  is the function defined by the indefinite integral. From Lemma 4.1, we have

$$\|(Es)^{(j)}\|_{L_p(\mathbb{R})} \leq C_9 \sum_{i=0}^{k-2} |s^{(i)}(0^+) - s^{(i)}(0^-)|, \quad 0 \leq j \leq k - 1,$$

and  $1 \leq p \leq \infty$ . If  $f = \sum_{-\infty}^{\infty} \alpha_i N_{i,k,t}$ , then  $Tf - f = \sum_{i=0}^{k-2} \alpha_i N_{i,k,t}$  and  $(Tf - f)' \equiv 0$  except possibly on  $[0, k - 1]$ . By assumption,  $f$  is continuous at zero and everywhere increasing; thus,  $\sum_{i=1}^{k-2} |f^{(i)}(0^+) - f^{(i)}(0^-)| \leq 1$  so that Lemma 4.1 (with  $j = 1$  and  $p = \infty$ ) assures

that

$$(Tf)' \geq 0 \quad \text{on } (-\infty, 0) \cup (k-1, \infty)$$

$$(Tf)' \geq -c_9 \quad \text{on } [0, k-1].$$

In particular,  $c_9 \sum_{i=-k+2}^{k-2} N_{i,k-1,\mathbf{z}} \equiv 1$  on  $[0, k-1]$  and the construction of  $s$  thus guarantees  $s'(x) \geq 0$  on  $(-\infty, \infty)$  and  $s(x) = f$  except on  $[0, 4k^2]$ . Since

$$\|f - s\|_{L_p[-d,d]} \leq \|s - Tf\|_{L_p[-d,d]} + \|M\|_{L_p[-d,d]}$$

and both terms on the right are uniformly bounded for all  $f \in F$  with  $\|f\|_{L_p[-d,d]} \geq \alpha$ , we have completed the proof of the lemma.

We need another lemma to relate the error estimates in (4.2) to the  $L_p$  moduli of continuity. As an application of Lemma 3.2, we note that for every  $f \in L_p^*[-1, 1]$ , there is a nondecreasing continuous function  $g$  on  $[-1, 1]$  such that  $g$  interpolates  $f$  at  $-1$  and  $1$ , the restrictions of  $g$  on  $[-1, 0]$  and  $[0, 1]$  are polynomials of degree  $\leq j$ , and such that

$$(4.3) \quad \|(f - g)^{(i)}\|_{L_p[-1, 0]} \leq c_8 \omega(f^{(i)}, 1; [-1, 0])_p$$

and

$$(4.4) \quad \|(f - g)^{(i)}\|_{L_p[0, 1]} \leq c_8 \omega(f^{(i)}, 1; [0, 1])_p,$$

for  $i = 0, \dots, j$ . Let  $\delta^i$  be the continuous linear functional defined on  $L_p^j[-1, 0]$  and  $L_p^j[0, 1]$  by  $\delta^i f = f^{(i)}(0)$ ,  $0 \leq i \leq j-1$ . Here we are using the well known fact that  $\|f^{(i)}\|_{L_\infty[0,1]} \leq c \|f\|_{L_p^j[0,1]}$  for  $i < j$ , and the norms

$$\|f\|_{L_p^j[-1,0]} = \sum_{i=0}^{j-1} \|f^{(i)}\|_{L_p[-1,0]}$$

and

$$\|f\|_{L_p^j[0,1]} = \sum_{i=0}^{j-1} \|f^{(i)}\|_{L_p[0,1]}$$

on  $L_p^j[-1, 0]$  and  $L_p^j[0, 1]$  respectively. Hence,

$$\begin{aligned} \sum_{i=1}^{j-1} |g^{(i)}(0^+) - g^{(i)}(0^-)| &\leq \sum_{i=1}^{j-1} \{|g^{(i)}(0^+) - f^{(i)}(0)| + |g^{(i)}(0^-) - f^{(i)}(0)|\} \\ &\leq \sum_{i=1}^{j-1} (\|\delta^i\|_1 \|f - g\|_{L_p^j[-1,0]} + \|\delta^i\|_2 \|f - g\|_{L_p^j[0,1]}) \\ &= \sum_{i=1}^{j-1} \sum_{l=0}^j (\|\delta^i\| \| (f - g)^{(l)} \|_{L_p[-1,0]} + \|\delta^i\|_2 \| (f - g)^{(l)} \|_{L_p[0,1]}) \\ &\leq (j+1)c_8 \left( \sum_{i=1}^{j-1} \|\delta^i\|_1 + \|\delta^i\|_2 \right) (\omega(f^{(j)}, 1; [-1, 0])_p \\ &\quad + \omega(f^{(j)}, 1; [0, 1])_p), \end{aligned}$$

where  $\|\delta^i\|_1$  and  $\|\delta^i\|_2$  are the norms of the linear functionals  $\delta^i$  on  $L_p^j[-1, 0]$  and  $L_p^j[0, 1]$  respectively. Hence, by using the Jensen's inequality

$$(|a| + |b|)^p \leq 2^{p-1}(|a|^p + |b|^p)$$

for all  $p$ ,  $1 \leq p < \infty$ , and applying Lemma 2.3, we have the following result.

LEMMA 4.4. *Let  $1 \leq p \leq \infty$  and  $k \geq 2$ . There is a positive constant  $c_{11}$ , depending only on  $p$  and  $k$ , such that for every  $f \in L_p^*[ -1, 1 ]$  and for every  $g \in C^*[ -1, 1 ]$  whose restrictions on  $[ -1, 0 ]$  and  $[ 0, 1 ]$  are polynomials of degree  $\leq j$  and which satisfies (4.3) and (4.4), the inequality*

$$\sum_{i=1}^{j-1} |g^{(i)}(0^+) - g^{(i)}(0^-)| \leq c_{11} \omega(f^{(i)}, 1; [-1, 1])_p,$$

is satisfied for all  $1 \leq j \leq k - 1$  and  $1 \leq p \leq \infty$ .

We are now ready to prove the main result of this paper.

*Proof of Theorem 1.1.* Let  $f \in L_p^*[ 0, 1 ]$  and  $0 \leq j \leq k - 1$ . It is sufficient to consider  $j \geq 1$  and  $N > 3d$ , where  $d = 4k^2$  as defined in proof of Lemma 4.3, since (1.1) holds for all  $N = 1, 2, \dots$ , if it holds for all sufficiently large  $N$  by a standard compactness argument. Set  $\hat{f}(t) = f(t/N)$  and let  $M$  be the integer part of  $N/(3d)$ , so that  $N - 3d < 3Md \leq N$ . By applying Lemma 3.2 to each of the intervals  $I_1 = [0, 3d]$ ,  $I_2 = [3d, 6d]$ ,  $\dots$ ,  $I_{M-1} = [3d(M-2), 3d(M-1)]$ , and  $I_M = [3d(M-1), N]$ , we can find a nondecreasing continuous function  $\hat{g}$  such that the restriction of  $\hat{g}$  to each of the intervals  $I_1, \dots, I_M$  is a polynomial of degree  $\leq k - 1$  and such that

$$(4.5) \quad \|(\hat{g} - \hat{f})^{(i)}\|_{L_p[I_l]} \leq c_{12} \omega(\hat{f}^{(i)}, 1; I_l)_p,$$

for all  $l = 1, \dots, M$ ,  $1 \leq j \leq k - 1$  and  $1 \leq p \leq \infty$ , where  $c_{12}$  is a positive constant depending only on  $k$ . Note that  $c_{12}$  depends only on  $c_8$  and the lengths of intervals  $I_1, \dots, I_M$ . Hence, by applying Lemma 4.4, there exists a positive constant  $c_{13}$ , depending only on  $k, d$ , and  $p$ , such that

$$(4.6) \quad \sum_{i=1}^{j-1} |\hat{g}^{(i)}(3ld^+) - \hat{g}^{(i)}(3ld^-)| \leq c_{13} \omega(\hat{f}^{(i)}, 1; I_l \cup I_{l+1})_p$$

for all  $l = 1, \dots, M - 1$ , and  $1 \leq j \leq k - 1$ . Also, by applying Lemma 4.3 to each of the intervals  $I_l \cup I_{l+1}$ ,  $l = 1, \dots, M - 1$ , there are nondecreasing splines  $\hat{s}_l \in Y^*$  such that  $\hat{s}_l(x) = \hat{g}(x)$  for all  $x \notin [(3l - 1)d, (3l + 1)d]$  and that

$$(4.7) \quad \|\hat{s}_l - \hat{g}\|_{L_p[I_l \cup I_{l+1}]} \leq c_{10} \sum_{i=1}^{j-1} |\hat{g}^{(i)}(3ld^+) - \hat{g}^{(i)}(3ld^-)|.$$

Let  $\hat{s} \in Y^*$  such that  $\hat{s} = \hat{s}_l$  on  $I_l \cup I_{l+1}$  for all  $l = 1, \dots, M - 1$ , and  $s(t) = \hat{s}(tN)$ ,  $0 \leq t \leq 1$ . Then  $s \in \mathcal{S}^*(k, N)$  (cf. § 1). For  $p = \infty$ , we have, using (4.6),

$$\begin{aligned} \|\hat{s} - \hat{g}\|_{L_\infty(\mathbb{R})} &\leq \max_l \|\hat{s}_l - \hat{g}\|_{L_\infty(\mathbb{R})} \\ &\leq c_{13} N^{-k} \omega(f^{(j)}, 1/N)_\infty. \end{aligned}$$

This, combined with (4.5) gives

$$(4.8) \quad \|s - f\|_\infty \leq c N^{-j} \omega(f, 1/N)_\infty,$$

with  $c = c_{12} + c_{13}$ . Now let  $1 \leq p < \infty$ . We have

$$\begin{aligned} \|s - f\|_p^p &= \frac{1}{N} \sum_{l=1}^M \|\hat{s} - \hat{f}\|_{L_p[I_l]}^p \\ &\leq \frac{1}{N} \sum_{l=1}^{M-1} \|\hat{s} - \hat{f}\|_{L_p[I_l \cup I_{l+1}]}^p \\ &\leq \frac{2^{p-1}}{N} \sum_{l=1}^{M-1} (\|\hat{s} - \hat{g}\|_{L_p[I_l \cup I_{l+1}]}^p + \|\hat{g} - \hat{f}\|_{L_p[I_l \cup I_{l+1}]}^p) \end{aligned}$$

by Jensen's inequality. Hence, by using (4.5), (4.6) and (4.7), we have

$$\|s - f\|_p^p \leq c_{14} N^{-1} \sum_{l=1}^{M-1} \omega(\hat{f}^{(l)}, 1; I_l \cup I_{l+1})_p^p$$

where  $c_{14}$  depends only on  $k$  and  $p$ . By Lemma 2.3, with partition  $\{0, 3d/N, 6d/N, \dots, 3d(M-1)/N, 1\}$ , we have proved that

$$(4.9) \quad \|s - f\|_p \leq CN^{-k} \omega(f^{(j)}, 1/N)_p$$

for some constant  $C$  depending only on  $k$  and  $p$ . With (4.8) and (4.9), we have completed the proof of the theorem.

**5. Remarks.** We feel that the techniques introduced in this paper are more important than the results. In particular, by judiciously using these techniques, it may be possible to successfully attack more complicated constrained approximation problems (such as approximation of convex functions by convex splines, etc.). One problem which we are investigating and, we feel, may fall to these techniques, is the problem mentioned by De Vore [4], namely show that

$$E_{N,p}^*(f, k) \leq c\omega_k(f, 1/N)_p,$$

where  $\omega_k$  is the  $k$ th order  $L_p$  modulus of smoothness. This inequality would yield inverse theorems, for instance, of the type considered by K. Scherer [11].

The analogous problem in monotone approximation by polynomials in  $L_p$  is also of interest. In the case  $p = \infty$ , Lorentz [7] and Lorentz and Zeller [8, 9] have studied this problem for  $j = 0, 1$  and De Vore [5] has obtained the most complete results to date.

#### REFERENCES

- [1] C. DE BOOR, *On calculating with B-splines*, J. Approximation Theory, 6 (1972), pp. 50–62.
- [2] C. DE BOOR AND G. J. FIX, *Spline approximation by quasiinterpolants*, Ibid., 8 (1973), pp. 19–45.
- [3] R. DE VORE, *Degree of approximation*, Approximation Theory II, G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 117–162.
- [4] ———, *Monotone approximation by splines*, this Journal, 8 (1977), pp. 891–905.
- [5] ———, *Monotone approximation by polynomials*, this Journal, 8 (1977), pp. 906–921.
- [6] H. JOHNEN, *Inequalities connected with the moduli of smoothness*, Math. Vestnik, 9, pp. 289–303.
- [7] G. G. LORENTZ, *Monotone approximation*, Inequalities, III, O. Shisha, ed., Academic Press, New York, 1972, pp. 201–215.
- [8] G. G. LORENTZ AND K. ZELLER, *Degree of approximation by monotone polynomials I*, J. Approximation Theory, 1 (1968), pp. 501–504.
- [9] ———, *Degree of approximation by monotone polynomials II*, Ibid., 2 (1969), pp. 265–269.
- [10] J. PEETRE, *A Theory of Interpolation of Normed Spaces*, Lecture Notes, Brazilia, 1963.
- [11] K. SCHERER, *On best approximation of continuous functions by splines*, SIAM J. Numer. Anal., 7 (1970), pp. 418–423.

## A SEMIGROUP ON THE SPACE OF COMPACT CONVEX BODIES\*

STEPHEN J. WILLSON†

**Abstract.** Let  $C_0$  denote the space of all compact convex subsets of  $R^n$  with nonempty interior. Give  $C_0$  its natural topology. It is proved that any continuous nonnegative function  $g$  on the unit sphere may be used to define a continuous semigroup on  $C_0$ ; i.e., a continuous map  $F: C_0 \times [0, \infty) \rightarrow C_0$  so  $F(F(X, s), t) = F(X, s + t)$  and  $F(X, 0) = X$ . If  $g$  is strictly positive, it is proved that there is a  $W$  in  $C_0$  depending only on  $g$  so for each  $X$  and for  $t$  large it follows  $F(X, t)$  is approximately  $tW$ . Indications are given for applications to crystal growth.

**1. Introduction.** Let  $C_0$  denote the set of compact convex subsets of  $R^n$  with nonempty interior. Give  $C_0$  its natural metric, that of the Blaschke selection theorem. If  $g$  is any continuous nonnegative function on the unit sphere  $S^{n-1}$ , we define a map  $F: C_0 \times [0, \infty) \rightarrow C_0$ , denoted  $(X, t) \rightarrow F_t X$ , by the formula

$$F_t X = \bigcap_{v \in S^{n-1}} H(v, h_X(v) + tg(v)).$$

Here  $h_X(v)$  is the support function of  $X$ ;  $H(v, a(v))$  denotes the closed half space

$$H(v, a(v)) = \{w \in R^n : \langle v, w \rangle \leq a(v)\}.$$

Thus  $F_t X$  is obtained by moving each supporting hyperplane of  $X$  an amount  $tg(v)$  away from  $X$ . (More precise definitions are given later.)

Our major results are as follows:

**THEOREM 5.7.** *F is a semigroup on  $C_0$ . More precisely, F is continuous, and if  $s, t \geq 0$  and  $X \in C_0$ , then  $F_s F_t X = F_{s+t} X$ .*

**THEOREM 6.1.** (Rough statement). *Suppose  $g(v)$  is always positive. There exists  $W \in C_0$ , depending only on  $g$ , so that if  $X \in C_0$ , then  $F_t X$  approximately equals  $tW$  for large  $t$ .*

The difficulty in proving Theorem 5.7 is that the support function  $h_{F_t X}(v)$  of  $F_t X$  need not equal  $h_X(v) + tg(v)$ ; it is quite possible that other hyperplanes intervene to cut off  $F_t X$  well before it reaches the hyperplane bounding  $H(v, h_X(v) + tg(v))$ . The definitions of  $F_s F_t X$  and  $F_{s+t} X$  then become quite different.

These results trivialize in the special case where there exists  $W \in C_0$  such that  $g(v) = h_W(v)$ . We then obtain  $F_t X = X + tW$ , using vector addition, and Theorems 5.7 and 6.1 become immediate. The interest in the above results thus lies in the case where  $g$  is not a support function.

The methods of the paper actually prove more than is indicated above. If  $g$  is merely assumed continuous but possibly negative, then for fixed  $X \in C_0$ ,  $F_t X$  will still be defined for positive  $t$  sufficiently near 0. In fact, we still obtain the formula of Theorem 5.7 for sufficiently small positive  $s$  and  $t$ ; this result is Theorem 5.6.

This research was motivated by analogy with the growth of physical crystals. Suppose that for a certain kind of crystal in a certain medium any facet with outward unit normal vector  $v$  has growth rate  $g(v)$ . If  $X$  is an arbitrary "seed crystal", one would guess that after time  $t$  the crystal has shape  $F_t X$ . Theorems 5.6 and 5.7 say that this mathematical model is plausible, even if  $g(v) < 0$  and the crystal is dissolving.

\* Received by the editors April 12, 1978, and in revised form July 10, 1979.

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

In 1901 G. Wulff [9] described for physical crystals a characteristic “equilibrium shape”  $W$ : for certain crystals  $W$  is a rectangular parallelepiped, for others an octahedron, etc. The shape  $W$  is defined in terms of the surface free energy function, which is assumed to be proportional to  $g(v)$ . The physics literature has focused on whether  $W$  minimizes the total surface free energy among all convex bodies of the same volume. Assuming this result, physicists have inferred convergence to  $W$  by means of “surface tension.” This paper gives an alternate, more direct proof of the convergence in Theorem 6.1. For a discussion, the reader may refer to Burton et al. [2], von Laue [5], and Herring [4].

It is well-known that the characteristic equilibrium shape  $W$  for real crystals tends to be polyhedral. In Proposition 6.6 we give conditions on  $g$  sufficient to ensure that  $W$  be a convex polytope.

In [8] the author has obtained a discrete analogue of Theorem 6.1 using cellular automata. The discrete model has the advantage of applying to nonconvex sets but the disadvantage that one must treat bothersome “edge effects.” The assumptions in the current paper imply that we are ignoring those edge effects.

In § 2 we fix notations. In § 3 we study convex sets given in the form  $\cap H(v, a(v))$ . In § 4 we study the continuity of  $F$  and in § 5 the semigroup property. The last section treats convergence to  $W$ .

**2. Notations.** Let  $R^n$  denote  $n$ -dimensional Euclidean space, equipped with the standard inner product denoted  $\langle x, y \rangle$  for  $x, y \in R^n$ . For  $x \in R^n$ , let  $|x| = \sqrt{\langle x, x \rangle}$ . Let  $S^{n-1} = \{v \in R^n : |v| = 1\}$  denote the unit sphere. If  $\varepsilon > 0$ , let  $B(\varepsilon) = \{v \in R^n : |v| < \varepsilon\}$ . If  $v \in R^n$  and  $a \in R$ , set  $H(v, a) = \{x \in R^n : \langle x, v \rangle \leq a\}$ ; if  $v \neq 0$  then  $H(v, a)$  is a closed half space bounded by  $\pi(v, a) = \{x \in R^n : \langle x, v \rangle = a\}$ .

If  $X$  is a nonempty compact convex subset of  $R^n$ , let  $h_X$  be the support function of  $X$ :  $h_X(v) = \sup_{x \in X} \langle x, v \rangle$ . It is well-known that  $h_X$  is a convex function, continuous and positively homogeneous. It is also known that  $X = \cap_{v \in S^{n-1}} H(v, h_X(v))$ .

Let  $C$  denote the set of nonempty compact convex subsets of  $R^n$  and let  $C_0$  denote those members of  $C$  with nonempty interior.

If  $X$  and  $Y$  are nonempty convex sets and  $a, b \in R$ , let  $aX + bY = \{ax + by : x \in X, b \in Y\}$ . Then  $aX + bY$  is a convex subset of  $R^n$ . Observe that we have the following identities: (i)  $h_{aX} = ah_X$  if  $a \geq 0, X \in C$ ; (ii)  $h_{X+Y} = h_X + h_Y$ , if  $X, Y \in C$ . If  $y \in R^n$ , then  $H(v, a) + y = H(v, a + \langle v, y \rangle)$ .

If  $X \in C$ , then  $\text{Bd } X$  denotes its boundary,  $\text{Int } X$  denotes its interior, and  $\text{Ext } X$  its extreme points. If  $X \subset R^n$ , then  $\mathcal{H}(X)$  denotes the convex hull of  $X$ ,  $\sim X$  its complement, and  $\text{cl } X$  its closure. The polar of  $X$  is  $X^* = \{y \in R^n : \langle x, y \rangle \leq 1 \text{ for all } x \in X\}$ . It is known that  $X^{**} \equiv (X^*)^* = \mathcal{H}(X \cup \{0\})$ ; hence if  $0 \in X \in C, X^{**} = X$ .

A good general reference for the above concepts is Eggleston [3].

**3. Dual descriptions.** Let  $a(v)$  be a continuous function from  $S^{n-1}$  to  $R$ . We are concerned in this section with subsets of  $R^n$  of the form  $X = \cap_{v \in S^{n-1}} H(v, a(v))$ . The function  $a(v)$  gives a “dual description” of  $X$ . Note that either  $X$  is empty or  $X$  is a compact convex subset of  $R^n$ .

If  $X$  is nonempty, clearly  $h_X(v) \leq a(v)$  for each  $v \in S^{n-1}$ . It is, however, quite possible that  $h_X(v) < a(v)$ . Similarly there might exist a closed proper subset  $D$  of  $S^{n-1}$  such that  $X = \cap_{v \in D} H(v, a(v))$ , in which case any  $v \in S^{n-1} - D$  “does not really affect  $X$ .” We are thus led to the problem of computing the support function  $h_X$  and of distinguishing in some sense the minimal possible  $D$ . The first problem is solved in Proposition 3.1 and the second in Proposition 3.2.

Given the continuous function  $a(v)$  on  $S^{n-1}$ , we may extend  $a(v)$  over  $R^n$  by



requiring that the extension be positively homogeneous; i.e., we may define  $a(tv) = ta(v)$  for  $t \in [0, \infty)$ ,  $v \in S^{n-1}$ . We shall usually assume that this extension has been made. Note then that  $X = \bigcap_{v \in S^{n-1}} H(v, a(v)) = \bigcap_{v \in R^n} H(v, a(v))$  and we shall usually denote this set  $\bigcap H(v, a(v))$ .

If  $y \in R^n$ , then  $X + y = \bigcap H(v, a(v) + \langle v, y \rangle)$ . Hence if  $X \neq \emptyset$  we shall often be free to translate  $X$ , for example, so as to assume  $0 \in X$  and  $a(v) \geq 0$ . If  $X$  has nonempty interior, we may often translate  $X$  to assume  $0 \in \text{Int } X$  and  $a(v) > 0$ .

**PROPOSITION 3.1.** *Assume  $a(v)$  is continuous and nonnegative on  $S^{n-1}$  (and extended to be positively homogeneous on  $R^n$ ). Let  $X = \bigcap H(v, a(v))$ ,  $K = \{v \in R^n : a(v) \leq 1\}$ . Define  $c(v) \in R \cup \{\infty\}$  by  $c(v) = \sup \{t : tv \in \mathcal{H}(K)\}$ . Then*

- (i)  $\mathcal{H}(K) = X^*$  and  $X = \mathcal{H}(K)^*$ .
- (ii) If  $a(v) = h_X(v)$  for all  $v$ , then  $K$  is convex.
- (iii)  $c(v) > 0$  for all  $v$  and  $h_X(v) = 1/c(v)$ .  
(Here  $1/\infty = 0$ .)
- (iv) If  $a(v)$  is a positively homogeneous convex function, then  $a(v) = h_X(v)$  for all  $v$ .

*Proof.* (i) If  $v \in K$  and  $x \in X$ , then  $\langle v, x \rangle \leq h_X(v) \leq a(v) \leq 1$ ; hence  $K \subset X^*$ . Since  $X^*$  is convex, it follows  $\mathcal{H}(K) \subset X^*$ .

Conversely, if  $w \in K^*$  and  $v \in R^n$ , we show  $\langle w, v \rangle \leq a(v)$ : if  $a(v) > 0$ , then  $\langle w, v \rangle = \langle w, v/a(v) \rangle a(v) \leq a(v)$  since  $v/a(v) \in K$ ; if  $a(v) = 0$  then for all  $t > 0$ ,  $tv \in K$ , so  $\langle w, tv \rangle \leq 1$ ,  $\langle w, v \rangle \leq 1/t$ , and  $\langle w, v \rangle \leq 0$  by letting  $t$  go to  $\infty$ . Thus  $w \in X$  and  $K^* \subset X$ . By duality  $X^* \subset K^{**} = \mathcal{H}(K)$ . Hence  $X^* = \mathcal{H}(K)$  and (i) is proved.

(ii) Since  $a(v) = h_X(v)$ ,  $a(v)$  is a convex function, whence  $K$  is convex, for example, by Rockafellar [6, p. 29].

(iii) Since  $K \subset \mathcal{H}(K)$ ,  $c(v) > 0$ . Observe that  $c(v)$  is the reciprocal of the ‘‘distance function’’ of  $\mathcal{H}(K) = X^*$ , so (iii) follows from Eggleston [3, p. 55].

(iv) Since  $a(v)$  is convex,  $K = \mathcal{H}(K)$ . Thus  $c(v) = \sup \{t : tv \in K\} = \sup \{t : a(tv) \leq 1\} = \sup \{t : ta(v) \leq 1\} = 1/a(v)$ . From (iii) it follows  $h_X(v) = a(v)$ .  $\square$

**PROPOSITION 3.2.** *Assume  $a(v)$  is continuous and strictly positive on  $S^{n-1}$ . Let  $X$  and  $K$  be as in Proposition 3.1, so that  $K$  is compact and  $\mathcal{H}(K)$  is a compact convex set. Define  $A = \{y/|y| : y \in \text{Ext } \mathcal{H}(K)\}$ . Then*

- (i)  $X = \bigcap_{v \in A} H(v, a(v))$ .
- (ii) If  $v \in A$ ,  $h_X(v) = a(v)$ .
- (iii) If  $E$  is a closed subset of  $S^{n-1}$  which does not contain  $A$ , then  $X \neq \bigcap_{v \in E} H(v, a(v))$ .

*Proof.* (i) If  $y \in \text{Ext } \mathcal{H}(K)$ , then  $y \in K$  and  $|y| = 1/a(y/|y|)$ . Since  $K$  is compact,  $\mathcal{H}(K) = \mathcal{H}(\text{Ext } \mathcal{H}(K))$ . By Proposition 3.1,  $X = \mathcal{H}(K)^* = (\mathcal{H}(\text{Ext } \mathcal{H}(K)))^* = (\text{Ext } \mathcal{H}(K))^* = \bigcap_{y \in \text{Ext } \mathcal{H}(K)} H(y, 1) = \bigcap_{y \in \text{Ext } \mathcal{H}(K)} H(y/|y|, 1/|y|) = \bigcap_{y \in \text{Ext } \mathcal{H}(K)} H(y/|y|, a(y/|y|)) = \bigcap_{v \in A} H(v, a(v))$ .

(ii) Follows from Proposition 3.1 (iii).

(iii) Suppose  $v_0 \in A$ ,  $v_0 \notin E$ . Assume  $X = \bigcap_{v \in E} H(v, a(v))$ . Then  $X = \bigcap_{v \in E} H(v/a(v), 1) = \{v/a(v) : v \in E\}^*$ , so  $\mathcal{H}(K) = X^* = \mathcal{H}(\{v/a(v) : v \in E\} \cup \{0\})$  by Eggleston [3, p. 25]. Thus every extreme point of  $\mathcal{H}(K)$  is of form  $v/a(v)$  for some  $v \in E$ . In particular,  $v_0/a(v_0)$  is of this form, so  $v_0 \in E$ , a contradiction.  $\square$

**COROLLARY 3.3.** *Make the same assumptions as in Proposition 3.2. The smallest closed subset  $E$  of  $S^{n-1}$  such that  $X = \bigcap_{v \in E} H(v, a(v))$  is  $E = \text{cl } A$ .*

**COROLLARY 3.4.** *Assume  $a(v)$  is strictly positive. Let  $X = \bigcap H(v, a(v))$  and let  $D = \{v \in S^{n-1} : a(v) = h_X(v)\}$ . Then  $X = \bigcap_{v \in D} H(v, a(v))$ .*

*Remark.* In Corollary 3.4 we cannot omit the hypothesis that  $a(v)$  be strictly positive. For example, if  $n = 2$  and  $a((\cos \theta, \sin \theta)) = \cos^2 \theta$ , then we can compute  $X = \{0\}$ ,  $D = \{(0, \pm 1)\}$ , and  $\bigcap_{v \in D} H(v, 0)$  is the entire real axis.

We may, however, strengthen Corollary 3.4 slightly as follows:

**PROPOSITION 3.5.** *Suppose  $a(v)$  is continuous on  $S^{n-1}$ . Let  $X = \bigcap H(v, a(v))$  and suppose  $X \neq \emptyset$ . Set  $D = \{v \in S^{n-1} : h_X(v) = a(v)\}$ , set  $Y = \bigcap_{v \in D} H(v, a(v))$ , and assume  $\text{Int } Y \neq \emptyset$ . Then  $X = Y$ .*

*Proof.* If  $X$  has nonempty interior, by translation we may assume  $0 \in \text{Int } X$ , and then Corollary 3.4 implies  $X = Y$ . Thus we may assume  $\text{Int } X = \emptyset$ , and so we may choose a point  $z \in \text{Int } Y \cap \sim X$ . Since  $X$  is closed, we may find the point  $x$  of  $X$  closest to  $z$ . Since  $X \subseteq Y$ , it follows that the closed line segment joining  $x$  and  $z$  is in  $Y$  by convexity; and all points of that line segment except for  $x$  itself lie in  $\sim X$ . By translation we may assume  $x = 0$  so  $\lambda z \in \sim X$  for  $0 < \lambda \leq 1$ . Since  $0 \in X$ ,  $a(v) \geq 0$  for all  $v \in S^{n-1}$ .

Let  $L = \{v \in S^{n-1} : \langle v, z \rangle \geq 0\}$ . I claim that  $a(v) > 0$  for all  $v \in L$ . To see this, since  $a(v) \geq 0$ , we assume  $a(v) = 0$  for some  $v \in L$ . Then  $0 \leq h_X(v) \leq a(v) = 0$  so  $a(v) = h_X(v) = 0$  and  $v \in D$ . But since  $z \in \text{Int } Y$ ,  $\langle v, z \rangle < a(v) = 0$ , contradicting that  $v \in L$ .

Since  $a(v)$  is continuous and positive on the compact set  $L$ , there exists a positive number  $\delta$  so  $a(v) \geq \delta$  for all  $v \in L$ . Let  $\lambda$  be the minimum of 1 and  $\delta/|z|$ . Then for  $v \in L$ ,  $\lambda z \in H(v, a(v))$  since  $\langle \lambda z, v \rangle \leq \delta \langle z/|z|, v \rangle \leq \delta \leq a(v)$ ; and for  $v \in S^{n-1} - L$ ,  $\lambda z \in H(v, a(v))$  since  $\langle \lambda z, v \rangle < 0 \leq a(v)$ . It follows  $\lambda z \in X$  and  $0 < \lambda \leq 1$ , a contradiction. This proves Proposition 3.5.  $\square$

**4. Convergence.** Suppose  $a_i(v)$  converges to  $a(v)$ . In this section we study whether  $\bigcap H(v, a_i(v))$  converges to  $\bigcap H(v, a(v))$ .

Let  $\mathcal{B}$  denote the set of compact nonempty subsets of  $R^n$ . If  $X \in \mathcal{B}$  and  $\varepsilon > 0$ , let  $U(X, \varepsilon) = \{y \in R^n : \text{there exists } x \in X \text{ with } |y - x| < \varepsilon\}$ . If  $X, Y \in \mathcal{B}$ , let  $\delta_1 = \inf \{\delta > 0 : Y \subset U(X, \delta)\}$  and  $\delta_2 = \inf \{\delta > 0 : X \subset U(Y, \delta)\}$ . Define  $\Delta(X, Y) = \delta_1 + \delta_2$ . It is well-known (See Eggleston [3, p. 60]) that  $\Delta$  defines a metric on  $\mathcal{B}$ . We may give  $C$  the topology induced as a subspace of  $\mathcal{B}$ . If  $X_i, X \in \mathcal{B}$  and  $X_i$  converges to  $X$  in this topology, we write  $X_i \rightarrow X$ . The Blaschke selection theorem (See Eggleston [3, p. 64].) applies to  $C$  with this topology.

The major theorem of this section is the following:

**THEOREM 4.1.** *Let  $a_i(v)$  be a sequence of continuous functions on  $S^{n-1}$  which converges uniformly to the continuous function  $a(v)$ . Let  $X_i = \bigcap H(v, a_i(v))$  and  $X = \bigcap H(v, a(v))$ . If  $X$  has nonempty interior, then for sufficiently large  $i$ ,  $\text{Int } X_i \neq \emptyset$ , and  $X_i$  converges to  $X$  in  $C$ .*

We remark that if  $X$  has no interior, then  $X_i$  need not converge to  $X$ . For example, if  $a_i(v) \equiv -1/i$  then  $X_i = \emptyset$  for all  $i$  while  $X = \{0\}$ . For a less trivial example let  $n = 2$  and let  $v_\theta$  be the unit vector in the direction of the angle  $\theta$ . Define  $a(v_\theta) = \max(\cos \theta, 0)$ . Choose a sequence  $a_i(v_\theta)$  for  $-\pi \leq \theta \leq \pi$  so that (1)  $a_i(v_\theta) = a(v_\theta)$  unless  $-\pi/2 \leq \theta \leq -(\pi/2) + 2^{-i}$  or  $(\pi/2) - 2^{-i} \leq \theta \leq \pi/2$ ; (2)  $a_i(v_\theta) = 0$  if  $-\pi/2 \leq \theta \leq -(\pi/2) + 2^{-i-1}$  or  $(\pi/2) - 2^{-i-1} \leq \theta \leq \pi/2$ ; (3)  $a_i(v_\theta)$  is continuous and always satisfies  $0 \leq a_i(v_\theta) \leq a(v_\theta)$ . Then  $a_i(v_\theta)$  converges uniformly to  $a(v_\theta)$ . On the other hand it is easy to verify that  $X_i = \{0\}$  for all  $i$  while  $X$  is the unit interval  $[0, 1]$  on the  $x$ -axis. If we add suitable positive constants to  $a_i(v_\theta)$ , we obtain an example where  $0 \in \text{Int } X_i$ ,  $X_i \rightarrow \{0\}$ ,  $X = [0, 1]$ .

In Corollary 4.3 below there is a weaker result which applies even if  $X$  has no interior.

The proof of Theorem 4.1 will occupy the remainder of this section.

**LEMMA 4.2.** *Let  $X_i, X \in C$ . Then  $X_i \rightarrow X$  in  $C$  if and only if  $h_{X_i}$  converges uniformly to  $h_X$  on  $S^{n-1}$ .*

*Proof.* This is a result of Bonnesen and Fenchel [1, p.35].  $\square$

**COROLLARY 4.3.** *Let  $a_i(v)$  be a sequence of continuous functions on  $S^{n-1}$  which converges pointwise to the function  $a(v)$ . Let  $X_i = \bigcap H(v, a_i(v))$  and  $X = \bigcap H(v, a(v))$ . Suppose  $X_i \rightarrow Y$  in  $C$ . Then  $Y \subseteq X$ .*

*Proof.* For  $v \in S^{n-1}$ ,  $h_{X_i}(v) \leq a_i(v)$ . Applying Lemma 4.2 and taking limits, we see  $h_Y(v) \leq a(v)$  for each  $v \in S^{n-1}$ . Hence  $Y \subseteq X$ .  $\square$

The following two results are well-known.

(4.4) Let  $X_i \rightarrow X$  in  $\mathcal{B}$ . Then  $\mathcal{H}(X_i) \rightarrow \mathcal{H}(X)$  in  $C$ .

(4.5) Let  $X_i \rightarrow X$  in  $C$ . Suppose  $0 \in \text{Int } X$ . Then for sufficiently large  $i$ ,  $0 \in \text{Int } X_i$ , and  $X_i^* \rightarrow X^*$ .

*Proof of Theorem 4.1.* We may assume by translation if necessary that  $0 \in \text{Int } X$ . Then  $a(v)$  has a strictly positive lower bound. Since  $a_i(v)$  converges uniformly to  $a(v)$ , it follows that for sufficiently large  $i$ ,  $a_i(v)$  is positive for all  $v \in S^{n-1}$ . Hence  $0 \in \text{Int } X_i$  for sufficiently large  $i$ , and we shall assume henceforth  $0 \in \text{Int } X_i$  for all  $i$ .

Let  $K_i = \{v/a_i(v) : v \in S^{n-1}\}$ ,  $K = \{v/a(v) : v \in S^{n-1}\}$ . Then  $K_i, K \in \mathcal{B}$  and by uniform convergence of  $a_i(v)$  to  $a(v)$  it follows  $K_i \rightarrow K$  in  $\mathcal{B}$ . By (4.4),  $\mathcal{H}(K_i) \rightarrow \mathcal{H}(K)$  in  $C$ . But  $0 \in \text{Int } \mathcal{H}(K)$  since  $a(v)$  is bounded above on  $S^{n-1}$ . Hence (4.5) implies  $\mathcal{H}(K_i)^* \rightarrow \mathcal{H}(K)^*$ . By (3.1)  $X_i = \mathcal{H}(K_i)^*$  and  $X = \mathcal{H}(K)^*$ . The theorem follows.  $\square$

**5. The semigroup property.**

**DEFINITION.** Let  $g$  be a fixed continuous real-valued function on  $S^{n-1}$ . If  $t \in R$  and  $X$  is a compact convex subset of  $R^n$ , define

$$F_t X = \bigcap H(v, h_X(v) + tg(v)).$$

This formula gives a dual description of  $F_t X$ . Either  $F_t X = \emptyset$  or  $F_t X$  is a compact convex set. Note that if  $y \in R^n$ , then  $F_t(X + y) = (F_t X) + y$ .

Theorem 4.1 has the following consequence.

**THEOREM 5.1.** *If  $X_i \rightarrow X$  in  $C$ ,  $t_i \rightarrow t$  in  $R$ , and  $F_{t_i} X_i$  has nonempty interior, then  $\text{Int } F_{t_i} X_i \neq \emptyset$  for large  $i$ , and  $F_{t_i} X_i \rightarrow F_t X$  in  $C$ .*

*Proof.* By Lemma 4.2 and the boundedness of  $g$  on  $S^{n-1}$ ,  $h_{X_i}(v) + t_i g(v)$  converges uniformly to  $h_X(v) + tg(v)$  on  $S^{n-1}$ . The result then follows from Theorem 4.1.  $\square$

In the special case where  $g(v) > 0$  for all  $v \in S^{n-1}$  and  $t > 0$ , it is easy to see that  $F_t X$  has nonempty interior, so Theorem 5.1 applies. In another extreme case we have the following result.

**PROPOSITION 5.2.** *Suppose  $g(v) \leq 0$  for all  $v \in S^{n-1}$ . Suppose  $s > 0$ ,  $X \in C$ , and  $F_s X \neq \emptyset$ . Then, if  $t$  increases to the limit  $s$ , it follows  $F_t X \rightarrow F_s X$ .*

*Proof.* If  $0 \leq t \leq s$ , then  $h_X(v) + sg(v) \leq h_X(v) + tg(v)$  for all  $v \in S^{n-1}$ . Hence  $F_s X \subseteq F_t X$  and  $F_t X \neq \emptyset$ . Suppose  $F_t X$  does not converge to  $F_s X$  as  $t$  increases to  $s$ . By the Blaschke selection theorem we may find a sequence  $t_i$  increasing to  $s$  so  $F_{t_i} X$  converges to some  $Y \in C$ , where  $Y \neq F_s X$ . Since  $F_{t_i} X \supseteq F_s X$  it follows  $Y \supseteq F_s X$ . But by Corollary 4.3  $Y \subseteq F_s X$ . Hence  $Y = F_s X$ , a contradiction. This proves the proposition.  $\square$

The major result of this section asserts that  $F_s F_t X = F_{s+t} X$  when  $s, t \geq 0$  and  $\text{Int } F_{s+t} X \neq \emptyset$ . The inclusion  $F_s F_t X \subseteq F_{s+t} X$  is trivial. For the opposite inclusion we require some intermediate results.

**PROPOSITION 5.3.** *Let  $X \in C$ . If  $0 \leq s \leq t$  and  $F_t X \neq \emptyset$ , then  $F_s X \neq \emptyset$ . Moreover, if  $0 < s \leq t$  and  $\text{Int } F_t X \neq \emptyset$ , then  $\text{Int } F_s X \neq \emptyset$ .*

*Proof.* Let  $a, b > 0$ . By translation, if necessary, we may assume  $0 \in X$ , so  $X \supseteq (b/(a+b))X$ . Then

$$\begin{aligned} F_b(X) &\supseteq F_b((b/(a+b))X) = \bigcap H(v, h_{(b/(a+b))X}(v) + bg(v)) \\ &= \bigcap H(v, (b/(a+b))h_X(v) + bg(v)) \\ &= (b/(a+b))[\bigcap H(v, h_X(v) + (a+b)g(v))] \\ &= (b/(a+b))F_{a+b}X. \end{aligned}$$

Hence  $F_{a+b}X \neq \emptyset$  implies  $F_bX \neq \emptyset$ ; and  $\text{Int } F_{a+b}X \neq \emptyset$  implies  $\text{Int } F_bX \neq \emptyset$ . The result follows.  $\square$

**THEOREM 5.4.** *Let  $X$  be a compact convex subset of  $R^n$  and let  $A = \{t \geq 0: F_tX \neq \emptyset\}$ . Then  $A$  is convex, and for any fixed  $v \in R^n$ , the function on  $A$  which takes  $t$  to  $h_{F_tX}(v)$  is concave.*

*Proof.*  $A$  is convex by Proposition 5.3. Fix  $t \in A$  and let  $k_t(v)$  be the ‘‘convex hull’’ of the function which takes  $v$  to  $h_X(v) + tg(v)$ ; i.e.,  $k_t(v)$  is the greatest convex function on  $R^n$  majorized by  $h_X(v) + tg(v)$ . By Rockafellar [6, p. 36] we obtain

$$(*) \quad k_t(v) = \inf \left\{ \sum \lambda_i (h_X(v_i) + tg(v_i)) : v = \lambda_1 v_1 + \dots + \lambda_m v_m, \lambda_i \geq 0, \lambda_1 + \dots + \lambda_m = 1 \right\},$$

where the functions are all regarded as defined on all of  $R^n$ . It follows that for fixed  $v$ , the function taking  $t$  to  $k_t(v)$  is an infimum of a family of affine functions of  $t$  and is therefore concave. (See Rockafellar [6; p. 35].)

To complete the proof of Theorem 5.4, we show  $k_t(v) = h_{F_tX}(v)$ . But from (\*), since  $h_X(v)$  and  $g(v)$  are positively homogeneous, it follows that  $k_t(v)$  is positively homogeneous. By definition,  $h_{F_tX}(v) \leq k_t(v)$  since  $h_{F_tX}(v)$  is convex. Hence  $F_tX \subseteq \bigcap H(v, k_t(v)) \subseteq \bigcap H(v, h_X(v) + tg(v)) = F_tX$ . It follows  $F_tX = \bigcap H(v, k_t(v))$  so  $k_t(v) = h_{F_tX}(v)$  by Proposition 3.1(iv).  $\square$

**COROLLARY 5.5.** *Let  $s$  and  $t$  be positive real numbers. Assume  $X \in C$  and  $F_sF_tX \neq \emptyset$ . Let*

$$E = \{v \in S^{n-1} : h_{F_tX}(v) = h_X(v) + tg(v)\},$$

$$D = \{v \in S^{n-1} : h_{F_sF_tX}(v) = h_{F_tX}(v) + sg(v)\}.$$

Then  $D \subset E$ .

*Proof.* Note  $F_tX \neq \emptyset$ , so  $E$  and  $D$  are defined. Write  $t = (s/(s+t))0 + (t/(s+t))(s+t)$ ; by Theorem 5.4 for any  $v$  it follows that  $(s/(s+t))h_X(v) + (t/(s+t))h_{F_s+tX}(v) \leq h_{F_tX}(v)$ , whence  $(t/s)(h_{F_s+tX}(v) - h_{F_tX}(v)) \leq h_{F_tX}(v) - h_X(v)$ .

Now suppose  $v \in D$ . Since  $F_sF_tX \subseteq F_{s+t}X$  we obtain  $tg(v) = (t/s)sg(v) = (t/s)(h_{F_sF_tX}(v) - h_{F_tX}(v)) \leq (t/s)(h_{F_s+tX}(v) - h_{F_tX}(v)) \leq h_{F_tX}(v) - h_X(v)$ . Thus  $h_{F_tX}(v) \geq h_X(v) + tg(v)$ , and since the opposite inclusion is trivial it follows  $v \in E$ .  $\square$

**THEOREM 5.6.** *Let  $s$  and  $t$  be nonnegative real numbers. Let  $X \in C$  and assume  $F_{s+t}X$  has nonempty interior. Then  $F_sF_tX = F_{s+t}X$ .*

*Proof.* The result is immediate if either  $s = 0$  or  $t = 0$ , so we shall assume both are positive. It is trivial that  $F_sF_tX \subseteq F_{s+t}X$ , so we need only prove the opposite inclusion.

We first complete the proof under the additional hypothesis that  $F_sF_tX \neq \emptyset$ : Since  $F_sF_tX \neq \emptyset$ ,  $E$  and  $D$  may be defined as in Corollary 5.5. Then

$$F_{s+t}X = \bigcap_{v \in S^{n-1}} H(v, h_X(v) + (s+t)g(v)) \subseteq \bigcap_{v \in E} H(v, (h_X(v) + tg(v)) + sg(v))$$

$$= \bigcap_{v \in E} H(v, h_{F_tX}(v) + sg(v))$$

$$\subseteq \bigcap_{v \in D} H(v, h_{F_tX}(v) + sg(v)) \quad [\text{since } D \subseteq E \text{ by Corollary 5.5}].$$

Letting  $a(v) = h_{F_tX}(v) + sg(v)$ , we note  $\bigcap_{v \in D} H(v, a(v))$  has nonempty interior since it contains  $\text{Int } F_{s+t}X$ . Hence by (3.5),  $\bigcap_{v \in D} H(v, a(v)) = \bigcap_{v \in S^{n-1}} H(v, a(v)) = F_sF_tX$ , so  $F_{s+t}X \subseteq F_sF_tX$ .

Thus Theorem 5.6 is true if  $F_sF_tX \neq \emptyset$ . We now show that  $F_sF_tX \neq \emptyset$  under the hypotheses of Theorem 5.6. Fix  $t > 0$ , and let  $J = \{a \in R : 0 \leq a \leq s \text{ and } F_aF_tX \neq \emptyset\}$ . We shall show that  $J$  is nonempty, open, and closed in  $[0, s]$ ; it will then follow by connectedness that  $J = [0, s]$  and the proof of Theorem 5.6 will be complete.

By Proposition 5.3,  $\text{Int } F_t X \neq \emptyset$ , so  $0 \in J$ . If  $a \in J$ , then  $\text{Int } F_{a+t} X \neq \emptyset$  by Proposition 5.3, so  $F_a F_t X = F_{a+t} X$  by the portion of Theorem 5.6 already proved. Since  $h_{F_t X}(v) + b g(v)$  converges uniformly to  $h_{F_t X}(v) + a g(v)$  as  $b \rightarrow a$ , it follows from Theorem 4.1 that  $\text{Int } F_b F_t X \neq \emptyset$  for  $b$  close to  $a$ . Hence  $J$  is open. Finally, if  $a_i \rightarrow a$  with  $a_i \in J$  and  $x_i \in F_{a_i} F_t X$ , then the sequence  $x_i$  has a limit point which is easily seen to lie in  $F_a F_t X$ . Thus  $J$  is closed and Theorem 5.6 is proved.  $\square$

*Remark.* If  $s < 0, t > 0, s + t \geq 0, g \geq 0$ , it need not follow that  $F_s F_t X = F_{s+t} X$  even if both have interior. Examples are easy to construct in the plane where  $X \neq Y$  but  $F_1 X = F_1 Y$ .

We may summarize part of this section in terms of semigroups.

**DEFINITION.** Let  $K$  be a topological space. A *semigroup* on  $K$  is a continuous map  $G: [0, \infty) \times K \rightarrow K$  so  $G(s, G(t, X)) = G(s + t, X)$  for all  $X \in K, s, t \geq 0$ ; and  $G(0, X) = X$  for all  $X \in K$ .

Recall that  $C_0$  denotes the set of compact convex subsets of  $R^n$  with nonempty interior. Give  $C_0$  the topology as a subspace of  $C$ . We then have the following theorem.

**THEOREM 5.7.** *Suppose  $g(v)$  is a continuous real valued function on  $S^{n-1}$  such that  $\bigcap H(v, g(v)) \neq \emptyset$ . Then the map  $F(t, X) = F_t X = \bigcap H(v, h_X(v) + t g(v))$  defines a semigroup on  $C_0$ .*

*Proof.* It is immediate that  $F_0 X = X$ . If  $z \in \text{Int } X$  and  $y \in \bigcap H(v, g(v))$  one easily verifies that  $z + t y \in \text{Int } F_t X$ . Hence the result follows from Theorems 5.1 and 5.6.  $\square$

*Remark.* Note that the hypothesis of Theorem 5.7 is equivalent to the existence of  $z \in R^n$  so  $\langle z, v \rangle \leq g(v)$  for all  $v \in S^{n-1}$ . This hypothesis is satisfied if, for example,  $g$  is nonnegative and continuous.

**6. The Wulff shape.** In this section we study the behavior of  $F_t X$  as  $t$  gets large.

**DEFINITION.** Let  $X, K \in C$ . We say  $F_t X$  approaches the shape  $K$  (as  $t \rightarrow \infty$ ) if there exist compact convex sets  $A$  and  $B$  in  $R^n$  so, for all  $t \geq 0$ ,

$$A + tK \subseteq F_t X \subseteq B + tK.$$

If  $K$  has interior, then  $tK$  grows arbitrarily large in all directions; the deviation of  $F_t X$  from  $tK$  remains bounded and hence becomes proportionally negligible. This explains the definition. If  $K = \{0\}$ , then  $F_t X$  remains bounded for all  $t$ .

**DEFINITION.** The *Wulff shape*  $W$  of  $g$  is

$$W = \bigcap_{v \in S^{n-1}} H(v, g(v)).$$

We call this the Wulff shape after G. Wulff who in [9] noticed its significance for physical crystals.

The major theorem of this section is the following.

**THEOREM 6.1.** *Suppose the Wulff shape  $W$  for  $g$  has nonempty interior and  $X \in C$ . Then  $F_t X$  approaches the shape  $W$ .*

We note that the hypotheses are satisfied if, for example,  $g(v)$  is always strictly positive.

*Remark.* If  $W$  does not have interior, then  $F_t X$  need not approach the shape  $W$ . For example, in  $R^2$  define  $g$  by  $g((\cos \theta, \sin \theta)) = \cos^2 \theta$ . The choices  $\theta = \pm \pi/2$  show that  $W$  is contained in the  $x$ -axis. The line through  $(\cos^2 \theta)(\cos \theta, \sin \theta)$  normal to  $(\cos \theta, \sin \theta)$  meets the  $x$ -axis at  $x = \cos \theta$ . Hence  $W = \{0\}$ . Now let  $X = \bigcap_{v \in S^1} H(v, 1)$

be the unit disk. It is a simple argument from elementary calculus to show that the point  $(2\sqrt{t}, 0)$  lies in  $F_t X$  for all  $t \geq 0$ . Hence  $F_t X$  is unbounded and cannot approach the shape  $W = \{0\}$ .

We now proceed to prove Theorem 6.1. We need two preliminary results.

PROPOSITION 6.2. *Let  $X, Y \in C$ . If  $X \subseteq Y$ , then  $F_t X \subseteq F_t Y$ .*

*Proof.*  $F_t X = \bigcap H(v, h_X(v) + tg(v)) \subseteq \bigcap H(v, h_Y(v) + tg(v)) = F_t Y$ .  $\square$

LEMMA 6.3. *Let  $\mu \geq 0, t \geq 0$ . Let  $W$  be the Wulff shape and assume  $W \neq \emptyset$ . Then*

$$F_t(\mu W) = (\mu + t)W.$$

*Proof.* Clearly  $h_{\mu W}(v) \leq \mu g(v)$  for all  $v$ . Hence  $F_t(\mu W) = \bigcap H(v, h_{\mu W}(v) + tg(v)) \subseteq \bigcap H(v, \mu g(v) + tg(v)) = (\mu + t)W$ .

Conversely, if  $\mu > 0$ , then

$$\begin{aligned} (\mu + t)W &= ((\mu + t)/\mu)(\mu W) \\ &= ((\mu + t)/\mu) \bigcap H(v, h_{\mu W}(v)) \\ &= \bigcap H(v, ((\mu + t)/\mu)h_{\mu W}(v)) \\ &= \bigcap H(v, h_{\mu W}(v) + (t/\mu)h_{\mu W}(v)) \\ &= \bigcap H(v, h_{\mu W}(v) + th_W(v)) \\ &\subseteq \bigcap H(v, h_{\mu W}(v) + tg(v)) = F_t(\mu W). \end{aligned}$$

Finally, if  $\mu = 0, F_t(\mu W) = F_t(\{0\}) = \bigcap H(v, tg(v)) = tW$ .  $\square$

*Proof of Theorem 6.1.* Since  $W$  has interior we may find  $b \in R^n$  and  $\mu > 0$  so  $X \subseteq b + \mu W$ . Choose  $a \in X$ . Then

$$a + 0W \subseteq X \subseteq b + \mu W.$$

By Proposition 6.2,  $F_t(a + 0W) \subseteq F_t X \subseteq F_t(b + \mu W)$ . By Lemma 6.3 and invariance under translation,

$$a + tW \subseteq F_t X \subseteq b + (\mu + t)W.$$

Let  $A = \{a\}, B = b + \mu W$ . Then

$$A + tW \subseteq F_t X \subseteq B + tW \quad \text{for all } t \geq 0.$$

Theorem 6.1 then follows.  $\square$

We have remarked that Theorem 6.1 can fail if  $W$  has no interior. Even then, however, we do have the following weaker theorem. (Note that if  $W$  has interior, the result follows immediately from Theorem 6.1.)

PROPOSITION 6.4. *Suppose the Wulff shape  $W$  for  $g$  is nonempty and  $X \in C$ . Then for  $t > 0 (F_t X)/t \in C$  and  $\lim_{t \rightarrow \infty} (F_t X)/t = W$ .*

*Proof.* Let  $Y(t) = (F_t X)/t$ . It is immediate that for  $v \in S^{n-1}, h_{F_t X}(v) \leq h_X(v) + tg(v)$ ; hence  $h_{Y(t)}(v) \leq g(v) + h_X(v)/t$ . Taking  $t \geq 1$ , we see we may assume  $Y(t)$  lies in a bounded region of  $R^n$ . By the Blaschke selection theorem to prove  $\lim_{t \rightarrow \infty} Y(t) = W$  we need only show that if  $t_i \rightarrow \infty$  and  $Y(t_i)$  converges to  $K \in C$ , then  $K = W$ .

Suppose  $Y(t_i)$  converges to  $K$  while  $t_i \rightarrow \infty$ . Then from the above,  $h_K(v) \leq g(v)$  for all  $v \in S^{n-1}$ , so  $K \subset W$ . On the other hand, if  $a \in X$ , then  $a + 0W \subset X$  so  $a + tW \subset F_t X$  by Proposition 6.2 and Lemma 6.3; hence  $a/t + W \subset Y(t)$  and  $h_W(v) \leq h_K(v)$  for all  $v \in S^{n-1}$ . It follows  $W \subset K$  and hence  $W = K$ . This proves the proposition.  $\square$

Since the conclusion of Theorem 6.1 is stronger than that of Proposition 6.4 it is convenient to have different hypotheses which imply the former even if  $W$  has no interior.

**PROPOSITION 6.5.** *Suppose  $W \neq \emptyset$  and there exists a finite set  $E \subset S^{n-1}$  so  $W = \bigcap_{v \in E} H(v, g(v))$ . Suppose  $X \subset \mathbb{R}^n$  has nonempty interior. Then  $F_t X$  approaches the shape  $W$ .*

*Proof.* By translation we may assume  $0 \in W$ , so that  $g(v) \geq 0$ . The result then follows from Proposition 6.5 of Willson [8] together with the proof of Theorem 6.1 above.  $\square$

Using Proposition 3.1, one can check that the hypotheses of Proposition 6.5 are satisfied if  $g$  is a continuous positively homogeneous polyhedral function (see Rockafellar [6, p. 172]). The next result gives another condition on  $g$  which implies the hypotheses of Proposition 6.5.

**DEFINITION.** The polar graph  $G$  of  $g$  is  $\{g(v)v : v \in S^{n-1}\} \subset \mathbb{R}^n$ . (It corresponds to the graph of  $g$  in polar coordinates.) If  $c \in \mathbb{R}^n$ , the standard sphere  $S$  on center  $c$  is  $S = \{x \in \mathbb{R}^n : |x - c| = |c|\}$  and the standard disk  $D$  on center  $c$  is  $D = \{x \in \mathbb{R}^n : |x - c| \leq |c|\}$ . Note  $0 \in S$ . If  $c = 0$ ,  $S = D = \{0\}$ .

**PROPOSITION 6.6.** *Let  $S_1, \dots, S_m$  be the standard spheres on centers  $c_1, \dots, c_m$  respectively. Assume  $g(v) \geq 0$  for all  $v \in S^{n-1}$  and the polar graph  $G$  of  $g$  lies in  $S_1 \cup \dots \cup S_m$ . Then there exists a finite set  $E \subset S^{n-1}$  so the Wulff shape  $W$  satisfies*

$$W = \bigcap_{v \in E} H(v, g(v)).$$

*In particular,  $W$  is a convex polytope. If, in addition, for each  $i$ ,  $c_i$  is a rational vector, then  $W$  is a convex rational polytope.*

*Remark.* A rational vector is an element of  $\mathbb{R}^n$  each of whose coordinates is a rational number. A convex rational polytope is the convex hull of finitely many rational vectors.

The main interest of Proposition 6.6 lies in the application to growth of physical crystals. In general, one assumes that the growth function  $g(v)$  is proportional to the surface free energy in direction  $v$ . (See Herring [4].) But the surface free energy is often obtained by combining inner products  $\langle v, e_i \rangle$  where the vectors  $e_i$  are the locations of atoms in the crystal lattice. The polar graph of  $g_i(v) = \langle v, e_i \rangle$  is a standard sphere, and frequently the resulting  $g(v)$  satisfies the hypotheses of Proposition 6.6. In any event, it is routine to study free energies in terms of their polar graphs.

*Proof.* The existence of  $E$  is essentially the result (Theorem 4.1) of Willson [7]. More specifically for  $i = 1, \dots, m$  let  $D_i$  be the standard disk on center  $c_i$ ;  $D_0 = \{0\} \subset \mathbb{R}^n$ ;  $\mathcal{G} = \{av : v \in S^{n-1}, 0 \leq a \leq g(v)\}$ ;  $K$  be the set of nonempty subsets  $T$  of  $\{0, \dots, m\}$  such that  $\bigcap_{i \in T} D_i \subset \mathcal{G}$ . Then the proof of (4.1) in [7] applies to show the existence of  $E$  once we prove  $\bigcup_{T \in K} \bigcap_{i \in T} D_i = \mathcal{G}$ .

We indicate the proof that  $\mathcal{G} \subset \bigcup_{T \in K} \bigcap_{i \in T} D_i$ ; the other inclusion is immediate. For any  $v_0 \in S^{n-1}$ , let  $T = \{i : 0 \leq i \leq m, g(v_0)v_0 \in D_i\}$ ; then  $av_0 \in \bigcap_{i \in T} D_i$  for  $0 \leq a \leq g(v_0)$  and we need only show  $T \in K$ . If  $g(v_0) = 0$ , then  $0 \in T$  so  $T \in K$  trivially; hence we may assume  $g(v_0) > 0$ . Define  $k : S^{n-1} \rightarrow [0, \infty)$  by  $k(v) = \sup \{\lambda \in [0, \infty) : \lambda v \in \bigcap_{i \in T} D_i\}$ . Then  $k$  is continuous and to show  $T \in K$  we need only show  $k(v) \leq g(v)$  for all  $v \in S^{n-1}$ .

Suppose  $x \in S^{n-1}$  and  $k(x) > g(x)$ . Clearly  $x \neq \pm v_0$  and we may let  $P$  denote the 2-dimensional plane through 0 spanned by  $v_0$  and  $x$ . Choose polar coordinates for  $P$  so  $v_0$  has polar angle 0 and  $x$  lies in the upper half plane. Let  $v_\theta$  be the unit vector in the direction  $\theta$ . Since  $g(v_0)v_0$  lies on some  $S_i$ ,  $h(v_0) = g(v_0)$ . Find  $\theta_1$  and  $\varepsilon > 0$  so  $0 \leq \theta_1 < \pi$ ,  $k(v_\theta) \leq g(v_\theta)$  for  $0 \leq \theta \leq \theta_1$ , but  $k(v_\theta) > g(v_\theta)$  for  $\theta_1 < \theta < \theta_1 + \varepsilon$ . By the definition of  $k$  and by renumbering, we may assume  $k(v_\theta)v_\theta$  lies on  $S_1$  for  $\theta_1 < \theta < \theta_1 + \varepsilon$  where  $1 \in T$ ; and  $g(v_\theta)v_\theta$  lies on  $S_2$  for  $\theta_1 < \theta < \theta_1 + \varepsilon$  where  $2 \notin T$ .

By definition of  $T$ ,  $g(v_0)v_0 \in D_1 - D_2$ . By simple geometry,  $D_1 - D_2 \subset H(c_2 - c_1, 0)$ . But by choice of  $\theta_1$  it is easy to see that all points of  $H(c_2 - c_1, 0) \cap P$  have polar angle  $\theta$

satisfying  $\theta_1 \cong \psi \cong \theta_1 + \pi$ . This excludes the positive  $x$ -axis and contradicts that  $g(v_0)v_0 \in H(c_2 - c_1, 0)$ . The contradiction proves the existence of  $E$ .

The last sentence of Proposition 6.6 follows by a study of each step of the proof of Theorem 4.1 in [7]; for more details, the reader may consult Willson [8, § 3].  $\square$

**Acknowledgment.** I wish to thank the referee for numerous simplifications of this paper. In particular, Theorem 5.4 is due to the referee.

## REFERENCES

- [1] T. BONNESEN AND W. FENCHEL, *Theorie der Konvexen Körper*, Chelsea, New York, 1971.
- [2] W. K. BURTON, N. CABRERA AND F. C. FRANK, *The growth of crystals and the equilibrium structure of their surfaces*, Philos. Trans. Roy. Soc. London Ser. A, 243 (1951), pp. 299–358.
- [3] H. G. EGGLESTON, *Convexity*, Cambridge University Press, London, 1969.
- [4] C. HERRING, *Some theorems on the free energies of crystal surfaces*, Phys. Rev., second series, 82 (1951), pp. 87–93.
- [5] M. VON LAUE, *Der Wulffsche Satz für die Gleichgewichtsform von Kristallen*, Z. Krist. 105 (1943), pp. 124–133.
- [6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [7] S. J. WILLSON, *Limiting shapes for configurations*, J. Comput. System Sci., 15 (1977), pp. 243–261.
- [8] ———, *On convergence of configurations*, Discrete Math. 23 (1978), pp. 279–300.
- [9] G. WULFF, *Zur Frage der Geschwindigkeit des Wachstums und der Auflösung der Krystallflächen*, Z. Krist. 34 (1901), pp. 449–530.



## A NOTE ON THE ASYMPTOTIC EXPANSION OF EIGENVALUES\*

JAMES MURDOCK† AND CLARK ROBINSON‡

**Abstract.** Under certain conditions  $k$ -term asymptotic expansions of the eigenvalues of a matrix can be deduced from a  $k$ -term asymptotic expansion of the matrix.

Suppose  $L_\varepsilon \sim L_0 + \varepsilon L_1 + \varepsilon^2 L_2 + \dots$  is an asymptotic expansion of a matrix function of a small parameter  $\varepsilon$ , and it is desired to find a few terms of the expansion of the eigenvalues. Does it suffice to take  $k$  terms in the expansion of  $L_\varepsilon$  to obtain  $k$  terms in the expansion of the eigenvalues? The example

$$L_\varepsilon = \begin{bmatrix} \varepsilon^k & \varepsilon^{k-1} \\ \delta \varepsilon^{k+1} & -\varepsilon^k \end{bmatrix}$$

which has eigenvalues  $\pm \varepsilon^k \sqrt{1 + \delta}$ , shows that this is not always the case. The following theorem gives a sufficient condition for this to be true.

**THEOREM.** Let  $L_\varepsilon$  and  $N_\varepsilon$  be continuous real or complex matrix functions of  $\varepsilon$  defined for  $\varepsilon \geq 0$ , and let  $M_\varepsilon = L_\varepsilon + \varepsilon^{k+1} N_\varepsilon$ . Suppose there exists a matrix  $C_\varepsilon$  defined in some interval  $0 \leq \varepsilon < \varepsilon_0$ , continuous in  $\varepsilon$  and nonsingular, such that  $C_\varepsilon^{-1} L_\varepsilon C_\varepsilon = D_\varepsilon = \text{diag}(\lambda_1(\varepsilon), \dots, \lambda_n(\varepsilon))$ . Suppose further that each pair of eigenvalues  $\lambda_i(\varepsilon), \lambda_j(\varepsilon)$  satisfies either  $\lambda_i(\varepsilon) = \lambda_j(\varepsilon) + O(\varepsilon^{k+1})$  or  $|\lambda_i(\varepsilon) - \lambda_j(\varepsilon)| \geq c\varepsilon^k$  for some  $c > 0$  (this condition is satisfied automatically if each eigenvalue  $\lambda_i(\varepsilon)$  is a  $C^{k+1}$  function of  $\varepsilon$ ). Then  $M_\varepsilon$  has  $n$  eigenvalues of the form

$$\nu_i(\varepsilon) = \lambda_i(\varepsilon) + \varepsilon^{k+1} \sigma_i(\varepsilon)$$

for  $i = 1, \dots, n$ .

*Remarks.* The hypotheses are satisfied for  $L_\varepsilon = L_0 + \varepsilon L_1 + \dots + \varepsilon^k L_k$  if  $L_0$  has distinct eigenvalues, or if  $L_0 = I$  and  $L_1$  has distinct eigenvalues. The referee has informed us that according to a theorem of Rellich, the hypotheses are also satisfied if  $L_0, \dots, L_k$  are Hermitian; see [4, p. 376]. In the example preceding the theorem,  $C_\varepsilon$  exists for  $\varepsilon > 0$  but either becomes unbounded or singular as  $\varepsilon \rightarrow 0$ . Thus it is necessary to insist on the continuity and nonsingularity of  $C_\varepsilon$  at  $\varepsilon = 0$  even if  $L_0$  is already diagonal.

The proof is based on a degree argument of Levinson [2], previously exploited by Coppel and Howe [1]. We first obtained this theorem in connection with our work on asymptotic expansions in dynamical systems ([3]). Although we eventually used a different argument there, we thought this result might have independent interest.

*Proof.* The eigenvalues of  $L_\varepsilon$  may be partitioned into equivalence classes,  $\lambda_i$  and  $\lambda_j$  being equivalent if  $\lambda_i(\varepsilon) = \lambda_j(\varepsilon) + O(\varepsilon^{k+1})$ . By re-numbering the eigenvalues and permuting the columns of  $C_\varepsilon$ , we may assume that  $\lambda_1, \dots, \lambda_p$  are equivalent and that none of these are equivalent to  $\lambda_{p+1}, \dots, \lambda_n$ . We shall show the existence of  $p$  eigenvalues of the form  $\nu_i(\varepsilon) = \lambda_i(\varepsilon) + \varepsilon^{k+1} \sigma_i(\varepsilon)$ ,  $i = 1, \dots, p$ . The existence of  $n$  such eigenvalues follows by repeating the argument with different equivalence classes of eigenvalues placed first.

Let  $\lambda(\varepsilon) = \lambda_1(\varepsilon)$  and observe that for  $i = 1, \dots, p$  we have  $\lambda_i(\varepsilon) = \lambda(\varepsilon) + \varepsilon^{k+1} \phi_i(\varepsilon)$ , with  $\phi_i(\varepsilon)$  continuous, hence  $\lambda_i(\varepsilon) = \lambda(\varepsilon) + \varepsilon^{k+1} \phi_i(0) + o(\varepsilon^{k+1})$ . Let

\* Received by the editors February 28, 1979, and in revised form August 13, 1979.

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

‡ Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

$\Lambda(\varepsilon) = \lambda(\varepsilon)I_p$ , where  $I_p$  is the  $p \times p$  identity matrix, and let  $\Phi = \text{diag}(\phi_1(0), \dots, \phi_p(0))$ . Then

$$D_\varepsilon = \left[ \begin{array}{c|c} \Lambda_\varepsilon + \varepsilon^{k+1}\Phi + o(\varepsilon^{k+1}) & 0 \\ \hline 0 & \hat{D}_\varepsilon \end{array} \right]$$

where  $\hat{D}_\varepsilon = \text{diag}(\lambda_{p+1}(\varepsilon), \dots, \lambda_n(\varepsilon))$ . Now

$$C_\varepsilon^{-1}M_\varepsilon C_\varepsilon = D_\varepsilon + \varepsilon^{k+1}C_0^{-1}N_0C_0 + o(\varepsilon^{k+1}),$$

and we write

$$C_0^{-1}N_0C_0 = B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where  $B_{11}$  is a  $p \times p$  block. Now  $\nu$  is an eigenvalue of  $M_\varepsilon$  if and only if it is an eigenvalue of  $C_\varepsilon^{-1}M_\varepsilon C_\varepsilon$ , hence if and only if

$$\det \left[ \begin{array}{c|c} \Lambda_\varepsilon + \varepsilon^{k+1}(\Phi + B_{11}) + o(\varepsilon^{k+1}) - \nu I_p & \varepsilon^{k+1}B_{12} + o(\varepsilon^{k+1}) \\ \hline \varepsilon^{k+1}B_{21} + o(\varepsilon^{k+1}) & \hat{D}_\varepsilon + \varepsilon^{k+1}B_{22} + o(\varepsilon^{k+1}) - \nu I_{n-p} \end{array} \right] = 0.$$

This equation has the form  $f(\varepsilon, \nu) = 0$ , to be solved for  $\nu = \nu(\varepsilon)$ . Make the  $\varepsilon$ -dependent change of variables  $\nu \leftrightarrow \sigma$  defined by  $\nu = \lambda(\varepsilon) + \varepsilon^{k+1}\sigma$ ; this will yield an equation  $g(\varepsilon, \sigma) = 0$  which we now determine. First note that  $\Lambda_\varepsilon - \nu I_p = -\varepsilon^{k+1}\sigma I_p$ . From the manner of partitioning the  $\lambda_i$  we see that there exist constants  $c > 0, \varepsilon_0 > 0$  such that for each  $i > p$ ,

$$\frac{|\lambda_i(\varepsilon) - \lambda(\varepsilon)|}{\varepsilon^k} \geq c \quad \text{for } 0 < \varepsilon < \varepsilon_0.$$

Hence  $\hat{D}_\varepsilon - \nu I_{n-p} = \varepsilon^k T_\varepsilon - \varepsilon^{k+1}\sigma I_{n-p}$  where  $T_\varepsilon$  is diagonal with each diagonal element bounded away from zero as  $\varepsilon \rightarrow 0$ . Inserting these relations in our determinant and canceling  $\varepsilon^{k+1}$  from the top  $p$  rows and  $\varepsilon^k$  from the remainder we find

$$\begin{aligned} 0 &= \det \left[ \begin{array}{c|c} (\Phi + B_{11}) - \sigma I_p & B_{12} \\ \hline 0 & T_\varepsilon \end{array} \right] + o(1) \\ &= \det [(\Phi + B_{11}) - \sigma I_p] \det T_\varepsilon + o(1). \end{aligned}$$

Since  $\det T_\varepsilon$  is bounded away from zero this reduces to

$$\det [(\Phi + B_{11}) - \sigma I_p] + o(1) = 0.$$

When  $\varepsilon = 0$  there exist  $p$  roots for  $\sigma$  by the fundamental theorem of algebra; these persist for small  $\varepsilon$  by Rouché's theorem. Q.E.D.

REFERENCES

[1] W. A. COPPEL AND A. HOWE, *Dynamical instability of linear canonical systems*, Austr. Math. Soc. J., 7 (1967), pp. 247-251.  
 [2] N. LEVINSON, *The stability of linear, real, periodic self-adjoint systems of differential equations*, J. Math. Anal. Appl., 6 (1963), pp. 473-482.  
 [3] J. A. MURDOCK AND C. ROBINSON, *Qualitative dynamics from asymptotic expansions: local theory*, J. Differential Equations, submitted.  
 [4] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.

## NONEXISTENCE OF CONTINUOUS SELECTIONS OF THE METRIC PROJECTION AND WEAK CHEBYSHEV SYSTEMS\*

GÜNTHER NÜRNBERGER†

**Abstract.** We show that an  $n$ -dimensional subspace  $G$  of  $C[a, b]$  which admits a continuous selection for the metric projection has to be weak Chebyshev. For  $C[a, b]$  this solves one part of a problem posed by Lazar, Morris and Wulbert [*Continuous selections for metric projections*, J. Functional Analysis, 3 (1969), pp. 193–216] who have proved this result for one-dimensional subspaces of  $C(X)$ ,  $X$  compact. Combining our theorem with known results we obtain a complete characterization of the existence of continuous selections for the metric projection for a certain class of  $n$ -dimensional subspaces in  $C[a, b]$ .

**1. Introduction.** We consider the following approximation problem: If  $G$  is an  $n$ -dimensional subspace of  $C_0(X)$ , the space of real-valued continuous functions  $f$  on a locally compact space  $X$  vanishing at infinity, i.e., for each  $\varepsilon > 0$  the set  $\{x \in X : |f(x)| \geq \varepsilon\}$  is compact, endowed with the norm  $\|f\| = \sup \{|f(x)| : x \in X\}$ , then for each  $f \in C_0(X)$  we are interested in the set  $P_G(f) = \{g_0 \in G : \|f - g_0\| = \inf \{\|f - g\| : g \in G\}$  which is called the set of *best approximations* of  $f$  from  $G$ . This defines a set-valued mapping  $P_G$  from  $C_0(X)$  into  $2^G$  which is called the *metric projection* onto  $G$ . A continuous mapping  $s$  from  $C_0(X)$  onto  $G$  is called *continuous selection* for  $P_G$ , if  $s(f) \in P_G(f)$  for each  $f \in C_0(X)$ .

In the last years many authors have investigated continuity properties of the set-valued metric projection, in particular selection problems (see e.g., Singer [9] and Vlasov [12]). The question, if continuous selections for  $P_G$  exist, is relevant for the convergence of algorithms for computing best approximations.

Lazar, Morris and Wulbert [4] were the first to characterize those one-dimensional subspaces  $G$  in  $C(X)$ ,  $X$  compact, which admit continuous selections for  $P_G$ . They posed the problem to characterize the corresponding  $n$ -dimensional subspaces. This question has also been raised in the book of Holmes [2]. Applying new methods, namely the theory of weak Chebyshev subspaces, Nürnberger and Sommer [7] were able to establish the existence of continuous selections for  $P_G$  for a class of  $n$ -dimensional weak Chebyshev subspaces  $G$  in  $C[a, b]$ ,  $n$  arbitrary, from which a result of Brown [1] for five-dimensional subspaces in  $C[-1, 1]$  follows. The result of Nürnberger and Sommer [7] has been extended by Nürnberger [6] to the case of  $n$ -dimensional subspaces of  $C_0(X)$ ,  $X$  locally compact, where  $X$  is a subset of the real line, if  $n \geq 2$ . Nürnberger and Sommer [8] used their selection theorem in [7] to give a complete characterization of those spline spaces  $G$  which admit a continuous selection for  $P_G$ . This result has been extended by Sommer [11] to “generalized” splines which are also weak Chebyshev.

The literature shows that in all theorems about the existence of continuous selections for the metric projection from  $C[a, b]$  onto  $n$ -dimensional subspaces  $G$  the spaces  $G$  are weak Chebyshev. In this paper we can actually prove the following theorem:

*Let  $G$  be an  $n$ -dimensional subspace of  $C[a, b]$  such that there exists a continuous selection for  $P_G$ . Then  $G$  is weak Chebyshev.*

This solves one part of the problem, posed by Lazar, Morris and Wulbert [4], for  $C[a, b]$ .

Now combining our theorem and results in Nürnberger and Sommer [7], and Sommer [10], we obtain a complete characterization of those  $n$ -dimensional subspaces

\* Received by the editors February 28, 1979.

† Institute für Angewandte Mathematik der Universität Erlangen-Nürnberg, 8520 Erlangen, Germany.

$G$  in  $C[a, b]$  with the property that no  $g \in G, g \neq 0$ , vanishes on an interval, which admit a continuous selection for  $P_G$ :

For an  $n$ -dimensional subspace  $G$  in  $C[a, b]$  with the property that no  $g \in G, g \neq 0$ , vanishes on an interval, the following statements are equivalent:

- (i) There exists a continuous selection for  $P_G$ .
- (ii) (a)  $G$  is weak Chebyshev. (b) Each  $g \in G, g \neq 0$ , has at most  $n$  distinct zeros in  $[a, b]$ .

This represents a complete solution of the problem, posed by Lazar, Morris and Wulbert [4], for the case of  $n$ -dimensional subspaces  $G$  in  $C[a, b]$  with the property that no  $g \in G, g \neq 0$ , vanishes on an interval.

That (ii) implies (i) has been shown by Nürnberger, Sommer [7], that (i) implies (iib) has been shown by Sommer [10], under the assumption that  $G$  is weak Chebyshev and that (i) implies (iia) is a consequence of our theorem.

**2. The main result.** To prove our theorem we use the following notation:

For  $f, f_1, f_2 \in C[a, b]$  and  $A \subset X$  we denote by  $Z(f) = \{x \in X : f(x) = 0\}$ ,  $Z(P_G(f)) = \{x \in X : g(x) = 0 \text{ for each } g \in P_G(f)\}$  and by  $\text{bd } A$  the boundary of  $A$ . Furthermore  $f_1 = f_2$  (respectively  $f_1 \leqq f_2$ ) on  $A$  means that  $f_1(x) = f_2(x)$  (respectively  $f_1(x) \leqq f_2(x)$ ) for each  $x \in A$ . If  $x \in X$  then by  $U(x)$  we denote the system of all neighborhoods of  $x$ .

A set  $\{g_1, \dots, g_n\}$  of  $n$  linearly independent real-valued functions, defined on a set  $Y$ , which contains at least  $n + 1$  distinct points, is called a *Chebyshev-system*, if for each  $n$  distinct points  $y_1, \dots, y_n$  in  $Y$  we have  $\det (g_i(y_j)) \neq 0$ .

DEFINITION. An  $n$ -dimensional subspace  $G$  of  $C[a, b]$  is called *weak Chebyshev*, if each  $g \in G$  has at most  $n - 1$  sign changes, i.e., there do not exist  $n + 1$  distinct points  $x_1, \dots, x_{n+1} \in [a, b]$ , where  $x_1 < \dots < x_{n+1}$ , such that  $\varepsilon(-1)^i g(x_i) > 0, i = 1, \dots, n + 1, \varepsilon = \pm 1$ .

We remark that Jones and Karlovitz [3] have pointed out the importance of weak Chebyshev subspaces by showing that an  $n$ -dimensional subspace  $G$  of  $C[a, b]$  is weak Chebyshev if and only if for each  $f \in C[a, b]$  there exists a function  $g_0 \in P_G(f)$  such that  $f - g_0$  alternates  $n + 1$  times, i.e., there exist  $n + 1$  points  $x_1 < \dots < x_{n+1}$  such that  $\varepsilon(-1)^i (f - g_0)(x_i) = \|f - g_0\|, i = 1, \dots, n + 1, \varepsilon = \pm 1$ . This result has been extended by Deutsch, Nürnberger and Singer [13] to  $C_0(X)$ , where  $X$  is a locally compact subset of the real line.

In the proof of our theorem we use the following necessary condition for the existence of continuous selections for  $P_G$ , proved by Lazar, Morris and Wulbert [4] for  $C(X), X$  compact, which we only formulate for  $C[a, b]$ .

LEMMA (Lazar, Morris and Wulbert [4]). *Let  $G$  be an  $n$ -dimensional subspace of  $C[a, b]$ , such that there exists a continuous selection for  $P_G$  and let  $f \in C[a, b]$  with  $\|f\| = 1$  and  $0 \in P_G(f)$ . Then there exists a function  $g_0 \in P_G(f)$  such that:*

- (i) *For each  $x \in \text{bd } Z(P_G(f)) \cap f^{-1}(1)$  and each  $g \in P_G(f)$  there exists a neighborhood  $U \in U(x)$  such that  $g_0 \geqq g$  on  $U$ .*
- (ii) *For each  $x \in \text{bd } Z(P_G(f)) \cap f^{-1}(-1)$  and each  $g \in P_G(f)$  there exists a neighborhood  $U \in U(x)$  such that  $g_0 \leqq g$  on  $U$ .*

Now we are in position to prove, that only the weak Chebyshev subspaces  $G$  in the class of  $n$ -dimensional subspaces of  $C[a, b]$  admit a continuous selection for  $P_G$ .

THEOREM. *Let  $G$  be an  $n$ -dimensional subspace of  $C[a, b]$ , such that there exists a continuous selection for  $P_G$ . Then  $G$  is weak Chebyshev.*

*Proof.* Assume that there exists a continuous selection for  $P_G$  and that  $G$  is not weak Chebyshev, i.e., there exists a function  $h_n \in G$  and there exist  $n + 1$  distinct points  $z_1 < \dots < z_{n+1}$  such that

$$\varepsilon(-1)^i h_n(z_i) > 0, \quad i = 1, \dots, n + 1, \quad \varepsilon = \pm 1.$$

By scaling we may assume that  $\|h_n\| = 1$ .

Case 1. The function  $h_n$  changes sign at  $n$  distinct points  $x_1 < \dots < x_n$  with  $z_i < x_i < z_{i+1}$ ,  $i = 1, \dots, n$ , i.e. for each  $i \in \{1, \dots, n\}$  and each  $U \in U(x_i)$  the function  $h_n$  attains both strictly positive and strictly negative values and  $h_n(x_i) = 0$ .

Let  $\{h_1, \dots, h_n\}$  be a basis of  $G$ . We show:

(1) There exist  $k$  distinct points  $y_1, \dots, y_k \in \{x_1, \dots, x_n\}$ ,  $1 \leq k \leq n$ , and a function  $f$  defined on  $\{x_1, \dots, x_n\} \setminus \{y_2, \dots, y_k\}$  such that  $|f| = 1$  on  $\{x_1, \dots, x_n\} \setminus \{y_2, \dots, y_k\}$  and we have the choice to define  $f$  such that  $f(y_1) = 1$  or  $f(y_1) = -1$  such that for each continuous extension of  $f$  to  $[a, b]$  with  $\|f\| = 1$  we have  $P_G(f) \subset \text{span}\{g_1, \dots, g_k\}$ , where  $\{g_1, \dots, g_k\} \subset \{h_1, \dots, h_n\}$  are linearly independent and  $g_1 = h_n$  and  $y_1 \in \bigcap_{i=1}^k Z(g_i)$ .

*Proof of (1).* We consider the subspace  $\text{span}\{h_1, \dots, h_{n-1}\}$ . Either  $\text{span}\{h_1, \dots, h_{n-1}\}$  is a Chebyshev-system on  $\{x_1, \dots, x_n\}$  or, by choosing a new basis of  $\text{span}\{h_1, \dots, h_{n-1}\}$  and renumbering the points  $x_1, \dots, x_n$ , if necessary, we may assume that  $h_{n-1} = 0$  on  $\{x_1, \dots, x_{n-1}\}$ . Again, either  $\text{span}\{h_1, \dots, h_{n-2}\}$  is a Chebyshev-system on  $\{x_1, \dots, x_{n-1}\}$  or we may assume that  $h_{n-2} = 0$  on  $\{x_1, \dots, x_{n-2}\}$ . If we continue this method by induction, at each step we may assume that  $h_i = 0$  on  $\{x_1, \dots, x_i\}$ ,  $1 \leq i \leq n - 1$ , until the induction stops.

Therefore we get that either the induction stops, i.e. there exists a number  $m \in \{1, \dots, n - 1\}$  such that  $\text{span}\{h_1, \dots, h_m\}$  is a Chebyshev-system on  $\{x_1, \dots, x_{m+1}\}$  and  $x_1, \dots, x_{m+1} \in \bigcap_{i=m+1}^n Z(h_i)$ , or  $\text{span}\{h_1\}$  is not a Chebyshev-system on  $\{x_1, x_2\}$ , i.e., we may assume that  $h_1(x_1) = 0$  and therefore  $x_1 \in \bigcap_{i=1}^n Z(h_i)$ .

If  $x_1 \in \bigcap_{i=1}^n Z(h_i)$ , we set  $y_1 = x_1$ , choose a basis  $\{g_1, \dots, g_n\}$  of  $G$  such that  $g_1 = h_n$  and get (1) with  $k = n$ .

Otherwise we conclude as follows:

Since  $\text{span}\{h_1, \dots, h_m\}$  is a Chebyshev-system on  $\{x_1, \dots, x_{m+1}\}$ , by choosing a new basis of  $\text{span}\{h_1, \dots, h_m\}$ , if necessary, we may assume that for each  $i \in \{1, \dots, m\}$  we have  $h_i(x_{m+1}) = 1$ ,  $h_i(x_j) = 0$  for each  $j \in \{1, \dots, m\} \setminus \{i\}$  and  $h_i(x_i) \neq 0$  for each  $i \in \{1, \dots, m\}$ . Now we define  $f$  on  $\{x_1, \dots, x_{m+1}\}$  as follows: For each  $i \in \{1, \dots, m\}$  set  $f(x_i) = \text{sgn } h_i(x_i)$  and  $f(x_{m+1}) = -1$ . If we extend  $f$  (arbitrarily) to  $[a, b]$  such that  $\|f\| = 1$ , we get for  $g = \sum_{i=1}^m a_i h_i$  in  $P_G(f)$ : If  $a_i < 0$  for some  $i = 1, \dots, m$ , then

$$|f(x_i) - g(x_i)| = |\text{sgn } h_i(x_i) - a_i h_i(x_i)| = |1 - a_i| |h_i(x_i)| > 1.$$

Therefore  $\|f - g\| > \|f - 0\|$ , which is a contradiction.

Therefore for each  $i \in \{1, \dots, m\}$  we have  $a_i \geq 0$ . If  $a_i > 0$  for some  $i = 1, \dots, m$ , then

$$|f(x_{m+1}) - g(x_{m+1})| = \left| -1 - \sum_{i=1}^m a_i \right| = 1 + \sum_{i=1}^m |a_i| > 1.$$

and we again have a contradiction.

Therefore for each  $i \in \{1, \dots, m\}$  we have  $a_i = 0$ . But this shows that  $P_G(f) \subset \text{span}\{h_{m+1}, \dots, h_n\}$ . Obviously we also can define  $f(x_{m+1}) = 1$ , if we set  $h_i(x_{m+1}) = -1$  for each  $i \in \{1, \dots, m\}$  and, as above, we get  $P_G(f) \subset \text{span}\{h_{m+1}, \dots, h_n\}$ . Now if we set

$$g_1 = h_n, \quad g_2 = h_{n-1}, \dots, g_{n-m} = h_{m+1}$$

and

$$y_1 = x_{m+1}, \quad y_2 = x_{m+2}, \dots, y_{n-m} = x_n,$$

we obtain (1) with  $k = n - m$ .

If  $k = 1$ , then we obviously can extend  $f$  continuously to  $[a, b]$  such that  $0 \leq f \leq \min\{1 + g_1, 1\}$  on a neighborhood of  $y_i$ , if  $f(y_i) = 1$  and  $\max\{-1 + g_1, -1\} \leq f \leq 0$  on a neighborhood of  $y_i$ , if  $f(y_i) = -1$ , and  $f = 0$  outside these neighborhoods. (We choose the neighborhoods to be disjoint.) Then  $\|f\| = 1$ ,  $0, g_1 \in P_G(f)$  and  $\dim P_G(f) = 1$ .

Then we consider  $f$  at a point  $y_i$ , where  $g_1$  changes sign. There we have  $f(y_i) = \varepsilon$  for some  $\varepsilon \in \{-1, 1\}$ . By the lemma there exists a function  $g \in P_G(f)$  and a neighborhood  $U_i \in U(y_i)$  such that  $g \geq \max\{0, g_1\}$  on  $U_i$ , if  $\varepsilon = 1$ , respectively  $g \leq \min\{0, g_1\}$  on  $U_i$ , if  $\varepsilon = -1$ . But this is not possible, since  $\text{span } P_G(f) = \text{span}\{g_1\}$  and  $g_1$  changes sign at  $x_i$ . Therefore we get a contradiction.

Therefore let  $k > 1$ . Now we make the following remark:

(2) If we define  $f$  as in (1) and on further points from  $\{y_2, \dots, y_k\}$  and possibly on a finite number of points in some neighborhoods of  $y_1, \dots, y_k$  such that  $f(x) = 1$  (respectively  $f(x) = -1$ ) for  $x \in [a, b]$ , where  $g_1(x) \geq 0$  (respectively  $g_1(x) \leq 0$ ), then there exists a continuous extension of  $f$  to  $[a, b]$  such that  $\|f\| = 1$  and  $0, g_1 \in P_G(f)$ . Because, if  $f$  is defined as above, then we obviously can extend  $f$  continuously such that  $0 \leq f \leq \min\{1 + g_1, 1\}$  on neighborhoods of points, where  $f = 1$ , and  $\max\{-1 + g_1, -1\} \leq f \leq 0$  on neighborhoods of points, where  $f = -1$ , and  $f = 0$  outside these neighborhoods. Of course we choose the neighborhoods to be disjoint. We make the following convention:

(3) Let  $f$  be defined (on a finite number of points) as in (2). Then, if we make a statement for  $f$ , we mean that this statement shall be true for each continuous extension of  $f$  to  $[a, b]$ , for which  $\|f\| = 1$  and  $0, g_1 \in P_G(f)$ . (Such an extension exists according to (2).)

(4) Now the method for our proof will be to show that either we can reduce the dimension of  $P_G(f)$ , by defining  $f$  as in (1) and additionally on further points from  $\{y_2, \dots, y_k\}$ , or we have  $y_1, \dots, y_k \in \bigcap_{i=1}^k Z(g_i)$ . Then from the fact that  $y_1, \dots, y_k \in \bigcap_{i=1}^k Z(g_i)$  we can deduce a contradiction or otherwise by reducing the dimension of  $P_G(f)$  after a finite number of steps we get a contradiction analogously as before.

Therefore we proceed as follows

(5) Let  $l$  be the maximal number of points in  $\{y_1, \dots, y_k\}$  which are in  $\bigcap_{i=1}^k Z(g_i)$ . By renumbering, if necessary, we may assume that  $y_1, \dots, y_l$  are in  $\bigcap_{i=1}^k Z(g_i)$ . Let  $f$  be as defined in (1).

For shortness we use the following notation:

We say that a function  $g \in G$  is an  $\varepsilon$ -function, where  $\varepsilon \in \{-1, 1\}$ , on  $U \in U(y_i)$ , where  $i \in \{1, \dots, l\}$ , if  $g \geq \max\{0, g_i\}$  (respectively  $g \leq \min\{0, g_i\}$ ) on  $U$ , if  $\varepsilon = 1$  (respectively  $\varepsilon = -1$ ).

We show:

(6) For each  $s \in \{l+1, \dots, k\}$ , each  $\varepsilon_1, \dots, \varepsilon_l \in \{-1, 1\}$  and each  $\varepsilon_{s+1}, \dots, \varepsilon_k \in \{-1, 1\}$  there exists a function  $g \in \text{span}\{g_1, \dots, g_k\}$  and there exist neighborhoods  $U_i \in U(y_i)$ ,  $i = 1, \dots, l$ , such that  $g$  is an  $\varepsilon_i$ -function on  $U_i$ ,  $i = 1, \dots, l$ ,  $g(x_i) = 0$ ,  $i = l+1, \dots, s$ , and  $\varepsilon_i g(y_i) \geq 0$ ,  $i = s+1, \dots, k$ .

*Proof of (6).* We prove (6) by induction on  $s$ . Let  $s = l+1$ ,  $\varepsilon_1, \dots, \varepsilon_l \in \{-1, 1\}$  and  $\varepsilon_{l+2}, \dots, \varepsilon_k \in \{-1, 1\}$  be given. Let  $f$  be defined as in (1) and extend  $f$  as follows:  $f(y_i) = \varepsilon_i$ ,  $i = 1, \dots, l$ ,  $f(y_s) = 1$  and  $f(y_i) = \varepsilon_i$ ,  $i = l+2, \dots, k$ . By the lemma there exists a function  $h_1 \in P_G(f) \subset \text{span}\{g_1, \dots, g_k\}$  and neighborhoods  $U_i \in U(y_i)$ ,  $i = 1, \dots, l$ , such that  $h_1$  is an  $\varepsilon_i$ -function on  $U_i$ ,  $i = 1, \dots, l$ ,  $h_1(y_s) \geq 0$  and  $\varepsilon_i h_1(y_i) \geq 0$ ,  $i = l+2, \dots, k$ .

Furthermore let  $f$  be defined as in (1) and extend  $f$  as follows:  $f(y_i) = \varepsilon_i$ ,  $i = 1, \dots, l$ ,  $f(y_s) = -1$  and  $f(y_i) = \varepsilon_i$ ,  $i = l+2, \dots, k$ . By the lemma there exists a function  $h_2 \in P_G(f) \subset \text{span}\{g_1, \dots, g_k\}$  and neighborhoods  $V_i \in U(y_i)$ ,  $i = 1, \dots, l$ , such that  $h_2$  is an

$\varepsilon_i$ -function on  $V_i, i = 1, \dots, l, h_2(y_s) \leq 0$  and  $\varepsilon_i h_2(y_i) \geq 0, i = l+2, \dots, k$ .

If  $h_2(y_s) = 0$ , then  $g = h_2$  has the desired property. If  $h_2(y_s) < 0$ , then there exists a scalar  $a \geq 0$  such that  $h_1(y_s) + ah_2(y_s) = 0$  and  $g = h_1 + ah_2$  has the desired property. In particular the function  $g$  is an  $\varepsilon_i$ -function on  $U_i \cap V_i, i = 1, \dots, l$ .

Let the statement be true for  $s - 1$  and let  $\varepsilon_1, \dots, \varepsilon_l \in \{-1, 1\}$  and  $\varepsilon_{s+1}, \dots, \varepsilon_k \in \{-1, 1\}$  be given and set  $\varepsilon_s = 1$  (respectively  $\varepsilon_s = -1$ ). By induction hypothesis there exists a function  $h_1 \in \text{span}\{g_1, \dots, g_k\}$  (respectively  $h_2 \in \text{span}\{g_1, \dots, g_k\}$ ) and neighborhoods  $U_i \in U(y_i)$  (respectively  $V_i \in U(y_i)$ ),  $i = 1, \dots, l$ , such that  $h_1$  (respectively  $h_2$ ) is an  $\varepsilon_i$ -function on  $U_i$  (respectively on  $V_i$ ),  $i = 1, \dots, l, h_1(y_i) = 0$  (respectively  $h_2(y_i) = 0$ ),  $i = l+1, \dots, s-1, h_1(y_s) \geq 0$  (respectively  $h_2(y_s) \leq 0$ ),  $\varepsilon_i h_i(y_i) \geq 0$  (respectively  $\varepsilon_i h_i(y_i) \leq 0$ ),  $i = s+1, \dots, k$ . If  $h_2(y_s) = 0$ , then  $g = h_2$  has the desired property. If  $h_2(y_s) < 0$ , then there exists a scalar  $a \geq 0$  such that  $h_1(y_s) + ah_2(y_s) = 0$  and  $g = h_1 + ah_2$  has the desired property. In particular the function  $g$  is an  $\varepsilon_i$ -function on  $U_i \cap V_i, i = 1, \dots, l$ . This shows (6).

From (6) we immediately get

(7) For each  $\varepsilon_1, \dots, \varepsilon_l \in \{-1, 1\}$  there exists a function  $g \in \text{span}\{g_1, \dots, g_k\}$  and there exists neighborhoods  $U_i \in U(y_i), i = 1, \dots, l$ , such that  $g$  is an  $\varepsilon_i$ -function on  $U_i, i = 1, \dots, l$ , and  $g(y_i) = 0, i = l+1, \dots, k$ .

Now we assume that for each  $i \in \{1, \dots, k\}$  there exist neighborhoods  $V_i \in U(y_i), i = 1, \dots, k$ , such that for each  $i \in \{1, \dots, k\}$  we have  $g_1(x) < 0$  (respectively  $g_1(x) > 0$ ) on  $V_i \cap \{x \in [a, b] : x < x_i\}$  (respectively  $V_i \cap \{x \in [a, b] : x > x_i\}$ ). The other cases follow analogously.

We show:

(8) For each  $\varepsilon_1, \dots, \varepsilon_l \in \{-1, 1\}$  we can choose an integer  $n_1$  as we want, provided we choose  $n_1$  large enough, such that there exists a function  $g \in \text{span}\{g_1, \dots, g_k\}$  and there exist neighborhoods  $U_i \in U(y_i), i = 1, \dots, l$ , such that  $g$  is an  $\varepsilon_i$ -function on  $U_i, i = 1, \dots, l, g(y_1 - 1/n_1) \leq 0$ , where  $y_1 - 1/n_1 \notin U_1$ , and  $g(y_i) = 0, i = l+1, \dots, k$ .

*Proof of (8).* We prove (8) analogously as (7) by showing that in (6) additionally  $g(y_1 - 1/n_1) \leq 0$  for some  $n_1$  with  $y_1 - 1/n_1 \notin U_1$ . This we do by defining  $f$  in the proof of (6) additionally to be  $-1$  at the point  $y_1 - 1/n_1$ . This shows (8).

We show:

(9) For each  $s \in \{1, \dots, l\}$  and each  $\varepsilon_{s+1}, \dots, \varepsilon_l \in \{-1, 1\}$  we can choose integers  $n'_1, n_1, \dots, n_l$  as we want, provided we choose them large enough, such that there exists a function  $g \in \text{span}\{g_1, \dots, g_k\}$  with  $g > 0$  on  $(y_1, y_1 + 1/n'_1]$ ,  $g(y_i - 1/n_i) = 0, i = 1, \dots, s, \varepsilon_i g \geq 0$  on  $[y_i - 1/n_i, y_i], i = s+1, \dots, l, g(y_i) = 0, i = l+1, \dots, k$ .

*Proof of (9).* We prove (9) by induction on  $s$ . Let  $s = 1$  and  $\varepsilon_2, \dots, \varepsilon_l \in \{-1, 1\}$  be given. Set  $\varepsilon_1 = 1$  (respectively  $\varepsilon_1 = -1$ ). By (8) we can choose an integer  $m_1$  (respectively  $n_1$ ) as we want, provided we choose the integer large enough, such that there exists a function  $h_1 \in \text{span}\{g_1, \dots, g_k\}$  (respectively  $h_2 \in \text{span}\{g_1, \dots, g_k\}$ ) and there exist neighborhoods  $U_i \in U(y_i)$  (respectively  $V_i \in U(y_i)$ ),  $i = 1, \dots, l$ , such that  $h_1$  (respectively  $h_2$ ) is an  $\varepsilon_i$ -function on  $U_i$  (respectively  $V_i$ ),  $i = 1, \dots, l, h_1(y_1 - 1/m_1) \geq 0$  (respectively  $h_2(y_1 - 1/n_1) \leq 0$ ), where  $y_1 - 1/n_1 \notin V_1$ , and  $h_1(y_i) = 0$  (respectively  $h_2(y_i) = 0$ ),  $i = l+1, \dots, k$ . We can choose the integers  $m_1$  and  $n_1$  to be equal. If  $h_2(y_1 - 1/n_1) = 0$  then it is easy to see that  $g = h_2$  has the desired property. If  $h_2(y_1 - 1/n_1) < 0$  then there exists a scalar  $a \geq 0$  such that  $h_1(y_1 - 1/n_1) + ah_2(y_1 - 1/n_1) = 0$  and it is easy to see that  $g = h_1 + ah_2$  has the desired property.

Let the induction be true for  $s - 1$  and  $\varepsilon_{s+1}, \dots, \varepsilon_l \in \{-1, 1\}$  be given. Set  $\varepsilon_s = 1$  (respectively  $\varepsilon_s = -1$ ). Then by induction hypothesis we can choose integers  $m'_1, m_1, \dots, m_l$  (respectively  $n'_1, n_1, \dots, n_l$ ) as we want, provided we choose them large enough, such that there exists a function  $h_1 \in \text{span}\{g_1, \dots, g_k\}$  (respectively  $h_2 \in \text{span}\{g_1, \dots, g_k\}$ ) such that  $h_1 > 0$  on  $(y_1, y_1 + 1/m'_1]$  (respectively  $h_2 > 0$  on

$(y_1, y_1 + 1/n'_1]$ ,  $h_1(y_i - 1/m_i) = 0$  (respectively  $h_2(y_i - 1/n_i) = 0$ ),  $i = 1, \dots, s - 1$ ,  $h_1 \geq 0$  on  $[y_s - 1/m_s, y_s]$  (respectively  $h_2 \leq 0$  on  $[y_s - 1/n_s, y_s]$ ),  $\varepsilon_i h_1 \geq 0$  on  $[y_i - 1/m_i, y_i]$  (respectively  $\varepsilon_i h_2 \geq 0$  on  $[y_i - 1/n_i, y_i]$ ),  $i = s + 1, \dots, l$ , and  $h_1(y_i) = 0$  (respectively  $h_2(y_i) = 0$ ),  $i = l + 1, \dots, k$ . We can choose the integers such that  $m'_1 = n'_1$  and  $m_i = n_i$ ,  $i = 1, \dots, l$ .

Now we can choose an integer  $k_s \geq m_s = n_s$ . If  $h_2(y_s - 1/k_s) = 0$ , then  $g = h_2$  has the desired property. If  $h_2(y_s - 1/k_s) < 0$ , then there exists a scalar  $a \geq 0$  such that  $h_1(y_s - 1/k_s) + ah_2(y_s - 1/k_s) = 0$  and  $g = h_1 + ah_2$  has the desired property. This proves (9).

From (9) we immediately get the following.

(10) We can choose integers  $n_1, \dots, n_l$  as we want, provided we choose them large enough, such that there exists a function  $g \in \text{span}\{g_1, \dots, g_k\}$ ,  $g \neq 0$ , with  $g > 0$  on  $[y_1, y_1 + 1/n'_1]$ ,  $g(y_i - 1/n_i) = 0$ ,  $i = 1, \dots, l$ , and  $g(y_i) = 0$ ,  $i = l + 1, \dots, k$ .

We show:

(11) For each  $j \in \{2, \dots, l\}$ , each  $s \in \{1, \dots, l\} \setminus \{j\}$  and each  $\varepsilon_i \in \{-1, 1\}$ ,  $i = s + 1, \dots, l$ ,  $i \neq j$ , we can choose integers  $n_1, \dots, n_l$  as we want, provided we choose them large enough, such that there exists a function  $\tilde{g}_j \in \text{span}\{g_1, \dots, g_k\}$  with  $\tilde{g}_j < 0$  on  $[y_i - 1/n_i, y_i]$ ,  $\tilde{g}_j(y_i - 1/n_i) = 0$ ,  $i = 1, \dots, s$ ,  $i \neq j$ ,  $\varepsilon_i \tilde{g}_j \geq 0$  on  $[y_i - 1/n_i, y_i]$ ,  $i = s + 1, \dots, l$ ,  $i \neq j$ , and  $\tilde{g}_j(y_i) = 0$ ,  $i = l + 1, \dots, k$ .

*Proof of (11).* We prove (11) by induction on  $s$ . Let  $j \in \{2, \dots, l\}$ ,  $s = 1$  and  $\varepsilon_i \in \{-1, 1\}$ ,  $i = 2, \dots, l$ ,  $i \neq j$ , be given. Set  $\varepsilon_j = -1$  and  $\varepsilon_1 = 1$  (respectively  $\varepsilon_1 = -1$ ). By (7) there exists a function  $h_j \in \text{span}\{g_1, \dots, g_k\}$  (respectively  $h'_j \in \text{span}\{g_1, \dots, g_k\}$ ) and there exist neighborhoods  $U_i \in U(y_i)$  (respectively  $V_i \in U(y_i)$ ),  $i = 1, \dots, l$ , such that  $h_j$  (respectively  $h'_j$ ) is an  $\varepsilon_i$ -function on  $U_i$  (respectively  $V_i$ ),  $i = 1, \dots, l$ , and  $h_j(y_i) = 0$  (respectively  $h'_j(y_i) = 0$ ),  $i = l + 1, \dots, k$ . Now we can choose an integer  $n_1$  as we want, provided we choose  $n_1$  large enough, such that  $y_1 - 1/n_1 \in U_1 \cap V_1$ . If  $h'_j(y_1 - 1/n_1) = 0$ , then  $\tilde{g}_j = h'_j$  has the desired property. If  $h'_j(y_1 - 1/n_1) < 0$ , then there exists a scalar  $a \geq 0$  such that  $h_j(y_1 - 1/n_1) + ah'_j(y_1 - 1/n_1) = 0$  and  $\tilde{g}_j = h_j + ah'_j$  has the desired property. We remark that we obviously can choose an integer  $n_j$  as we want, provided we choose  $n_j$  large enough, such that  $\tilde{g}_j < 0$  on  $[y_j - 1/n_j, y_j] \subset U_j \cap V_j$ . Now we proceed similarly as in the proof of (9) to prove the induction step. This shows (11).

From (11) we immediately get

(12) For each  $j \in \{2, \dots, l\}$  we can choose integers  $n_1, \dots, n_l \in \{-1, 1\}$  as we want, provided we choose them large enough, such that there exists a function  $\tilde{g}_j \in \text{span}\{g_1, \dots, g_k\}$  with  $\tilde{g}_j < 0$  on  $[y_i - 1/n_i, y_i]$  and  $\tilde{g}_j(y_i - 1/n_i) = 0$ ,  $i = 1, \dots, l$ ,  $i \neq j$ , and  $\tilde{g}_j(y_i) = 0$ ,  $i = l + 1, \dots, k$ .

Now we choose integers  $n_1, \dots, n_l$  large enough such that  $g_1(x_1 - 1/n_1) < 0$ . Furthermore let  $\tilde{g}_{l+1} \in \text{span}\{g_1, \dots, g_k\}$  be the corresponding function to  $n_1, \dots, n_l$  which exists according to (10) and let  $\tilde{g}_2, \dots, \tilde{g}_l \in \text{span}\{g_1, \dots, g_k\}$  be the corresponding function to  $n_1, \dots, n_l$  which exists according to (12).

We show:

(13) The functions  $g_1, \tilde{g}_2, \dots, \tilde{g}_{l+1}$  are linearly independent.

*Proof of (13).* Let  $a_1, \tilde{a}_2, \dots, \tilde{a}_{l+1}$  be scalars such that  $a_1 g_1 + \tilde{a}_2 \tilde{g}_2 + \dots + \tilde{a}_{l+1} \tilde{g}_{l+1} = 0$ . Then from  $g_1(y_1 - 1/n_1) < 0$  and  $\tilde{g}_i(y_1 - 1/n_1) = 0$ ,  $i = 2, \dots, l + 1$ , (compare (10) and (12)) it follows that  $a_1 = 0$ . Then from  $\tilde{g}_2(y_2 - 1/n_2) < 0$  and  $\tilde{g}_i(y_2 - 1/n_2) = 0$ ,  $i = 3, \dots, l + 1$ , it follows that  $\tilde{a}_2 = 0$ . We continue this method and consider the remaining linear combinations at  $\{y_i - 1/n_i : i = 3, \dots, l\}$  and get as above  $\tilde{a}_i = 0$ ,  $i = 3, \dots, l$ . Finally we have  $\tilde{a}_{l+1} \tilde{g}_{l+1} = 0$ , from which we get  $\tilde{a}_{l+1} = 0$ , since  $\tilde{g}_{l+1} \neq 0$ . This shows (13).

(14) If  $l = k$  then from (13) it follows that  $g_1, \tilde{g}_2, \dots, \tilde{g}_{k+1}$  are linearly independent functions in  $\text{span}\{g_1, \dots, g_k\}$  which is a contradiction.



(15) If  $l < k$  then by (13) the functions  $g_1, \tilde{g}_2, \dots, \tilde{g}_{l+1}$  are linearly independent and by (10) and (12) have the property that  $g_i(y_j) = \tilde{g}_i(y_j) = 0, i = 2, \dots, l+1, j = l+1, \dots, k$ . Therefore there exist functions  $\tilde{g}_{l+2}, \dots, \tilde{g}_k \in \text{span} \{g_1, \dots, g_k\}$  such that  $g_1, \tilde{g}_2, \dots, \tilde{g}_{l+1}, \tilde{g}_{l+2}, \dots, \tilde{g}_k$  form a basis of  $\text{span} \{g_1, \dots, g_k\}$ . Now we consider the  $k-l$  linearly independent functions  $g_1, \tilde{g}_{l+2}, \dots, \tilde{g}_k$  at the  $k-l$  distinct points  $y_{l+1}, \dots, y_k$  and, since  $y_i \notin Z(g_1) \cap \bigcap_{i=l+2}^k Z(g_i), i = l+1, \dots, k$ , because  $y_i \notin \bigcap_{i=1}^k Z(g_i), i = l+1, \dots, k$ , and  $\tilde{g}_i(y_j) = 0, i = 2, \dots, l+1, j = l+1, \dots, k$ , we can conclude as in (1) and extend the function  $f$  at further points from  $\{y_{l+1}, \dots, y_k\}$  to reduce the dimension of  $P_G(f)$ , more precisely there exist  $p$  distinct points  $y'_1, \dots, y'_p \in \{y_1, \dots, y_k\}$ , where  $y'_1 = y_1$  and  $p < k$ , such that  $f$  is defined on  $\{x_1, \dots, x_n\} \setminus \{y'_3, \dots, y'_p\}$ , such that  $|f| = 1$  on  $\{x_1, \dots, x_n\} \setminus \{y'_3, \dots, y'_p\}$  and we have the choice to define  $f$  such that  $f(y'_2) = 1$  or  $f(y'_2) = -1$ , and  $P_G(f) \subset \text{span} \{g'_1, \dots, g'_p\}$ , where  $\{g'_1, \dots, g'_p\} \subset \{g_1, g_2, \dots, g_k\}$  are linearly independent with  $g'_1 = g_1$ , and  $y'_2 \in \bigcap_{i=1}^p Z(g'_i)$ . Notice that also  $y'_1 = y_1 \in \bigcap_{i=1}^p Z(g'_i)$ .

We continue this method by considering the functions  $\{g'_1, \dots, g'_p\}$  on  $\{y'_1, \dots, y'_p\}$ , instead of  $\{g_1, \dots, g_k\}$  on  $\{y_1, \dots, y_k\}$  as before, starting with (5). It can be easily verified that after a finite number of steps we get a contradiction as in (14).

Case 2. There do not exist  $n$  distinct points  $x_1 < \dots < x_n$  with  $z_i < x_i < z_{i+1}, i = 1, \dots, n$ , such that  $h_n$  changes sign at  $x_i, i = 1, \dots, n$ .

Let  $i \in \{1, \dots, n\}$  and consider  $(z_i, z_{i+1})$ . Then either there exists a point  $x_i \in (z_i, z_{i+1})$  such that  $h_n$  changes sign at  $x_i$  or there exist two distinct points  $x_i, \tilde{x}_i \in [z_i, z_{i+1}], x_i < \tilde{x}_i$ , and there exist neighborhoods  $U_i \in U(x_i)$  and  $\tilde{U}_i \in U(\tilde{x}_i)$ , such that  $h_n = 0$  on  $[x_i, \tilde{x}_i], h_n < 0$  (respectively  $h_n > 0$ ) on  $U_i \cap \{x \in [a, b]: x < x_i\}$  and  $h_n > 0$  (respectively  $h_n < 0$ ) on  $\tilde{U}_i \cap \{x \in [a, b]: x > \tilde{x}_i\}$ . In the "either"-case we set  $\tilde{x}_i = x_i$ .

Therefore we get  $n$  pairs of points  $(x_1, \tilde{x}_1), \dots, (x_n, \tilde{x}_n)$ . Obviously there exists an integer  $i \in \{1, \dots, n\}$  with  $x_i < \tilde{x}_i$ . Now we argue similarly as in Case 1. We first consider the basis  $\{h_1, \dots, h_n\}$  on  $\{x_1, \dots, x_n\}$  and conclude as in (1) to obtain a function  $f$ , defined on  $\{x_1, \dots, x_n\} \setminus \{y_2, \dots, y_k\}$ , where  $\{y_1, \dots, y_k\} \subset \{x_1, \dots, x_n\}$ , such that  $P_G(f) \subset \text{span} \{g_1, \dots, g_k\}$ , where  $g_1, \dots, g_k$  are linearly independent functions in  $\{h_1, \dots, h_n\}$  with  $g_1 = h_n, y_1 \in \bigcap_{i=1}^k Z(g_i)$  and we have the choice to define  $f$  such that  $f(y_1) = 1$  or  $f(y_1) = -1$ . Since each  $y_i$  is equal to some  $x_j$ , we can set  $\tilde{y}_i = \tilde{x}_j$ .

Let  $s$  be the number of points in  $\{y_1, \dots, y_k\}$  for which we have  $y_i \neq \tilde{y}_i$ . Then we have  $y_i = \tilde{y}_i$  for the remaining  $k-s$  points. We may assume that there exist at most  $k-1$  distinct points in  $\{y_1, \dots, y_k, \tilde{y}_1, \dots, \tilde{y}_k\}$  which are not in the set  $\bigcap_{i=1}^k Z(g_i)$ , because otherwise we can apply (1) to reduce the dimension of  $P_G(f)$  (provided we do not already have  $k=1$ ). Then, as can be easily verified, there exists at least one integer  $i \in \{1, \dots, n\}$  such that  $y_i = \tilde{y}_i$  and  $y_i \in \bigcap_{i=1}^k Z(g_i)$  (respectively  $y_i \neq \tilde{y}_i$  and  $y_i, \tilde{y}_i \in \bigcap_{i=1}^k Z(g_i)$ ).

By renumbering the points, if necessary, we may assume that  $i = 1$ . If  $k > 1$  we consider  $\tilde{y}_1, y_1, \dots, y_k$  and conclude similarly as in case 1 to get, after a finite number of steps, a contradiction as in (14).

If  $k = 1$  there exists a function  $f$  with  $\|f\| = 1, 0, g_1 \in P_G(f)$  and  $\dim P_G(f) = 1$ . Then, if there exists an integer  $i \in \{1, \dots, n\}$  with  $x_i = \tilde{x}_i$ , we get a contradiction as in Case 1. If not, there exists an integer  $i \in \{1, \dots, n\}$  such that  $x_i \neq \tilde{x}_i$  and  $f$  has not been defined on  $\tilde{x}_i$ . If  $f(x_i) = \varepsilon$ , where  $\varepsilon \in \{-1, 1\}$ , then we define  $f(\tilde{x}_i) = \varepsilon$ . By the lemma there exists a function  $g$  in  $P_G(f)$  and neighborhoods  $U_i \in U(x_i)$  and  $\tilde{U}_i \in U(\tilde{x}_i)$  with  $g \geq \max \{0, g_1\}$  on  $U_i \cup \tilde{U}_i$ , if  $\varepsilon = 1$ , respectively  $g \leq \min \{0, g_1\}$  on  $U_i \cup \tilde{U}_i$ , if  $\varepsilon = -1$ . But this is not possible, since  $P_G(f) = \text{span} \{g_1\}$ , and we get a contradiction. This completes the proof of the theorem.

With our theorem we solve part of a problem, posed by Lazar, Morris and Wulbert [4], for  $C[a, b]$ , who have proved the theorem for one-dimensional subspaces in  $C(X)$ ,  $X$  compact.

Furthermore from our theorem and results of Nürnberger and Sommer [7], and Sommer [10], it follows a complete characterization of those  $n$ -dimensional subspaces  $G$  in  $C[a, b]$  with the property that no  $g \in G$ ,  $g \neq 0$ , vanishes on an interval, which admit a continuous selection for  $P_G$ .

**THEOREM.** *Let  $G$  be an  $n$ -dimensional subspace in  $C[a, b]$  with the property that no  $g \in G$ ,  $g \neq 0$ , vanishes on an interval. Then the following statements are equivalent:*

- (i) *There exists a continuous selection for  $P_G$ .*
- (ii)  *$G$  is weak Chebyshev and each  $g \in G$ ,  $g \neq 0$ , has at most  $n$  distinct zeros in  $[a, b]$ .*

This gives us a complete solution of a problem, posed by Lazar, Morris and Wulbert [4], for the case of  $n$ -dimensional subspaces  $G$  in  $C[a, b]$ , which have the property that no  $g \in G$ ,  $g \neq 0$ , vanishes on an interval.

#### REFERENCES

- [1] A. L. BROWN, *On continuous selections for metric projections in spaces of continuous functions*, J. Functional Analysis, 8 (1970), pp. 431–449.
- [2] R. B. HOLMES, *A course on optimization and best approximation*, Lecture Notes 257, Springer Verlag, Berlin, 1972.
- [3] R. C. JONES AND L. A. KARLOVITZ, *Equioscillation under nonuniqueness in the approximation of continuous functions*, J. Approximation Theory, 3 (1970), pp. 138–145.
- [4] A. J. LAZAR, P. D. MORRIS AND D. E. WULBERT, *Continuous selections for metric projections*, J. Functional Analysis, 3 (1969), pp. 193–216.
- [5] G. NÜRNBERGER, *Schnitte für die metrische Projektion*, J. Approximation Theory, 20 (1977), pp. 196–220.
- [6] ———, *Continuous selections for the metric projection and alteration*, J. Approximation Theory, to appear.
- [7] G. NÜRNBERGER AND M. SOMMER, *Weak Chebyshev subspaces and continuous selections for the metric projection*, Trans. Amer. Math. Soc., 238 (1978), pp. 129–138.
- [8] ———, *Characterization of continuous selections of the metric projection for spline functions*, J. Approximation Theory, 22 (1978), pp. 320–330.
- [9] I. SINGER, *The theory of best approximation and functional analysis*, CBMS Regional Conference Series in Applied Mathematics, No. 4, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
- [10] M. SOMMER, *Nonexistence of continuous selections of the metric projection for a class of weak Chebyshev spaces*, Trans. Amer. Math. Soc., to appear.
- [11] ———, *Characterization of continuous selections of the metric projection for generalized splines*, this Journal, to appear.
- [12] L. P. VLASOV, *Approximative properties of sets in normed linear spaces*, Russian Math. Surveys, 28 (1973), pp. 1–66.
- [13] F. DEUTSCH, G. NÜRNBERGER AND I. SINGER, *Weak Chebyshev subspaces and alternation*, Pacific J. Math., to appear.

## STATISTICAL INDEPENDENCE AND INVARIANT SUBSPACES FOR POINT TRANSFORMATIONS\*

ALAN LAMBERT†

**Abstract.** Suppose  $(X, \Sigma, m)$  is a probability space and  $\tau$  is a measure preserving transformation on  $X$ . If  $\phi$  is strictly positive on  $X$  with  $E(\phi|\tau^{-1}\Sigma) = 1$  a.e. and  $\{\phi, \phi \circ \tau, \phi \circ \tau^2, \dots\}$  is an independent process, then  $\mathcal{E}$ , the closed  $L^2$  span of  $\{1, \phi, \phi \circ \tau, \dots\}$  is a proper reducing subspace for the point transformation  $Uf = f \circ \tau$ . Moreover  $\mathcal{E}^\perp$  is infinite dimensional. It is shown that  $\mathcal{E}$  is never invariant for the operator  $Tf = \phi \cdot f \circ \tau$ . Necessary and sufficient conditions are established for the invariance of  $\mathcal{E}$  for  $T^*$ .

**1. Introduction.** This paper is concerned with the relation between a family of independent, identically distributed random variables and invariant subspaces of certain operators on Hilbert space. The operators considered are convolution, or point, transformations. The existence of invariant and reducing subspaces for such operators is of importance in ergodic theory, especially subspaces which are the closure of the orbit of a single function under the transformation (i.e., a cyclic subspace). This paper develops the relationship between the i.i.d. process and the point transformation in terms of a weighted point transformation. We show that if  $\tau$  is a measure preserving mapping on the probability space  $(X, \Sigma, m)$ ,  $\phi$  is a strictly positive measurable function on  $X$  with  $E(\phi|\tau^{-1}\Sigma) = 1$  a.e., and  $\{\phi, \phi\tau, \phi \circ \tau^2, \dots\}$  is a statistically independent process, then  $\mathcal{E}$ , the closed  $L^2$  span of  $\{1, \phi, \phi \circ \tau, \dots\}$ , is a proper subspace of  $L^2$ . Further,  $\mathcal{E}$  is a reducing subspace for the point transformation  $Uf = f \circ \tau$  and  $\mathcal{E}^\perp$  is infinite dimensional. We also establish necessary and sufficient conditions for the invariance of  $\mathcal{E}^\perp$  under the weighted composition operator  $Tf = \phi f \circ \tau$  and show that  $\mathcal{E}$  is never invariant for  $T$ . The paper concludes with an example illustrating all these properties.

**2. Preliminaries.** Let  $(X, \Sigma, m)$  be a probability measure space and for each  $p \geq 1$ ,  $L^p = L^p(X, \Sigma, m)$  over  $\mathbb{C}$ .  $\mathcal{B}(L^p)$  is the ring of bounded linear transformations from  $L^p$  into  $L^p$ . We shall be concerned in this paper with weighted composition operators  $T = T_{\phi, \tau}$  defined as follows: Let  $\tau$  be a measurable mapping of  $X$  onto  $X$  such that the measure  $m \circ \tau^{-1}(A) = m(\tau^{-1}(A))$  is absolutely continuous with respect to  $m$ . Further, suppose  $\phi$  is a strictly positive measurable mapping of  $X$  to  $\mathbb{R}$ . Then we set  $Tf = \phi \cdot f \circ \tau$  whenever the resultant function is in the appropriate space. The properties and notation of weighted composition operators are developed in [2]. For coherence we briefly review some of the pertinent information. We set  $\tau^n$  to be the  $n$ -fold composition of  $\tau$  with itself and let  $E_n(f)$  be the conditional expectation of  $f$  with respect to  $\tau^{-n}\Sigma$ . Define  $\phi_n(x) = \prod_{k=0}^{n-1} \phi \circ \tau^k(x)$ . One sees easily that

$$T^n f = \phi_n \cdot f \circ \tau^n.$$

Throughout this paper we will be concerned only with the case that  $\tau$  is a measure preserving, i.e.,  $dm \cdot \tau^{-1}/dm = 1$  a.e.  $dm$ . In this case  $\|T\|_p$  (the norm of  $T$  in  $\mathcal{B}(L^p)$ ) is seen to be  $\|E_1(\phi^p)\|_\infty^{1/p}$ . The conservative set for  $T$  is

$$C(T) = \left\{ x \mid \sum_{n=0}^{\infty} (T^n 1)(x) = \infty \right\}$$

\* Received by the editors August 31, 1978, and in revised form August 13, 1979.

† Department of Mathematics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223.

(note that  $T^n 1 = \phi_n$ ) and  $D(T) = X - C(T)$  is the dissipative set for  $T$ . See [1; Chap. II] for a general discussion of these sets. A set  $S \in \Sigma$  is an *invariant set* for  $T$  if  $T^* 1_S = 1_S$ .  $1_S$  is the indicator function for  $S$ . The following results are proved in [2].

PROPOSITION 2.1 [2; 4.3].  $C(T)$  is the largest  $\tau$ -invariant subset of  $\{x | \phi(x) = 1\}$ .

PROPOSITION 2.2 [2; 3.2]. *The following are equivalent:*

- (a)  $T$  is an isometry on  $L^1$ ;
- (b)  $X$  is an invariant set for  $T$ ;
- (c)  $E_1(\phi) = 1$ .

We shall on occasion write expressions such as  $E_1(f) \circ \tau^{-1}$ . This is well defined because one easily sees that since  $E_1(f)$  is  $\tau^{-1}\Sigma$  measurable, there is a  $\Sigma$ -measurable function  $h$  such that  $E_1(f) = h \circ \tau$ .

Throughout this paper function statements are to be interpreted as true a.e.  $dm$  and set statements are to be interpreted as true modulo the  $m$ -null sets. For example if  $m(A \cap B) = 0$  we will simply write  $A \cap B = \emptyset$ .

We shall refer to the composition operator  $U$  given by  $Uf = f \circ \tau$ . This operator is isometric and doubly stochastic on every  $L^p$  space,  $1 \leq p < \infty$ . We shall often make use of the fact that for any measurable  $f$  and  $g$  and any  $K \geq 1$   $E_1(f \cdot g \circ \tau^K) = [E_1(f)]g \circ \tau^K$ . We reserve the notation  $E(f)$  for the mean,  $\int f dm$ .

It is shown in [2] that  $T^*f = E_1(\phi f) \circ \tau^{-1}$ .

**3. Independence and invariance.** We assume throughout that  $T = T_{\phi, \tau}$  is an isometry on  $L^1$ , bounded on  $L^2$ , and that  $\phi$  is not identically one. We assume further that  $\{\phi, \phi \circ \tau, \phi \circ \tau^2, \dots\}$  is a statistically independent process. Since  $\tau$  is measure preserving we see that they are identically distributed. Set  $\rho = \|\phi - 1\|_2$ ,

$$e_n(x) = \begin{cases} 1, & n = 0, \\ (1/\rho)(\phi \circ \tau^{n-1} - 1), & n \geq 1. \end{cases}$$

Since  $E(\phi \circ \tau^K) = E(\phi) = 1$  and  $\|\phi \circ \tau^K - 1\|_2 = \|\phi - 1\|_2 = \rho$ ,  $\{e_n\}_{n=0}^\infty$  is an orthonormal sequence in  $L^2$ . Let  $\mathcal{E}$  be its closed linear span. Of course  $\mathcal{E}$  is the span of  $\{1, \phi, \phi \circ \tau, \dots\}$  as well.

THEOREM 3.1.  $\mathcal{E}$  is a proper reducing subspace for  $U$ .

*Proof.* It is clear that  $\mathcal{E} \neq 0$  and  $U\mathcal{E} \subseteq \mathcal{E}$ . Moreover  $U^*e_0 = e_0$  and in general  $U^*f = (E_1f) \circ \tau^{-1}$ . Thus  $U^*\phi = E_1(\phi) \circ \tau^{-1} = 1 \in \mathcal{E}$  and  $U^*(\phi \circ \tau^{K+1}) = E_1(\phi \circ \tau^{K+1}) \circ \tau^{-1} = \phi \circ \tau^K \in \mathcal{E}$ . This proves that  $\mathcal{E}$  is invariant for both  $U$  and  $U^*$ . It remains to show that  $\mathcal{E} \neq L^2$ .

Let  $K$  be a nonnegative integer. Then

$$\begin{aligned} (T(\phi \circ \tau^K), e_0) &= \int \phi \cdot \phi \circ \tau^{K+1} dm \\ &= \left( \int \phi \right) \left( \int \phi \circ \tau^{K+1} \right) \\ &= \left[ \int \phi \right]^2 = 1; \\ (T(\phi \circ \tau^K), e_1) &= (1/\rho) \left[ \int \phi \cdot \phi \circ \tau^{K+1} \cdot (\phi - 1) \right] \\ &= (1/\rho) \left[ \int \phi(\phi - 1) \right] \left[ \int \phi \circ \tau^{K+1} \right] \\ &= (1/\rho) \left[ \int (\phi - 1)^2 \right] \end{aligned}$$

$$\begin{aligned}
 &= (1/\rho)\rho^2 = \rho; \\
 (T(\phi \circ \tau^K), e_{K+2}) &= (1/\rho) \left[ \int \phi \cdot \phi \circ \tau^{K+1} \cdot (\phi \circ \tau^{K+1} - 1) \right] \\
 &= (1/\rho) \left[ \int \phi \right] \left[ \int \phi \circ \tau^{K+1} \cdot (\phi \circ \tau^{K+1} - 1) \right] \\
 &= (1/\rho) \left[ \int \phi \cdot (\phi - 1) \right] = \rho.
 \end{aligned}$$

However, for  $n$  other than 0, 1, or  $K + 2$ ,

$$\begin{aligned}
 (T(\phi \circ \tau^K), e_n) &= (1/\rho) \left[ \int \phi \cdot \phi \circ \tau^{K+1} \cdot (\phi \circ \tau^{n-1} - 1) \right] \\
 &= (1/\rho) \left[ \int (\phi - 1) \right] = 0.
 \end{aligned}$$

Thus, if  $P$  is the orthogonal projection of  $L^2$  onto  $\mathcal{E}$ , we have for any  $K \geq 0$

$$\begin{aligned}
 P[T(\phi \circ \tau^K)] &= P(\phi \cdot \phi \circ \tau^{K+1}) \\
 &= 1 + \rho e_1 + \rho e_{K+2} \\
 &= 1 + (\phi - 1) + (\phi \circ \tau^{K+1} - 1) \\
 &= \phi + \phi \circ \tau^{K+1} - 1.
 \end{aligned}$$

We will now use these equations to show that for some  $K$ ,  $\phi \cdot \phi \circ \tau^{K+1}$  is not in  $\mathcal{E}$ . Indeed, suppose all these functions are in  $\mathcal{E}$ . Then for all  $n \geq 1$  we have  $\phi \cdot \phi \circ \tau^n = \phi + \phi \circ \tau^n - 1$ , or equivalently  $(\phi - 1)(\phi \circ \tau^n - 1) = 0$  for all  $n \geq 1$ . Let  $S = \{x | \phi(x) = 1\}$  and  $S_n = \{x | \phi \circ \tau^n(x) = 1\} = \tau^{-n}S$ . Set  $\mathcal{B} = \bigcap_{n=1}^\infty S_n$ . Then  $\emptyset \neq X - S \subseteq \mathcal{B}$ . Now  $\tau^{-1}\mathcal{B} = \tau^{-1} \bigcap_{n=1}^\infty \tau^{-n}S = \bigcap_{n=2}^\infty \tau^{-n}S \supseteq \mathcal{B}$ . But since  $\tau$  is measure preserving we have  $\tau^{-1}B = B$ . Now by Proposition 2.1  $C(T) \subseteq S$  hence  $X - S \subseteq D(T) \cap B$ . However on  $B$   $\phi_n(x) = \phi(x) \cdot \dots \cdot \phi \circ \tau^{n-1}(x) = \phi(x)$  so  $\sum_{n=1}^\infty \phi_n = \infty$  a.e. on  $B$ , i.e.,  $B \subseteq C(T)$ . Thus  $X - S \subseteq D(T) \cap C(T) = \emptyset$ . That is  $X = S$ . But by assumption  $\phi$  is not identically one. Therefore  $\mathcal{E}$  is a proper subspace and the proof is complete.

We have just seen that  $\mathcal{E}^\perp \neq 0$ . We now strengthen this result considerably.

**THEOREM 3.2.**  $\mathcal{E}^\perp$  is infinite dimensional.

*Proof.* We know that  $\mathcal{E}^\perp \neq 0$  and reduces  $U$ . Suppose  $\mathcal{E}^\perp$  is finite dimensional. Then  $\mathcal{E}^\perp$  has an orthonormal basis of eigenvectors for  $U$  [3; Chap. 3]. Moreover if  $f \neq 0$  and  $Uf = \alpha f$ , then  $|\alpha| = 1$  and  $f = \bar{\alpha}f \circ \tau$ . Also from the proof of Theorem 3.1, there is a  $K \geq 1$  such that  $0 \neq \phi \cdot \phi \circ \tau^K - P(\phi \cdot \phi \circ \tau^K) = (\phi - 1)(\phi \circ \tau^K - 1) \in \mathcal{E}^\perp$ . But if  $Uf = \alpha f$  then

$$\begin{aligned}
 ((\phi - 1)(\phi \circ \tau^K - 1), f) &= \alpha \int (\phi - 1)(\phi \circ \tau^K - 1)\bar{f} \circ \tau \\
 &= \alpha \int E_1(\phi - 1)(\phi \circ \tau^K - 1)(\bar{f} \circ \tau) \\
 &= 0.
 \end{aligned}$$

This is a contradiction; hence  $\mathcal{E}^\perp$  is infinite dimensional.

*Remark.* Theorem 3.2 may be replaced by the stronger statement that the eigenfunctions of  $U$  don't span  $\mathcal{E}^\perp$ .

We have seen that  $\mathcal{E}$  is invariant for  $U$  and  $U^*$  but definitely not invariant for  $T$ .

We will show that  $\mathcal{E}^\perp$  may be invariant for  $T$ .

**THEOREM 3.3.** *The following are equivalent:*

- (a)  $T\mathcal{E}^\perp \subseteq \mathcal{E}^\perp$ ;
- (b)  $T^*\mathcal{E} \subseteq \mathcal{E}$ ;
- (c)  $T^*\phi$  is constant;
- (d)  $T^*\phi$  is in  $\mathcal{E}$ .

*Proof.* Statements (a) and (b) are equivalent for any Hilbert space operator. We compute  $P(T^*\phi)$  as follows:

$$\begin{aligned} (T^*\phi, e_0) &= \int T^*\phi = \int E_1(\phi^2) \circ \tau^{-1} = \int E_1(\phi^2) = \int \phi^2; \\ (T^*\phi, e_{n+1}) &= (\phi, Te_{n+1}) = (1/\rho)(\phi, \phi \cdot [\phi \circ \tau^{n+1} - 1]) \\ &= (1/\rho) \left[ \int \phi^2 \cdot \phi \circ \tau^{n+1} - \int \phi^2 \right] \\ &= (1/\rho) \left[ \left( \int \phi^2 \right) \left( \int \phi \right) - \int \phi^2 \right] = 0. \end{aligned}$$

Thus  $P(T^*\phi) = \int \phi^2$  a.e. so (b) implies (d) and (d) implies (c). The equivalence cycle will be closed if we show that (c) implies (b). We now proceed to do so. Assume (c). Since  $T^*e_0 = e_0$  we have as well that  $T^*e_1 = (1/\rho)(T^*\phi - T^*e_0)$  is in  $\mathcal{E}$ . But for  $n \geq 2$ ,

$$\begin{aligned} T^*e_n &= (1/\rho)E_1(\phi[\phi \circ \tau^{n-1} - 1]) \circ \tau^{-1} \\ &= (1/\rho)[E_1(\phi \cdot \phi \circ \tau^{n-1}) \circ \tau^{-1} - 1] \\ &= (1/\rho)\{[E_1(\phi) \circ \tau^{-1}]\phi \circ \tau^{n-2} - 1\} \\ &= (1/\rho)(\phi \circ \tau^{n-2} - 1) \\ &= e_{n-1}. \end{aligned}$$

This shows that if  $T^*\phi$  is in  $\mathcal{E}$  then  $T^*\mathcal{E} \subseteq \mathcal{E}$ , and the proof is complete.

**COROLLARY 3.4.**  $T^*\mathcal{E} \subseteq \mathcal{E}$  if and only if  $T$  is a scalar multiple of an isometry on  $L^2$ .

*Proof.* Since

$$\begin{aligned} \|Tf\|_2^2 &= \int \phi^2 |f \circ \tau|^2 \\ &= \int [E_1(\phi^2) \circ \tau^{-1}] |f|, \end{aligned}$$

$T$  is a scalar multiple of an isometry if and only if  $T^*\phi = E_1(\phi^2) \circ \tau^{-1}$  is constant almost everywhere. By Theorem 3.3 this holds if and only if  $T^*\mathcal{E} \subseteq \mathcal{E}$ .

The following example shows how all the restrictions placed on  $\phi$  in the above results may occur. We use Example 5.1 of [2].

*Example.* Let  $X = [0, 1]$  and let  $m$  be Lebesgue measure. Let  $\tau$  be given by

$$\tau(x) = \begin{cases} 2x, & 0 \leq x \leq \frac{1}{2}, \\ 2 - 2x, & \frac{1}{2} < x \leq 1 \end{cases}$$

and

$$\phi(x) = \begin{cases} \frac{3}{2}, & 0 \leq x \leq \frac{1}{2}, \\ \frac{1}{2}, & \frac{1}{2} < x \leq 1. \end{cases}$$

Then  $T = T_{\phi, \tau}$  is an isometry on  $L^1$  and  $\tau$  is measure preserving. We show first that  $\phi, \phi \circ \tau, \dots$  are statistically independent. Let  $f_i = \phi \circ \tau^i, i = 0, 1, \dots$ . It suffices to show that for any finite collections  $A_1, \dots, A_n$  of (nonempty) open intervals and any collection  $K_1, K_2, \dots, K_n$  of nonnegative integers with only  $K_1$  possibly 0, that

$$m(f_{K_1}^{-1}(A_1) \cap f_{K_1+K_2}^{-1}(A_2) \cap \dots \cap f_{K_1+\dots+K_n}^{-1}(A_n)) = m(f_{K_1}^{-1}(A_1)) \cdot \dots \cdot m(f_{K_1+\dots+K_n}^{-1}(A_n)).$$

Note that  $f_K^{-1}(A) = \tau^{-K}(\phi^{-1}(A))$ . We proceed by induction on  $n$ . Suppose  $n = 2$ . We must show that

$$(1) \quad m(\tau^{-K_1}(\phi^{-1}(A_1)) \cap \tau^{-(K_1+K_2)}(\phi^{-1}(A_2))) = m(\tau^{-K_1}(\phi^{-1}(A_1))) \cdot m(\tau^{-(K_1+K_2)}(\phi^{-1}(A_2))).$$

This last expression reduces to  $m(\phi^{-1}(A_1)) \cdot m(\phi^{-1}(A_2))$  since  $\tau$  is measure preserving. The left-hand side of the above equation may be rewritten as

$$(2) \quad m(\phi^{-1}(A_1) \cap \tau^{-K_2}(\phi^{-1}(A_2))).$$

Note that for any interval  $A$ ,  $\phi^{-1}(A)$  is one of the four sets  $\emptyset, [0, 1], [0, \frac{1}{2}], [\frac{1}{2}, 1]$ . If  $\phi^{-1}(A_1)$  is either  $\emptyset$  or  $[0, 1]$  (1) holds. Moreover, if (1) holds whenever  $\phi^{-1}(A_1) = [0, \frac{1}{2}]$  then it is easy to see it holds whenever  $\phi^{-1}(A_1)$  is  $[\frac{1}{2}, 1]$ . Thus we assume  $\phi^{-1}(A_1)$  is  $[0, \frac{1}{2}]$ . Then (2) becomes  $m([0, \frac{1}{2}] \cap \tau^{-K_2}(\phi^{-1}(A_2)))$ . But  $K_2 \geq 1$  and the graph of  $\tau$  (hence  $\tau^K, K \geq 1$ ) is symmetric about  $\frac{1}{2}$ . Thus

$$m([0, \frac{1}{2}] \cap \tau^{-K_2}(\phi^{-1}(A_2))) = \frac{1}{2}m(\tau^{-K_2}(\phi^{-1}(A_2))) = \frac{1}{2}m(\phi^{-1}(A_2)),$$

showing (1) holds.

Now assume the result holds for  $n - 1$ . As before we assume  $\phi^{-1}(A_1) = [0, \frac{1}{2}]$ . Then

$$\begin{aligned} m(\phi^{-1}(A_1) \cap \tau^{-K_{2-1}}(A_2) \cap \dots \cap \tau^{-(K_2+\dots+K_n)}(A_n)) \\ = m([0, \frac{1}{2}] \cap \tau^{-K_2}[\phi^{-1}(A_2) \cap \dots \cap \tau^{-(K_3+\dots+K_n)}(A_n)]) \\ = \frac{1}{2}m(\phi^{-1}(A_2) \cap \dots \cap \tau^{-(K_3+\dots+K_n)}(A_n)). \end{aligned}$$

This last expression is, according to the induction hypothesis,  $\frac{1}{2}m(\phi^{-1}(A_2)) \cdot \dots \cdot m(\phi^{-1}(A_n))$ . The proof of independence is complete.

This example in fact illustrates the invariance of  $\mathcal{E}$  under  $T^*$ . By Theorem 3.3 we need only show that  $E_1(\phi^2)$  is constant. Routine calculation shows that for any  $f$

$$E_1(f) = \frac{f(x) + f(1-x)}{2}.$$

Now

$$\phi^2(x) = \begin{cases} \frac{9}{4}, & 0 \leq x \leq \frac{1}{2}, \\ \frac{1}{4}, & \frac{1}{2} < x \leq 1 \end{cases}$$

so  $E_1(\phi^2) = (\frac{9}{4} + \frac{1}{4})/2 = \frac{5}{4}$  on  $[0, 1]$ . Using other step functions and the same  $\tau$  as in the preceding example one may generate examples where  $E_1(\phi) = 1$  but  $\{\phi, \phi \circ \tau, \phi \circ \tau^2, \dots\}$  is not an independent process. It would be interesting to see if the resulting space  $\mathcal{E}$  is proper under looser hypotheses than those imposed in this paper.

**Acknowledgment.** The author wishes to express his gratitude to Joseph Quinn and Barnet Weinstock for their helpful suggestions related to this work.

REFERENCES

[1] SHAUL R. FOGUEL, *The Ergodic Theory of Markov Processes*, Van Nostrand-Reinhold, New York, 1969.  
 [2] T. HOOVER, A. LAMBERT AND J. QUINN, *The Markov Process Determined by a Weighted Composition Operator*, preprint.  
 [3] PETER WALTERS, *Ergodic Theory-Introduction Lectures*, Springer-Verlag Lecture Notes in Mathematics, No. 458, Springer-Verlag, New York, 1975.

## HOMOGENEOUS GENERALIZED TEMPERATURES\*

*Dedicated to the memory of Joaquin B. Diaz*

DEBORAH TEPPER HAIMO†

**Abstract.** For the generalized heat equation  $u_{xx} + (2v/x)u_x = u_t$ ,  $v > 0$ , there are two basic kinds of homogeneous solutions of nonnegative integral degree  $n$  and two of negative degree  $-n - 1 - 2v$ ,  $0 < v < \frac{1}{2}$ . Criteria are established for the expansion of generalized homogeneous temperatures in series of these basic solutions. The results parallel, in part, those of D. V. Widder for the classical heat equation.

**1. Introduction.** In a number of papers [9]–[12], D. V. Widder studied those solutions of the classical heat equation  $u_{xx} = u_t$  that are homogeneous of either positive or negative integral degree. He determined that there are two fundamental kinds of each, and established various criteria for the representation of any solution of the heat equation in series of these basic homogeneous solutions.

We seek to extend some of the Widder results to the generalized heat equation  $u_{xx} + (2v/x)u_x = u_t$ ,  $v > 0$ , where the homogeneous solutions to be considered will be of nonnegative integral degree  $n$  or of negative degree  $\lambda = -n - 1 - 2v$ ,  $0 < v < \frac{1}{2}$ .

**2. Definitions and preliminary results.** The generalized heat equation is given by

$$(2.1) \quad \Delta_x u(x, t) = \frac{\partial}{\partial t} u(x, t),$$

where the operator  $\Delta_x$  is defined by

$$(2.2) \quad \Delta_x f(x) = f''(x) + \frac{2v}{x} f'(x), \quad v > 0.$$

We denote by  $H$  the class of all  $C^2$  solutions of (2.1), and call a member  $u(x, t)$  of  $H$  a generalized temperature.

The fundamental solution of (2.1) is the function

$$(2.3) \quad G(x; t) = \left(\frac{1}{2t}\right)^{\nu+(1/2)} e^{-(x^2/4t)},$$

and its associated function is

$$(2.4) \quad G(x, y; t) = \left(\frac{1}{2t}\right)^{\nu+(1/2)} e^{-((x^2+y^2)/4t)} \mathcal{J}\left(\frac{xy}{2t}\right),$$

where

$$(2.5) \quad \mathcal{J}(z) = 2^{\nu-(1/2)} \Gamma(\nu + \frac{1}{2}) z^{(1/2)-\nu} I_{\nu-(1/2)}(z),$$

$I_\alpha(z)$  being the modified Bessel function of order  $\alpha$ .

We denote by  $H^*$  the class of all those members  $u(x, t)$  of  $H$  for  $a < t < b$  which, for all  $t, t'$  with  $a < t' < t < b$ , have the semigroup property

$$(2.6) \quad u(x, t) = \int_0^\infty G(x, y; t-t') u(y, t') d\mu(y),$$

$$d\mu(y) = \frac{y^{2\nu} dy}{2^{\nu-(1/2)} \Gamma(\nu + \frac{1}{2})},$$

\* Received by the editors October 3, 1978, and in revised form August 13, 1979.

† Department of Mathematics, University of Missouri, St. Louis, Missouri 63121.



the integral converging absolutely for  $a < t < b$ . A member of  $H^*$  is said to have the Huygens property.

We introduce the Appell transform  $u^A(x, t)$  of a generalized temperature  $u(x, t)$  given by

$$(2.7) \quad u^A(x, t) = G(x; t)u(x/t, -1/t).$$

It is well known that if  $u(x, t) \in H^*$  for  $0 < a < t < b$ , then  $u^A(x, t) \in H^*$  for  $-1/a < t < -1/b$ . See, for example, [8, Lemma 2.6].

A generalized temperature  $u(x, t)$  is said to be homogeneous of degree  $\alpha$  if, for any  $\lambda > 0$ ,

$$(2.8) \quad u(\lambda x, \lambda^2 t) = \lambda^\alpha u(x, t).$$

The definition corresponds to that of the homogeneity of  $u(x, t^2)$  in the classical analytic sense. We denote by  $M_\alpha$  the class of all homogeneous generalized temperatures of degree  $\alpha$ .

The generalized heat polynomials are the polynomials

$$(2.9) \quad P_{n,\nu}(x, t) = \sum_{k=0}^n 2^{2k} \binom{n}{k} \frac{\Gamma(\nu + \frac{1}{2} + n)}{\Gamma(\nu + \frac{1}{2} + n - k)} x^{2n-2k} t^k,$$

and their Appell transforms are the functions

$$(2.10) \quad W_{n,\nu}(x, t) = G(x; t)P_{n,\nu}(x/t, -1/t).$$

In earlier papers [3]–[8], various criteria, particularly those involving membership in  $H^*$ , were established for the representation of generalized temperatures in series of the generalized heat polynomials  $P_{n,\nu}(x, t)$  and of their Appell transforms  $W_{n,\nu}(x, t)$ . We note here that the  $P_{n,\nu}(x, t)$  are homogeneous of degree  $2n$ , whereas the  $W_{n,\nu}(x, t)$  are of homogeneity  $-2n - 1 - 2\nu$ .

Function theoretic properties will be central in the characterization of generalized temperatures that are expandable in series of basic homogeneous generalized temperatures. To this end, we introduce the class  $\{\sigma, \tau\}$  of all entire functions  $\varphi$  that are of order at most  $\sigma$ , and if of order  $\sigma$ , of type at most  $\tau$ . A function  $\varphi \in \{\sigma, \tau\}$  is said to have growth  $\{\sigma, \tau\}$ . We note that for  $\varphi \in \{\sigma, \tau\}$  with  $\varphi$  of order less than  $\sigma$ ,  $\tau = 0$ . Hence we have  $e^{\tau z} \in \{1, \tau\}$ , and also,  $e^{\tau z} \in \{1 + \varepsilon, 0\}$  for every  $\varepsilon > 0$ .

It is clear that a function  $\varphi$  has growth  $\{\sigma, \tau\}$  if and only if, for any  $\varepsilon > 0$ ,

$$(2.11) \quad \varphi(z) = O(e^{(\tau+\varepsilon)|z|^\sigma}), \quad |z| \rightarrow \infty.$$

Further, if  $\varphi$  has the Maclaurin expansion

$$(2.12) \quad \varphi(z) = \sum_{n=0}^{\infty} a_n z^n,$$

then  $\varphi \in \{\sigma, \tau\}$  if and only if

$$(2.13) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{\sigma/n} \leq e\sigma\tau.$$

**3. Basic homogeneous temperatures.** A function  $u(x, t)$  belongs to class  $M_n$ ,  $n$  a positive integer, if and only if, for  $\lambda > 0$ ,

$$(3.1) \quad u(\lambda x, \lambda^2 t) = \lambda^n u(x, t).$$

By choosing  $\lambda = 1/\sqrt{2t}$ , we have

$$u(x/\sqrt{2t}, 1/2) = (1/\sqrt{2t})^n u(x, t),$$

or

$$(3.2) \quad u(x, t) = (2t)^{n/2} f(z),$$

where

$$(3.3) \quad f(z) = u(z, 1/2), \quad z = x/\sqrt{2t}.$$

Noting that  $u(x, t)$  in (3.2) satisfies the generalized heat equation  $\Delta_x u = u_t$ , we find that  $f(z)$  must satisfy the ordinary linear differential equation

$$(3.4) \quad f''(z) + \left(\frac{2\nu}{z} + z\right) f'(z) - n f(z) = 0.$$

If  $f_1(z)$  and  $f_2(z)$  are linearly independent solutions of (3.4), then all functions  $u(x, t)$  of  $M_n$  must be of the form

$$u(x, t) = (2t)^{n/2} [c_1 f_1(z) + c_2 f_2(z)],$$

with  $c_1, c_2$  arbitrary constants, or of the form

$$(3.5) \quad u(x, t) = c_1 u_1(x, t) + c_2 u_2(x, t),$$

where  $u_1(x, t) = (2t)^{n/2} f_1(z)$ ,  $u_2(x, t) = (2t)^{n/2} f_2(z)$  are a linearly independent set of homogeneous generalized temperatures of degree  $n$ . To determine such functions, we recall that the generalized heat polynomials  $P_{n,\nu}(x, t)$  are homogeneous of degree  $2n$  and, as established in [5, Lemma 2.1], have the integral representation

$$(3.6) \quad P_{n,\nu}(x, t) = \int_0^\infty G(x, y; t) y^{2n} d\mu(y), \quad t > 0.$$

This suggests the introduction of the functions

$$(3.7) \quad S_{n,\nu}(x, t) = \int_0^\infty G(x, y; t) y^n d\mu(y), \quad t > 0,$$

which clearly belong to  $M_n$ , and which we take as  $u_1(x, t)$ . We call these functions basic homogeneous generalized temperatures of degree  $n$  of the first kind.

To determine functions  $u_2(x, t)$  that will satisfy our needs, we seek ones that will have symmetry of form with  $u_1(x, t)$ . To this end, we define, for  $0 < \nu < \frac{1}{2}$ ,

$$(3.8) \quad \mathcal{H}(z) = \frac{2^{\nu+(1/2)}}{\Gamma(\frac{1}{2}-\nu)} z^{(1/2)-\nu} K_{\nu-(1/2)}(z),$$

where  $K_\alpha(z)$  is the modified Bessel function of the third kind. In addition, we introduce the function

$$(3.9) \quad H(x, y; t) = \left(\frac{1}{2t}\right)^{\nu+(1/2)} e^{-((x^2+y^2)/4t)} \mathcal{H}\left(\frac{xy}{2t}\right),$$

the modified associated fundamental solution. We then take for  $u_2(x, t)$  the functions

$$(3.10) \quad R_{n,\nu}(x, t) = \int_0^\infty H(x, y; t) y^n d\mu(y), \quad t > 0.$$

readily confirmed to be linearly independent of  $S_{n,\nu}(x, t)$ . These functions belong to  $M_n$  and will be called homogeneous generalized temperatures of degree  $n$  of the second kind.

We thus have the following result.

**THEOREM 3.1.** *A function  $u(x, t)$  is a homogeneous generalized temperature of degree  $n$  if and only if*

$$(3.11) \quad u(x, t) = AS_{n,\nu}(x, t) + BR_{n,\nu}(x, t), \quad t > 0,$$

$A, B$  arbitrary constants.

For homogeneous generalized temperatures of negative degree  $-n-1-2\nu$ , we introduce the functions

$$(3.12) \quad s_{n,\nu}(x, t) = \int_0^\infty y^n e^{ty^2} \mathcal{F}(xy) d\mu(y), \quad t < 0,$$

of the first kind, and

$$(3.13) \quad r_{n,\nu}(x, t) = \int_0^\infty y^n e^{ty^2} \mathcal{K}(xy) d\mu(y), \quad t < 0,$$

of the second kind.

As may be readily verified, the Appell transformation establishes a duality between homogeneous generalized temperatures of positive degree and those of negative degree. For the basic homogeneous generalized temperatures, we have, in particular, the following result.

**THEOREM 3.2.** *For  $n = 0, 1, 2, \dots$ ,*

$$(3.14) \quad s_{n,\nu}^A(x, t) = 2^{-n-1-2\nu} S_{n,\nu}(x, t),$$

$$(3.15) \quad r_{n,\nu}^A(x, t) = 2^{-n-1-2\nu} R_{n,\nu}(x, t).$$

As a consequence of Theorems 3.1 and 3.2, we obtain the totality of homogeneous generalized temperatures of degree  $-n-1-2\nu$ .

**THEOREM 3.3.** *A function  $u(x, t)$  is a homogeneous generalized temperature of degree  $-n-1-2\nu$ ,  $0 < \nu < \frac{1}{2}$ , if and only if*

$$u(x, t) = as_{n,\nu}(x, t) + br_{n,\nu}(x, t), \quad t < 0,$$

$a, b$  arbitrary constants.

**4. Properties of the basic homogeneous generalized temperatures.** The operator  $\Delta_x$  reduces the degree of homogeneity of a homogeneous function by two. In particular,  $\Delta_x$  transforms the basic homogeneous generalized temperatures of negative degree into themselves, but those of positive degree into combinations of themselves. A straightforward computation yields the following result.

**THEOREM 4.1.** *For  $n = 0, 1, 2, \dots$ ,*

$$(4.1) \quad \Delta_x S_{n,\nu}(x, t) = \left( \frac{x^2}{4t^2} + \frac{2n+2\nu+1}{2t} \right) S_{n,\nu}(x, t) - \frac{1}{4t^2} S_{n+2,\nu}(x, t), \quad t > 0;$$

$$(4.2) \quad \Delta_x R_{n,\nu}(x, t) = \left( \frac{x^2}{4t^2} + \frac{2n+2\nu+1}{2t} \right) R_{n,\nu}(x, t) - \frac{1}{4t^2} R_{n+2,\nu}(x, t), \quad t > 0,$$

$$(4.3) \quad \Delta_x s_{n,\nu}(x, t) = s_{n+2,\nu}(x, t), \quad t < 0,$$

$$(4.4) \quad \Delta_x r_{n,\nu}(x, t) = r_{n+2,\nu}(x, t), \quad t < 0.$$

Recurrence relations may likewise be derived for the basic homogeneous generalized temperatures as given in the following theorem.

THEOREM 4.2. For  $n = 0, 1, 2, \dots$ ,

$$(4.5) \quad \begin{aligned} S_{n+4,\nu}(x, t) &= [x^2 + 2t(2n + 2\nu + 5)]S_{n+2,\nu}(x, t) \\ &\quad - 4t^2(n + 2)(n + 2\nu + 1)S_{n,\nu}(x, t), \quad t > 0, \end{aligned}$$

$$(4.6) \quad \begin{aligned} R_{n+4,\nu}(x, t) &= [x^2 + 2t(2n + 2\nu + 5)]R_{n+2,\nu}(x, t) \\ &\quad - 4t^2(n + 2)(n + 2\nu + 1)S_{n,\nu}(x, t), \quad t > 0, \end{aligned}$$

$$(4.7) \quad \begin{aligned} s_{n+4,\nu}(x, t) &= \frac{1}{4t^2} \{ [x^2 - 2t(2n + 2\nu + 5)]s_{n+2,\nu}(x, t) \\ &\quad - (n + 2)(n + 2\nu + 1)s_{n,\nu}(x, t) \}, \quad t < 0, \end{aligned}$$

$$(4.8) \quad \begin{aligned} r_{n+4,\nu}(x, t) &= \frac{1}{4t^2} \{ [x^2 - 2t(2n + 2\nu + 5)]r_{n+2,\nu}(x, t) \\ &\quad - (n + 2)(n + 2\nu + 1)r_{n,\nu}(x, t) \}, \quad t < 0. \end{aligned}$$

We note that by evaluating the four functions  $s_{0,\nu}(x, t)$ ,  $s_{1,\nu}(x, t)$ ,  $t_{0,\nu}(x, t)$  and  $r_{1,\nu}(x, t)$ , and appealing to Theorems 3.2, 4.1 and 4.2, we may recursively derive all the basic homogeneous generalized temperatures  $s_{n,\nu}(x, t)$ ,  $r_{n,\nu}(x, t)$ ,  $S_{n,\nu}(x, t)$  and  $R_{n,\nu}(x, t)$ . The values of these four functions are

$$(4.9) \quad s_{0,\nu}(x, t) = \left(-\frac{1}{2t}\right)^{\nu+(1/2)} e^{-x^2/(4t)}, \quad t < 0,$$

$$(4.10) \quad s_{1,\nu}(x, t) = \frac{\Gamma(\nu + 1)}{2^{\nu+(1/2)}\Gamma(\nu + \frac{1}{2})} \left(-\frac{1}{t}\right)^{\nu+1} {}_1F_1\left(\nu + 1; \nu + \frac{1}{2}; -\frac{x^2}{4t}\right), \quad t < 0,$$

where  ${}_1F_1(a, c; x)$  is the confluent hypergeometric function,

$$(4.11) \quad r_{0,\nu}(x, t) = \left(-\frac{1}{2t}\right)^{\nu+(1/2)} e^{-x^2/(4t)} \frac{\Gamma(\frac{1}{2} - \nu, -x^2/(4t))}{\Gamma(\frac{1}{2} - \nu)}, \quad t < 0,$$

where  $\Gamma(a, x)$  is the incomplete gamma function  $\int_x^\infty e^{-u} u^{a-1} du$ ,

$$(4.12) \quad r_{1,\nu}(x, t) = s_{1,\nu}(x, t) - \frac{2^{\nu-(3/2)}\Gamma(\frac{3}{2})}{\Gamma(\frac{3}{2} - \nu)} x^{1-2\nu} (-t)^{-3/2} {}_1F_1\left(\frac{3}{2}; \frac{3}{2} - \nu; -\frac{x^2}{4t}\right), \quad t < 0.$$

By an appeal to [2, p. 197(20) and p. 199(37)], the integrals defining the basic homogeneous generalized temperatures may be evaluated explicitly. We then have

$$(4.13) \quad S_{n,\nu}(x, t) = 2^n \frac{\Gamma(n/2 + \nu + \frac{1}{2})}{\Gamma(\nu + \frac{1}{2})} t^{n/2} {}_1F_1\left(-\frac{n}{2}; \nu + \frac{1}{2}; -\frac{x^2}{4t}\right),$$

$$(4.14) \quad \begin{aligned} R_{n,\nu}(x, t) &= S_{n,\nu}(x, t) - 2^{n+2\nu-1} \frac{\Gamma(n/2 + 1)}{\Gamma(\frac{3}{2} - \nu)} x^{1-2\nu} t^{n/2 + \nu - (1/2)} \\ &\quad \cdot {}_1F_1\left(-\frac{n}{2} - \nu + \frac{1}{2}; \frac{3}{2} - \nu; -\frac{x^2}{4t}\right), \end{aligned}$$

$$(4.15) \quad s_{n,\nu}(x, t) = (-t)^{-(n/2 + \nu + (1/2))} 2^{-\nu - (1/2)} \frac{\Gamma(n/2 + \nu + \frac{1}{2})}{\Gamma(\nu + \frac{1}{2})} {}_1F_1\left(\frac{n}{2} + \nu + \frac{1}{2}; \nu + \frac{1}{2}; -\frac{x^2}{4t}\right),$$

$$(4.16) \quad r_{n,\nu}(x, t) = s_{n,\nu}(x, t) - 2^{\nu - (3/2)} \frac{\Gamma(n/2 + 1)}{\Gamma(\frac{3}{2} - \nu)} x^{1-2\nu} (-t)^{-n/2 - 1} {}_1F_1\left(\frac{n}{2} + 1; \frac{3}{2} - \nu; -\frac{x^2}{4t}\right),$$

with these equations serving to define the functions  $S_{n,\nu}, R_{n,\nu}, s_{n,\nu}, r_{n,\nu}$  for all  $t$ .

It is then immediate that, for  $t > 0$ ,

$$(4.17) \quad S_{n,\nu}(x, -t) = e^{(n/2)\pi i} S_{n,\nu}(ix, t),$$

$$(4.18) \quad R_{n,\nu}(x, -t) = e^{(n/2)\pi i} R_{n,\nu}(ix, t),$$

$$(4.19) \quad s_{n,\nu}(x, -t) = e^{(n+2\nu+1)(\pi i/2)} s_{n,\nu}(ix, t),$$

$$(4.20) \quad r_{n,\nu}(x, -t) = e^{(n+2\nu+1)(\pi i/2)} r_{n,\nu}(ix, t).$$

**5. Estimates.** From definition (2.4), it is clear that, for any  $\delta > 0$ ,

$$(5.1) \quad G(x, y, t) = \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{x^2/(4\delta)-y^2/(4(t+\delta))} G\left(x\frac{t+\delta}{\delta}, y; \frac{t(t+\delta)}{\delta}\right).$$

We then have that, for  $t > 0$ ,

$$(5.2) \quad \begin{aligned} S_{n,\nu}(x, t) &= \int_0^\infty G(x, y; t) y^n d\mu(y) \\ &= \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{x^2/(4\delta)} \int_0^\infty G\left(x\frac{t+\delta}{\delta}, y; \frac{t(t+\delta)}{\delta}\right) e^{-y^2/(4(t+\delta))} y^n d\mu(y). \end{aligned}$$

Appealing to the inequality

$$(5.3) \quad y^n e^{-Ay^2} \leq \left(\frac{n}{2Ae}\right)^{n/2},$$

we find that

$$(5.4) \quad \begin{aligned} S_{n,\nu}(x, t) &\leq \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{x^2/(4\delta)} \left[\frac{2n(t+\delta)}{e}\right]^{n/2} \int_0^\infty G\left(x\frac{t+\delta}{\delta}, y; \frac{t(t+\delta)}{\delta}\right) d\mu(y) \\ &= \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{x^2/(4\delta)} \left[\frac{2n(t+\delta)}{e}\right]^{n/2} S_{0,\nu}\left(x\frac{t+\delta}{\delta}, \frac{t(t+\delta)}{\delta}\right). \end{aligned}$$

We have, however, that

$$(5.5) \quad S_{0,\nu}(x, t) = 1.$$

Hence

$$(5.6) \quad S_{n,\nu}(x, t) \leq \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{x^2/(4\delta)} \left[\frac{2n(t+\delta)}{e}\right]^{n/2}.$$

We have, further, the readily established inequality

$$(5.7) \quad |\mathcal{H}(x)| \leq \mathcal{F}(x) \left[1 + \left|\frac{x}{2}\right|^{1-2\nu}\right]$$

so that, noting definitions (2.4) and (3.9), we have

$$(5.8) \quad |H(x, y; t)| \leq G(x, y; t) \left[1 + \left|\frac{xy}{4t}\right|^{1-2\nu}\right]$$

whence it follows that, for  $t > 0$ ,

$$(5.9) \quad |R_{n,\nu}(x, t)| \leq S_{n,\nu}(x, t) + \left(\frac{|x|}{4t}\right)^{1-2\nu} S_{n+1-2\nu,\nu}(x, t).$$

On the basis of (4.17), (4.18), (5.6) and (5.9), we have the following result.

**THEOREM 5.1.** For  $t > 0$  and  $\delta > 0$ ,

$$(5.10) \quad |S_{n,\nu}(x, \pm t)| \cong \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{\pm x^2/(4\delta)} \left[\frac{2n(t+\delta)}{e}\right]^{n/2},$$

$$(5.11) \quad |R_{n,\nu}(x, \pm t)| \cong \left(\frac{t+\delta}{\delta}\right)^{\nu+(1/2)} e^{\pm x^2/(4\delta)} \left[\frac{2n(t+\delta)}{e}\right]^{n/2} \cdot \left[1 + \left(\frac{|x|}{4t}\right)^{1-2\nu} \left(\frac{2(n+1-2\nu)(t+\delta)}{e}\right)^{(1/2)-\nu} \left(\frac{n+1+2\nu}{n}\right)^{n/2}\right].$$

From (4.13)–(4.16), noting that

$$(5.12) \quad {}_1F_1(a, c; x) = e^x {}_1F_1(c-a, c; -x),$$

we have

$$(5.13) \quad s_{n,\nu}(x, t) = \left(-\frac{1}{2t}\right)^{n+\nu+(1/2)} e^{-x^2/(4t)} S_{n,\nu}(x, -t),$$

$$(5.14) \quad r_{n,\nu}(x, t) = \left(-\frac{1}{2t}\right)^{n+\nu+(1/2)} e^{-x^2/(4t)} R_{n,\nu}(x, -t).$$

From these equalities and Theorem 5.1, we have the following estimates for the basic homogeneous generalized temperatures of negative degree.

**THEOREM 5.2.** For  $t > 0$  and  $\delta > 0$

$$(5.15) \quad |s_{n,\nu}(x, \pm t)| \cong \left(\frac{t+\delta}{2\delta t}\right)^{\nu+(1/2)} e^{\mp x^2(t+\delta)/(4t\delta)} \left[\frac{n(t+\delta)}{2et^2}\right]^{n/2},$$

$$(5.16) \quad |r_{n,\nu}(x, \pm t)| \cong \left(\frac{t+\delta}{2\delta t}\right)^{\nu+(1/2)} e^{\mp x^2(t+\delta)/(4t\delta)} \left[\frac{n(t+\delta)}{2et^2}\right] \cdot \left[1 + \left(\frac{|x|}{4t}\right)^{1-2\nu} \left(\frac{2(n+1-2\nu)(t+\delta)}{e}\right)^{(1/2)-\nu} \left(\frac{n+1-2\nu}{n}\right)^{n/2}\right].$$

Since  $\mathcal{J}(x) \cong 1$ , we also have, for  $t > 0$ , the inequality

$$(5.17) \quad \begin{aligned} S_{n,\nu}(x, t) &= \left(\frac{1}{2t}\right)^{\nu+(1/2)} e^{-x^2/(4t)} \int_0^\infty e^{-y^2/(4t)} \mathcal{J}\left(\frac{xy}{2t}\right) y^n d\mu(y) \\ &\cong \left(\frac{1}{2t}\right)^{\nu+(1/2)} e^{-x^2/(4t)} \int_0^\infty e^{-y^2/(4t)} y^n d\mu(y) \\ &= 2^n t^{n/2} \frac{\Gamma(n/2 + \nu + \frac{1}{2})}{\Gamma(\nu + \frac{1}{2})} e^{-x^2/(4t)}. \end{aligned}$$

Appealing to (5.13), we then have the following result.

**THEOREM 5.3.** For  $t > 0$ ,

$$(5.18) \quad |S_{n,\nu}(x, \pm t)| \cong 2^n t^{n/2} \frac{\Gamma(n/2 + \nu + \frac{1}{2})}{\Gamma(\nu + \frac{1}{2})} e^{\mp x^2/(4t)},$$

$$(5.19) \quad |s_{n,\nu}(x, \pm t)| \cong \left(\frac{1}{2t}\right)^{\nu+(1/2)} \frac{\Gamma(n/2 + \nu + \frac{1}{2})}{\Gamma(\nu + \frac{1}{2})} t^{-n/2}.$$

**6. Regions of convergence.** The series in terms of the basic functions of  $M_n$  converge for  $0 < |t| < \tau$ , whereas those in terms of the basic functions of  $M_{-n-1-2\nu}$  converge outside a strip, as established in the following results.

**THEOREM 6.1.** *If*

$$(6.1) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} = \frac{e}{2\tau},$$

*then the series*

$$(6.2) \quad \sum_{n=0}^{\infty} a_n S_{n,\nu}(x, t)$$

*converges absolutely for  $0 < |t| < \tau$  and uniformly on  $0 < |t| < b < \tau, |x| \leq c$ .*

*Proof.* We have that, for  $0 < \tau < \infty$ ,

$$(6.3) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} = \frac{e}{2\tau}$$

implies that, for  $0 < \theta < 1$ ,

$$(6.4) \quad a_n = O\left(\frac{e}{2\tau n \theta}\right)^{n/2}, \quad n \rightarrow \infty.$$

Using (6.4) and (5.10), we have that, for  $t > 0$ , the critical series that dominates

$$(6.5) \quad \sum_{n=0}^{\infty} |a_n S_{n,\nu}(x, \pm t)|$$

is the geometric series

$$(6.6) \quad \sum_{n=0}^{\infty} \left(\frac{b + \delta}{\tau \theta}\right)^{n/2}$$

which converges for

$$(6.7) \quad b + \delta < \tau \theta.$$

Since  $\delta$  may be taken arbitrarily close to 0, and  $\theta$  to 1, the absolute and uniform convergence of the series (6.2) follows by the Weierstrass M-test.

The proof of the analogous result for basic members of  $M_n$  of the second kind is similar and we have the following.

**THEOREM 6.2.** *If*

$$(6.8) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} = \frac{e}{2\tau},$$

*then the series*

$$(6.9) \quad \sum_{n=0}^{\infty} a_n R_{n,\nu}(x, t)$$

*converges absolutely for  $0 < |t| < \tau$  and uniformly on  $0 < a \leq |t| \leq b < \tau, |x| \leq c$ .*

For negative homogeneity, we have parallel results and will omit the proofs.

**THEOREM 6.3.** *If*

$$(6.10) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} = 2e\tau,$$

then each of the series

$$(6.11) \quad \sum_{n=0}^{\infty} a_n s_{n,\nu}(x, t)$$

and

$$(6.12) \quad \sum_{n=0}^{\infty} a_n r_{n,\nu}(x, t)$$

converges absolutely for  $|t| > \tau$ .

**7. Membership in class  $H$ .** Each of the series of basic functions belongs to class  $H$  within its region of convergence, as we establish next.

**THEOREM 7.1.** *Let*

$$(7.1) \quad u(x, t) = \sum_{n=0}^{\infty} a_n S_{n,\nu}(x, t),$$

the series converging at  $(x_0, t_0)$ ,  $0 < |t_0| < \tau$ . Then  $u(x, t)$  is a generalized temperature for  $0 < |t| < \tau$ , and

$$(7.2) \quad u(x, 0) = \sum_{n=0}^{\infty} a_n |x|^n$$

belongs to  $\{2, 1/(4\tau)\}$ .

*Proof.* Since the function  $S_{n,\nu}(x, t)$  belong to  $H$ , so will the series (7.1) if we can establish that

$$(7.3) \quad \frac{\partial}{\partial t} u(x, t) = \sum_{n=0}^{\infty} a_n \frac{\partial}{\partial t} S_{n,\nu}(x, t).$$

To justify differentiation under the summation sign, we must show that the series in (7.3) converges uniformly.

That this is so can be established on noting that, for  $t > 0$ , by (4.1),

$$(7.4) \quad \sum_{n=0}^{\infty} a_n \frac{\partial}{\partial t} S_{n,\nu}(x, t) = \sum_{n=0}^{\infty} a_n \left[ \left( \frac{x^2}{4t^2} + \frac{2n+2\nu+1}{2t} \right) S_{n,\nu}(x, t) - \frac{1}{4t^2} S_{n+2,\nu}(x, t) \right].$$

The convergence of the series (7.1) at  $(x_0, t_0)$  for any  $t_0$ ,  $0 < t_0 < \tau$ , implies that  $a_n S_{n,\nu}(x_0, t_0)$  tends to zero with  $n$ . It then follows that, by (5.18)

$$(7.5) \quad \begin{aligned} a_n &= O\left(\frac{1}{S_{n,\nu}(x_0, t_0)}\right), \quad n \rightarrow \infty, \\ &= O\left(\frac{\Gamma(\nu + \frac{1}{2})}{2^n t_0^{n/2} \Gamma((n/2) + \nu + \frac{1}{2})}\right), \quad n \rightarrow \infty, \end{aligned}$$

or, by Stirling's formula,

$$(7.6) \quad a_n = O\left(\frac{e}{2nt_0}\right)^{n/2}, \quad n \rightarrow \infty.$$

The uniform convergence of (7.4) then follows in the region  $|x| < R$ ,  $0 < t_1 < t < t_2 < t_0$  on applying the estimates (7.6) and (5.10) to the right of (7.4). Since  $R$  and  $t_0$  are arbitrary, we have the series (7.1) in  $H$  for  $0 < t < \tau$ , with an analogous computation establishing membership in  $H$  for negative time  $-\tau < t < 0$ .



Finally, from (7.1), the definition of  $S_{n,\nu}(x, t)$ , and its evenness as a function of  $x$ , we have (7.2). Moreover (7.6) implies that

$$(7.7) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} \leq \frac{2e}{4\tau}$$

which establishes that  $u(x, 0) \in \{2, 1/(4\tau)\}$ .

We can prove the following somewhat weaker theorem for  $R_{n,\nu}(x, t)$  in the same way.

THEOREM 7.2. *Let*

$$(7.8) \quad u(x, t) = \sum_{n=0}^{\infty} a_n R_{n,\nu}(x, t),$$

the series converging at  $(0, t_0)$ ,  $0 < |t_0| < \tau$ . Then  $u(x, t)$  belongs to class  $H$  for  $0 < |t| < \tau$ . Further,

$$(7.9) \quad u(x, 0) = \sum_{n=0}^{\infty} a_n x^n$$

is of growth  $\{2, 1/(4\tau)\}$ .

For series of basic functions of negative homogeneity, we have corresponding results which we state without proof.

THEOREM 7.3. *Each of the series*

$$(7.10) \quad u(x, t) = \sum_{n=0}^{\infty} a_n S_{n,\nu}(x, t), \quad |t| > \tau,$$

and

$$(7.11) \quad u(x, t) = \sum_{n=0}^{\infty} a_n r_{n,\nu}(x, t), \quad |t| > \tau,$$

belongs to class  $H$  for  $|t| > \tau$ .

**8. Representations in series of  $S_{n,\nu}(x, t)$ .** We now obtain, as a principal result, a characterization of those functions  $u(x, t)$  that can be expanded in series of the basic homogeneous generalized temperatures  $S_{n,\nu}(x, t)$ .

THEOREM 8.1. *A necessary and sufficient condition that a function  $u(x, t)$  have the series representation*

$$(8.1) \quad u(x, t) = \sum_{n=0}^{\infty} a_n S_{n,\nu}(x, t), \quad 0 < |t| < \tau,$$

is that

$$(8.2) \quad u(x, t) = \int_0^{\infty} G(x, y; t) \varphi(y) d\mu(y), \quad 0 < t < \tau,$$

where  $\varphi \in \{2, 1/(4\tau)\}$ . The coefficients  $a_n = \varphi^{(n)}(0)/n!$ .

*Proof.* Let  $u(x, t)$  have the representation (8.2) with

$$(8.3) \quad \varphi(x) = \sum_{n=0}^{\infty} a_n x^n$$

of growth  $\{2, 1/(4\tau)\}$  so that

$$(8.4) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} \leq \frac{2e}{4\tau}.$$

Then, provided that termwise integration is valid, we have

$$\begin{aligned} u(x, t) &= \int_0^\infty G(x, y; t) \left( \sum_{n=0}^\infty a_n y^n \right) d\mu(y) \\ &= \sum_{n=0}^\infty a_n \left( \int_0^\infty y^n G(x, y; t) d\mu(y) \right) \\ &= \sum_{n=0}^\infty a_n S_{n,\nu}(x, t). \end{aligned}$$

The interchange of summation and integration is justified provided that

$$(8.5) \quad \int_0^\infty G(x, y; t) \left( \sum_{n=0}^\infty |a_n| y^n \right) d\mu(y) < \infty.$$

We note, however, that  $\sum_{n=0}^\infty |a_n| y^n \in \{2, 1/(4\tau)\}$ , and hence, by (2.11), for  $t > 0$ ,

$$(8.6) \quad \sum_{n=0}^\infty |a_n| y^n = O(e^{(1/(4\tau)+\epsilon)y^2}), \quad y \rightarrow \infty.$$

It follows that the integral (8.5) converges for  $0 < t < \tau/(1 + 4\epsilon\tau)$  and so for  $0 < t < \tau$ . We thus have established the sufficiency of the condition for  $0 < t < \tau$ , and by Theorem 6.1, for  $0 < |t| < \tau$ .

Conversely, assume that (8.1) holds. Then, as in the proof of Theorem 7.1, for every  $t < \tau$ ,

$$(8.7) \quad \overline{\lim}_{n \rightarrow \infty} n |a_n|^{2/n} \leq \frac{e}{2t}$$

and so

$$(8.8) \quad \overline{\lim}_{\tau \rightarrow \infty} n |a_n|^{2/n} \leq \frac{e}{2\tau}.$$

Using  $a_n$  as coefficients, we now define a function  $\varphi$  by

$$(8.9) \quad \varphi(x) = \sum_{n=0}^\infty a_n x^n,$$

and note that as a consequence of (8.8),  $\varphi \in \{2, 1/(4\tau)\}$ . If we now evaluate

$$(8.10) \quad \int_0^\infty G(x, y; t) \varphi(y) d\mu(y),$$

we determine, by the first part of the proof, that the integral (8.10) is equal to  $u(x, t)$ . We thus have established the necessity of the condition.

The determination of the coefficients  $a_n$  as equal to  $\varphi^{(n)}(0)/n!$  is immediate.

The theorem may be illustrated by the function

$$(8.11) \quad u(x, t) = \left( \frac{\tau}{\tau - t} \right)^{\nu+(1/2)} e^{x^2/(4(\tau-t))}.$$

As established in [5, Lemma 2.4],  $u(x, t)$  has the series expansion

$$(8.12) \quad \sum_{k=0}^\infty \frac{1}{2^{2k} \tau^k k!} S_{2k,\nu}(x, t)$$

and, further, it is the explicit value of the integral

$$(8.13) \quad \int_0^\infty G(x, y; t) e^{y^2/(4\tau)} d\mu(y),$$

where  $e^{y^2/(4\tau)}$  clearly has growth  $\{2, 1/(4\tau)\}$  as predicted by the theorem. We note that the series (8.12) converges for  $0 < t < \tau$ , and indeed, by Theorem 6.1, for  $0 < |t| < \tau$ . We find, however, that

$$\sum_{k=0}^\infty \frac{1}{2^{2k} \tau^k k!} S_{2k, \nu}(0, t) = \sum_{k=0}^\infty \frac{t^k \Gamma(k - \nu - \frac{1}{2})}{\tau^k k! \Gamma(\nu + \frac{1}{2})}$$

diverges for  $t = \tau$ . It therefore follows that the strip of convergence of the series (8.1) cannot be extended. The integral (8.13) converges for  $0 < t < \tau$ , and, indeed, if  $\tau$  is negative, for all positive  $t$ .

**9. Representation in series of  $s_{n, \nu}(x, t)$ .** For a theorem analogous to that of the preceding section for series of basic homogeneous generalized temperatures of negative degree of the first kind, we take advantage of the duality provided by the Appell transform.

**THEOREM 9.1.** *A necessary and sufficient condition that a function  $u(x, t)$  have the representation*

$$(9.1) \quad u(x, t) = \sum_{n=0}^\infty a_n s_{n, \nu}(x, t), \quad |t| > \tau,$$

is that

$$(9.2) \quad u(x, t) = \int_0^\infty e^{iy^2} \mathcal{F}(xy) \varphi(y) d\mu(y), \quad -\infty < t < -\tau < 0,$$

where  $\varphi \in \{2, \tau\}$ . The coefficients  $a_n = \varphi^{(n)}(0)/n!$ .

*Proof.* We have that

$$(9.3) \quad u(x, t) = \int_0^\infty e^{iy^2} \mathcal{F}(xy) \varphi(y) d\mu(y), \quad -\infty < t < -\tau < 0,$$

with  $\varphi \in \{2, \tau\}$  if and only if

$$(9.4) \quad \begin{aligned} u^A(x, t) &= G(x; t) \int_0^\infty e^{-y^2/t} \mathcal{F}\left(\frac{xy}{t}\right) \varphi(y) d\mu(y), & 0 < t < \frac{1}{\tau}, \\ &= \int_0^\infty G(x, y; t) \Phi(y) d\mu(y), \end{aligned}$$

where

$$(9.5) \quad \Phi(y) = \frac{1}{2^{2\nu+1}} \varphi\left(\frac{y}{2}\right),$$

so that  $\Phi \in \{2, \tau/4\}$ . By Theorem 8.1, however, (9.4) holds if and only if

$$(9.6) \quad u^A(x, t) = \sum_{n=0}^\infty b_n s_{n, \nu}(x, t), \quad |t| < \frac{1}{\tau},$$

with

$$(9.7) \quad b_n = \frac{\Phi^{(n)}(0)}{n!}.$$

Substituting (3.14) in (9.6), we have

$$(9.8) \quad u^A(x, t) = \sum_{n=0}^{\infty} 2^{n+2\nu+1} b_n s_{n,\nu}^A(x, t), \quad |t| < \frac{1}{\tau},$$

or

$$(9.9) \quad u(x, t) = \sum_{n=0}^{\infty} a_n s_{n,\nu}(x, t), \quad |t| > \tau,$$

where

$$(9.10) \quad \begin{aligned} a_n &= 2^{n+2\nu+1} b_n \\ &= 2^{n+2\nu+1} \frac{\Phi^{(n)}(0)}{n!} \\ &= \frac{\varphi^{(n)}(0)}{n!} \end{aligned}$$

and the proof is complete.

The theorem may be illustrated by the function

$$(9.11) \quad u(x, t) = G(ix, -t - \tau), \quad t < -\tau < 0,$$

which has the integral representation

$$(9.12) \quad \int_0^{\infty} e^{(t+\tau)y^2} \mathcal{J}(xy) d\mu(y),$$

where  $e^{\tau y^2} \in \{2, \tau\}$ . The function (9.11) also has the series expansion

$$(9.13) \quad \sum_{n=0}^{\infty} \frac{\tau^n}{n!} s_{2n,\nu}(x, t),$$

This series diverges for  $t = \tau$  so that the region of convergence of the series (9.1) cannot be increased beyond  $|t| > \tau$ . The integral (9.12) converges for  $t < -\tau < 0$ , and for all negative  $t$  if  $\tau$  is negative.

**10. Representations in series of  $R_{n,\nu}(x, t)$ .** We characterize those functions which have series expansions in terms of the basic homogeneous generalized temperatures of positive degree of the second kind. The proof is substantially that of Theorem 8.1 and will be omitted.

**THEOREM 10.1.** *A necessary and sufficient condition that a function  $u(x, t)$  have representation*

$$(10.1) \quad u(x, t) = \sum_{n=0}^{\infty} a_n R_{n,\nu}(x, t), \quad 0 < |t| < \tau,$$

is that

$$(10.2) \quad u(x, t) = \int_0^{\infty} H(x, y; t) \varphi(y) d\mu(y), \quad 0 < t < \tau,$$

with  $\varphi \in \{2, 1/(4\tau)\}$ . The coefficients  $a_n = \varphi^{(n)}(0)/n!$ .

We illustrate the theorem by the example

$$(10.3) \quad u(x, t) = \frac{1}{\Gamma(\frac{1}{2} - \nu)} \left( \frac{\tau}{\tau - t} \right)^{\nu + (1/2)} e^{x^2/(4(\tau-t))} \Gamma\left(-\nu + \frac{1}{2}, \frac{x^2 \tau}{4t(\tau-t)}\right),$$

which is the value of the integral

$$(10.4) \quad \int_0^\infty H(x, y; t) e^{y^2/(4\tau)} d\mu(y),$$

with  $e^{y^2/(4\tau)} \in \{2, 1/(4\tau)\}$  as predicted by the theorem. Further, the function (9.3) has the series expansion

$$(10.5) \quad \sum_{n=0}^\infty \frac{1}{2^{2n} \tau^n n!} R_{2n,\nu}(x, t),$$

whose strip convergence is limited to  $0 < |t| < \tau$ . The integral (10.4) converges for  $0 < t < \tau$ , and for all positive  $t$  if  $\tau$  is negative.

**11. Representations in series of  $r_{n,\nu}(x, t)$ .** The Appell transform can be invoked again and applied to Theorem 10.1 to obtain a corresponding characterization for functions expandable in series of basic homogeneous generalized temperatures of negative degree of the second kind. We state the result without proof.

**THEOREM 11.1.** *A necessary and sufficient condition that a function  $u(x, t)$  have the representation*

$$(11.1) \quad u(x, t) = \sum_{n=0}^\infty a_n r_{n,\nu}(x, t), \quad |t| > \tau,$$

is that

$$(11.2) \quad u(x, t) = \int_0^\infty e^{ty^2} \mathcal{H}(xy) \varphi(y) d\mu(y), \quad -\infty < t < -\tau < 0,$$

with  $\varphi \in \{2, \tau\}$ . The coefficients  $a_n = \varphi^{(n)}(0)/n!$ .

We illustrate the theorem with the example

$$(11.3) \quad u(x, t) = \frac{1}{\Gamma(\frac{1}{2} - \nu)} \left[ \frac{1}{-2(t + \tau)} \right]^{\nu + (1/2)} e^{-x^2/(4(t+\tau))} \Gamma\left(-\nu + \frac{1}{2}; -\frac{x^2}{4(t + \tau)}\right),$$

which is the value of the integral

$$(11.4) \quad \int_0^\infty e^{(t+\tau)y^2} \mathcal{H}(xy) d\mu(y), \quad -\infty < t < -\tau < 0,$$

and has the series expansion

$$(11.5) \quad \sum_{n=0}^\infty \frac{\tau^n}{n!} r_{2n,\nu}(x, t), \quad |t| > \tau,$$

as the theorem requires.

REFERENCES

[1] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. TRICOMI, *Higher Transcendental Functions*, vols. 1, 2, McGraw-Hill, New York, 1953.  
 [2] ———, *Tables of Integral Transforms*, vols. 1, 2, McGraw-Hill, New York, 1954.  
 [3] D. T. HAIMO, *Functions with the Huygens property*, Bull. Amer. Math. Soc., 71 (1965), pp. 528–532.  
 [4] ———, *L<sup>2</sup> Expansions in terms of generalized heat polynomials and of their Appell transforms*, Pacific J. Math., 5 (1965), pp. 865–875.  
 [5] ———, *Expansions in terms of generalized heat polynomials and of their Appell transforms*, J. Math. Mech., 15 (1966), pp. 735–758.

- [6] ———, *Equivalence of integral transform and series expansion representations of generalized temperatures*, Proc. of the Symposium on Analytic Methods in Mathematical Physics, Gordon and Breach, New York, 1969, pp. 453–459.
- [7] ———, *Series expansions and integral representations of generalized temperatures*, Illinois J. Math., 14 (1970), pp. 621–629.
- [8] D. T. HAIMO AND F. M. CHOLEWINSKI, *Integral representations of solutions of the generalized heat equation*, Ibid., 10 (1966), pp. 623–638.
- [9] D. V. WIDDER, *Series expansions in terms of the temperature functions of Poritsky and Powell*, Quart. Appl. Math., 20 (1963), pp. 41–47.
- [10] ———, *Expansions in terms of the homogeneous solutions of the heat equation*, Proceedings of the Edwardsville Conference, Southern Illinois University Press, 1968, pp. 171–196.
- [11] ———, *Expansions in series of homogeneous temperature functions of the first and second kinds*, Duke Math. J., 36 (1969), pp. 495–509.
- [12] ———, *Homogeneous solutions of the heat equation*, Proc. of the Symposium on Analytic Methods in Mathematical Physics, Gordon and Breach, New York, 1969, pp. 379–398.
- [13] D. V. WIDDER AND P. C. ROSENBLUM, *Expansions in terms of heat polynomials and associated functions*, Trans. Amer. Math. Soc., 92 (1959), pp. 220–266.

## HAMMING ASSOCIATION SCHEMES AND CODES ON SPHERES\*

STUART P. LLOYD†

**Abstract.** A Hamming associator is defined for the unit sphere  $S^n \subset E^{n+1}$ . The spherical harmonics are the eigenfunctions of the associator; and the ultraspherical polynomials, the eigenvalues, corresponding to the Krawtchouk polynomials in the binary case. A linear programming bound on the size of a minimum distance code on  $S^n$  generalizes the bound obtained by N. J. A. Sloane in the discrete case.

**1. Introduction.** A finite metric space is an association scheme if, given any  $a, b, c$ , for each pair of points  $x, y$  such that  $d(x, y) = a$ , the number of points  $z$  such that  $d(x, z) = b$  and  $d(z, y) = c$  is independent of  $x, y$ . (Typically the space is at least two-point homogeneous under its isometry group.) N. J. A. Sloane in [1] discusses algebraic systems induced by association schemes, and describes a linear programming bound on the size of a minimum distance code in a metric association scheme.

The central unit sphere  $S^n$  in  $(n + 1)$ -dimensional Euclidean space  $E^{n+1}$  has the orthogonal group  $O(n + 1)$  as isometry group, and is  $n + 1$ -point homogeneous. We shall give a simple description of an associator which corresponds to the Hamming associator in the binary case. We show that the spherical harmonics of  $S^n$  are the eigenfunctions of the associator; and the ultraspherical polynomials, the eigenvalues, corresponding to the Krawtchouk polynomials of the binary case.

The corresponding coding problem for  $S^n$  is to place the largest possible number of disjoint open spherical caps on  $S^n$ , each of specified angular radius  $\alpha$ . The special value  $\alpha = \pi/6$  corresponds to finding the largest number of nonoverlapping unit spheres in  $E^{n+1}$  which touch  $S^n$ . A linear programming bound on the size of such a code is set up, analogous to the bound of [1] in the discrete case.

**2. Polar coordinates in  $E^{n+1}$ .** We will actually need very little about the explicit form of the spherical harmonics on  $S^n$ . The formulas of [4, Chap. XI] are a challenge, however, and we digress in § 5 to present a tidy form for the spherical harmonics.

As a preliminary to this, we give a slightly modified version of polar coordinates on  $S^n$ , so as to obtain a uniform notation in all dimensions. We start with  $n = 0$ . The real line  $E^1 = \{x_0: -\infty < x_0 < \infty\}$  has unit sphere  $S^0 = \{x_0 \in E^1: x_0^2 = 1\}$ , i.e.,  $S^0 = \{-1, +1\}$  consists of two points. We introduce a two-valued polar coordinate  $\theta_0$  on  $S^0$  by

$$\cos \theta_0 = +1 \in S^0 \quad \text{for } \theta_0 = 0,$$

$$\cos \theta_0 = -1 \in S^0 \quad \text{for } \theta_0 = \pi.$$

The radial coordinate being  $r = +\sqrt{x_0^2}$ , the polar coordinate representation in  $E^1$  is  $x_0 = r \cos \theta_0$ ; the polar angle is discontinuous and undefined at the origin. In terms of polar coordinates the Euclidean volume element is  $dx_0 = dr[d\theta_0]$  where  $dr$  is Lebesgue measure on  $\{r > 0\}$  and where  $[d\theta_0]$  is two unit point measures, one each at  $\theta_0 \in \{0, \pi\}$ . The 0-dimensional surface area of  $S^0$  is thus

$$\sigma_0 = \int [d\theta_0] = 2.$$

Consider next  $n = 1$ . The rectangular coordinates in  $E^2$  being  $(x_0, x_1)$ , the unit sphere centered at the origin is  $S^1 = \{(x_0, x_1) \in E^2: x_0^2 + x_1^2 = 1\}$ , and the radial coord-

\* Received by the editors January 22, 1979, and in revised form August 20, 1979.

† Bell Laboratories, Whippany, New Jersey 07981.

dinate is  $r = +\sqrt{x_0^2 + x_1^2}$ . Instead of the conventional  $x_1 = r \cos \phi$ ,  $x_0 = r \sin \phi$ ,  $0 \leq \phi < 2\pi$ , we take as our polar coordinate representation

$$x_1 = r \cos \theta_1, \quad 0 \leq \theta_1 \leq \pi,$$

$$x_0 = r \sin \theta_1 \cos \theta_0, \quad \theta_0 \in \{0, \pi\}.$$

In the half-space  $\{x_0 > 0\}$  the angle  $\theta_0$  has value 0, and in  $\{x_0 < 0\}$  it has value  $\pi$ ; on  $\{x_0 = 0\}$  it is undefined. If  $0 \leq \theta_1 \leq \pi$  is thought of as longitude from the meridian then  $\theta_0$  is the analog of the familiar designation E, W for longitudes. In conventional polar coordinates the angle  $\phi$  winds clear around to  $2\pi$ , like right ascension or hour angle, and is related to our variables by  $\phi = |\theta_1 - 2\theta_0|$ . Alternatively, if  $\phi$  has range  $-\pi < \phi \leq \pi$  then  $\theta_1 = |\phi|$  and  $\cos \theta_0 = \text{sgn}(\phi)$ .

Let us now consider the general case. The unit sphere  $S^n$  in  $E^{n+1}$  is  $S^n = \{(x_0, \dots, x_n) : x_0^2 + \dots + x_n^2 = 1\}$ , and the radial coordinate is  $r = +\sqrt{x_0^2 + \dots + x_n^2}$ . Polar coordinates for a point in general position are determined by

$$x_n = r \cos \theta_n,$$

$$x_{n-1} = r \sin \theta_n \cos \theta_{n-1},$$

$$\vdots$$

$$x_1 = r \sin \theta_n \sin \theta_{n-1} \cdots \sin \theta_2 \cos \theta_1,$$

$$x_0 = r \sin \theta_n \sin \theta_{n-1} \cdots \sin \theta_2 \sin \theta_1 \cos \theta_0,$$

where  $0 \leq \theta_n, \dots, \theta_1 \leq \pi, \quad \theta_0 \in \{0, \pi\}$ .

Notice that this definition places  $\theta_1$  on an equal footing with  $\theta_2, \dots, \theta_n$ , which are the conventional colatitudes. In each  $E^{n+1}$ ,  $\cos \theta_0 = \text{sgn}(x_0)$ ,  $x_0 \neq 0$ , is constant on the half-spaces  $\{x_0 > 0\}$ ,  $\{x_0 < 0\}$ , and is discontinuous and undefined on  $\{x_0 = 0\}$ .

The Euclidean volume element  $dx_0 \cdots dx_n$  in  $E^{n+1}$  will be written  $r^n dr d\omega$  in polar coordinates, the  $n$ -dimensional element of area of  $S^n$  being

$$d\omega = [d\theta_0] \prod_1^n (\sin^{j-1} \theta_j d\theta_j),$$

for  $\theta_0, \dots, \theta_n$  defined above. The area of  $S^n$  is

$$\sigma_n = \int_{S^n} d\omega = \frac{2 \cdot \pi^{(n+1)/2}}{((n-1)/2)!}, \quad n \geq 0,$$

where  $z!$  always denotes  $\Gamma(z+1)$ . It is also convenient to define numbers  $\tau_1, \dots, \tau_n$  by  $\sigma_n = \sigma_0 \tau_1 \cdots \tau_n$ , i.e.,

$$\tau_j = \int_0^\pi \sin^{j-1} \theta d\theta = B\left(\frac{1}{2}, \frac{j}{2}\right)$$

$$= \frac{(-1/2)!((j-2)/2)!}{((j-1)/2)!}$$

$$= \frac{2^{j-1} [((j-2)/2)!]^2}{(j-1)!}$$

$$= \frac{(j-1)! \pi}{2^{j-1} [((j-1)/2)!]^2}, \quad j \geq 1;$$

$\tau_1 = \pi$ ,  $\tau_3 = \pi/2$ ,  $\dots$  and  $\tau_2 = 2$ ,  $\tau_4 = \frac{4}{3}$ ,  $\dots$ . We point out that  $\tau_1$  is not exceptional; the



2 in  $\sigma_1 = \sigma_0 \tau_1 = \sigma_0 \int_0^\pi d\theta_1 = 2\pi$  is the surface area of  $S^0$ . Again, the unit circle  $S^1$  is treated as the union of two semicircles, each coordinatized by  $0 \leq \theta_1 \leq \pi$ .

**3. The associator for  $S^n$ .** In the discrete case, the application of an adjacency matrix  $D_k$  has the effect of summing on those points at distance  $k$  from a given point [1, § 3]. For the continuous analog it is more convenient to average, as follows. Let  $C(S^n)$  denote the Banach space of continuous real functions on  $S^n$ , normed by  $\|f\| = \max_\omega |f(\omega)|$ . Let  $\omega \in S^n$  be fixed, and let the surface element  $d\omega'$  of  $S^n$  have polar coordinates referred to  $\omega$  as pole, with  $\theta$  denoting the polar angle. That is,

$$\begin{aligned} \omega \cdot \omega' &= \cos \theta \\ d\omega' &= \sin^{n-1} \theta \, d\theta \, d'\omega'; \end{aligned}$$

the small circle  $\{\omega': \omega \cdot \omega' = \cos \theta\}$  consisting of points  $\omega'$  at polar angle  $\theta$  from  $\omega$  is congruent to  $(\sin \theta) S^{n-1}$ , and  $d'\omega'$  is the element of area on the copy of  $S^{n-1}$ :

$$\begin{aligned} \int d'\omega' &= \sigma_{n-1} \quad \text{for any } 0 < \theta < \pi, \\ \int_0^\pi \sin^{n-1} \theta \, d\theta &= \tau_n, \\ \int_{S^n} d\omega' &= \tau_n \sigma_{n-1} = \sigma_n. \end{aligned}$$

For each  $0 < \theta < \pi$  we define the Hamming associator  $A_\theta: C(S^n) \rightarrow C(S^n)$  by

$$(A_\theta f)(\omega) = \int f(\omega') \, d'\omega' / \sigma_{n-1}, \quad \omega \in S^n, \quad f \in C(S^n), \quad 0 < \theta < \pi$$

with  $d'\omega'$  as above; that is,  $(A_\theta f)(\omega)$  is an average of  $f$  over the points on  $S^n$  at constant polar angle  $\theta$  from  $\omega$ . For  $\theta = 0$  we set  $(A_\theta f)(\omega) = f(\omega)$  and for  $\theta = \pi$  we set  $(A_\theta f)(\omega) = f(-\omega)$ ,  $\omega \in S^n$ ,  $f \in C(S^n)$ ; this makes  $A_\theta$  continuous in  $0 \leq \theta \leq \pi$  in the strong operator topology. Obviously  $A_\theta \geq 0$  and  $A_\theta 1 = 1$ , so the operator norm of  $A_\theta$  is  $\|A_\theta\| = 1$ ,  $0 \leq \theta \leq \pi$ . The property

$$\int_0^\pi [(A_\theta f)(\omega)] \sin^{n-1} \theta \, d\theta / \tau_n = \int_{S^n} f(\omega') \, d\omega' / \sigma_n \quad (\text{constant in } \omega), \quad f \in C(S^n)$$

is straightforward from the definition.

Let  $O(n+1)$  denote the group of  $(n+1) \times (n+1)$  real orthogonal matrices. Treating  $\omega \in S^n \subset E^{n+1}$  as a column vector, we have a group action  $\omega' = \rho\omega$ ,  $\omega \in S^n$ ,  $\rho \in O(n+1)$ , which is uniformly jointly continuous. There is then an induced action  $T_\rho: C(S^n) \rightarrow C(S^n)$ ,  $\rho \in O(n+1)$ , defined by

$$(T_\rho f)(\omega) = f(\rho\omega), \quad \omega \in S^n, \quad \rho \in O(n+1), \quad f \in C(S^n).$$

This is an antirepresentation:  $T_{\rho_2} T_{\rho_1} = T_{\rho_1 \rho_2}$ , with the properties  $T_\rho \geq 0$ ,  $T_\rho 1 = 1$ ,  $\|T_\rho\| = 1$ . (The set of adjoints  $T_\rho^*$  is a homomorphism which extends  $\rho\omega$  to the measures on  $S^n$ .)

We will use without formal proof the property  $T_\rho A_\theta = A_\theta T_\rho$ ,  $\rho \in O(n+1)$ ,  $0 \leq \theta \leq \pi$ . Geometrically the property is obvious. In  $(T_\rho A_\theta f)(\omega) = (A_\theta f)(\rho\omega)$  the averages over small circles of angular radius  $\theta$  are found first, then the set of average values is slid over  $S^n$  by  $\rho^{-1}$ . In  $(A_\theta T_\rho f)(\omega)$  the function is slid before the averaging is done; the rotational invariance of the family of averaging measures will be taken for granted.

Let  $d\rho$  denote Haar measure on  $O(n+1)$ , with normalization  $\int d\rho = 1$ . If  $\omega$  is fixed then  $d\rho$  sweeps  $\omega' = \rho\omega$  uniformly over  $S^n$ , in the sense

$$\int_{O(n+1)} f(\rho\omega) d\rho = \int_{S^n} f(\omega') d\omega'/\sigma_n \quad (\text{constant in } \omega), \quad f \in C(S^n);$$

this is well known [2], [3].

For  $1 \leq p \leq \infty$  we denote by  $L_p(S^n)$  the usual Lebesgue real function space  $L_p(S^n, d\omega/\sigma_n)$ ; for  $1 \leq p < \infty$  the norm is

$$\|f\|_p = \left[ \int_{S^n} |f(\omega)|^p d\omega/\sigma_n \right]^{1/p}, \quad f \in L_p(S^n).$$

All functions being real valued, the pairing (scalar product) is

$$(f, g) = \int_{S^n} f(\omega)g(\omega) d\omega/\sigma_n$$

symmetric in  $f, g$ . Operators  $A_\theta$  have extensions  $A_\theta : L_p(S^n) \rightarrow L_p(S^n)$  satisfying  $\|A_\theta\|_p = 1, 1 \leq p \leq \infty$ . This is clear for  $p = \infty$ ; for  $1 \leq p < \infty$  assume first that  $f \in C(S^n)$ . Then, for any  $\omega \in S^n$ ,

$$\begin{aligned} \int_{S^n} |(A_\theta f)(\omega')|^p d\omega'/\sigma_n &= \int_{O(n+1)} |(A_\theta f)(\rho\omega)|^p d\rho \\ &= \int_{O(n+1)} |(T_\rho A_\theta f)(\omega)|^p d\rho \\ &= \int_{O(n+1)} |(A_\theta T_\rho f)(\omega)|^p d\rho \\ &\leq \int_{O(n+1)} [A_\theta(T_\rho |f|^p)](\omega) d\rho \quad (\text{by Jensen's inequality}) \\ &= A_\theta \left[ \int_{S^n} |f(\omega')|^p d\omega'/\sigma_n \right](\omega) \\ &= [\|f\|_p]^p \quad (\text{constant in } \omega) \end{aligned}$$

after some interchanges of the order of integration. The result stated now follows from the usual approximation arguments. The property

$$\int_0^\pi (A_\theta f, g) \sin^{n-1} \theta d\theta/\tau_n = (f, 1)(1, g) = \left[ \int_{S^n} f(\omega) d\omega/\sigma_n \right] \left[ \int_{S^n} g(\omega') d\omega'/\sigma_n \right]$$

derives from a property of  $A_\theta$  given previously.

**4. Polar polynomials.** Polar (or ultraspherical or spherical) polynomials  $P_l^{(n)}(x)$  for  $S^n$  are defined by generating functions

$$n \geq 2: \frac{1}{[1+t^2-2tx]^{(n-1)/2}} = \sum_{l=0}^\infty \frac{(n-2+l)!}{(n-2)!l!} t^l P_l^{(n)}(x), \quad -1 \leq x \leq 1, \quad |t| < 1;$$

$$n = 1: P_0^{(1)}(x) = 1, \quad \log \frac{1}{[1+t^2-2tx]^{1/2}} = \sum_{l=1}^\infty \frac{1}{l} t^l P_l^{(1)}(x), \quad -1 \leq x \leq 1, \quad |t| < 1;$$

$$n = 0: [1+t^2-2tx]^{1/2} = 1-tx = \sum_{l=0}^1 (-1)^l t^l P_l^{(0)}(x), \quad x \in \{-1, 1\}, \quad |t| < 1.$$

In terms of the Gegenbauer functions  $C_\alpha^\nu(z)$  [4, Chap. 3] these are

$$P_l^{(n)}(x) = \frac{(n-2)!l!}{(n-2+l)!} C_l^{(n-1)/2}(x)$$

$$= {}_2F_1(-l, n+l-1; n/2; (1-x)/2), \quad n \geq 1$$

and this serves to define  $P_l^{(n)}(x)$  for complex  $n, l, x$  except for  $n = 0, -2, -4, \dots$ . We will always have nonnegative integer values for  $n, l$ , and if  $n \geq 1$  then  $P_l^{(n)}(x)$  is a polynomial in  $x$  of exact degree  $l$ , even or odd in  $x$  for even or odd  $l$ , respectively. Associated with  $P_l^{(n)}(x)$  is the important normalization constant

$$d_l^{(n)} = \frac{(n+l)!}{n!l!} - \frac{(n+l-2)!}{n!(l-2)!}, \quad n \geq 0, \quad l = 0, 1, \dots;$$

the second term vanishes for  $l = 0, 1$ . The  $d_l^{(n)}$  are integers, positive when  $n \geq 1$ , and will be discussed later.

Certain values are, for  $n = 1, 2, \dots$  and  $-1 \leq x \leq 1$ ,

$$P_l^{(n)}(1) = 1 \geq |P_l^{(n)}(x)|, \quad l \geq 0,$$

$$P_0^{(n)}(x) = 1, \quad d_0^{(n)} = 1,$$

$$P_1^{(n)}(x) = x, \quad d_1^{(n)} = n + 1,$$

$$P_2^{(n)}(x) = \frac{(n+1)x^2 - 1}{n}, \quad d_2^{(n)} = \frac{n(n+3)}{2},$$

$$d_l^{(n)} = \frac{(n-2+l)!}{(n-2)!l!} \left[ 1 + \frac{2l}{n-1} \right], \quad n \geq 2,$$

and for  $l = 0, 1, \dots$ ,

$$n = 0: \quad d_0^{(0)} = d_1^{(0)} = 1, \quad d_2^{(0)} = d_3^{(0)} = \dots = 0 \quad \text{and}$$

$$P_0^{(0)}(x) = 1, \quad P_1^{(0)}(x) = x, \quad P_l^{(0)}(x) \text{ undefined for } l > 1 \quad x \in \{-1, 1\},$$

$$n = 1: \quad d_0^{(1)} = 1, \quad d_1^{(1)} = d_2^{(1)} = \dots = 2 \quad \text{and}$$

$$P_l^{(1)}(\cos \theta) = \cos l\theta \text{ is the Chebycheff polynomial (1st kind),}$$

$$n = 2: \quad d_l^{(2)} = 2l + 1 \quad \text{and} \quad P_l^{(2)}(x) \text{ is the Legendre polynomial,}$$

$$n = 3: \quad d_l^{(3)} = (l+1)^2 \quad \text{and} \quad P_l^{(3)}(\cos \theta)$$

$$= \frac{\sin(l+1)\theta}{(l+1)\sin \theta} \text{ is the Chebycheff polynomial (2nd kind).}$$

The well-known recurrence relations and differential equations for  $C_\alpha^\nu(z)$  [4, Chap. 3] [5, Chap. 22] transcribe to properties of  $P_l^{(n)}(x)$ . We list, for  $n = 1, 2, \dots$  and  $l = 0, 1, \dots$ ,

$$xP_l^{(n)}(x) = \frac{n-1+l}{n-1+2l} P_{l+1}^{(n)}(x) + \frac{l}{n-1+2l} P_{l-1}^{(n)}(x),$$

$$\sum_{l=0}^{L-1} d_l^{(n)} P_l^{(n)}(x) P_l^{(n)}(y) = \frac{(n-2+L)!}{(n-1)!(L-1)!} \frac{P_L^{(n)}(x) P_{L-1}^{(n)}(y) - P_{L-1}^{(n)}(x) P_L^{(n)}(y)}{x-y},$$

$$P_l^{(n)}(\cos \theta \cos \gamma + \sin \theta \sin \gamma \cos \phi) = \sum_{m=0}^l d_{l-m}^{(n+2m)} [\sin^m \theta P_{l-m}^{(n+2m)}(\cos \theta)] \cdot [\sin^m \gamma P_{l-m}^{(n+2m)}(\cos \gamma)] [d_m^{(n-1)} P_m^{(n-1)}(\cos \phi)] / d_l^{(n)};$$

when  $n = 1$  there is a restriction  $\cos \phi = \pm 1$  in the last. The Poisson kernel for  $S^n \subset E^{n+1}$  derives from

$$\frac{1-t^2}{[1-2xt+t^2]^{(n+1)/2}} = \sum_{l=0}^{\infty} t^l d_l^{(n)} P_l^{(n)}(x), \quad -1 \leq x \leq 1, \quad |t| < 1.$$

The  $P_l^{(n)}(x)$  have the orthogonality property

$$\int_0^\pi P_l^{(n)}(\cos \theta) P_{l'}^{(n)}(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n = \delta_{ll'} / d_l^{(n)}, \quad n \geq 1, \quad 0 \leq l, l' < \infty,$$

so that  $\{\sqrt{d_l^{(n)}} P_l^{(n)}(\cos \theta)\}$  are real orthonormal in  $L_2^{(n)} = L_2([0, \pi], \sin^{n-1} \theta d\theta / \tau_n)$ . They are also complete, and each  $\nu \in L_2^{(n)}$  has orthogonal expansion

$$\begin{aligned} \nu(\cos \theta) &= \sum_{l=0}^{\infty} \nu_l P_l^{(n)}(\cos \theta) \quad \text{in mean in } L_2^{(n)}, \\ \nu_l &= d_l^{(n)} \int_0^\pi \nu(\cos \theta) P_l^{(n)}(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n, \quad l \geq 0. \end{aligned}$$

The completeness relation takes the form

$$\begin{aligned} [\|\nu\|_2]^2 &= \int_0^\pi \nu^2(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n \\ &= \int_{-1}^1 \nu^2(x) (1-x^2)^{(n/2)-1} dx / \tau_n \\ &= \sum_{l=0}^{\infty} \nu_l^2 / d_l^{(n)}, \quad \nu \in L_2^{(n)}. \end{aligned}$$

As indicated,  $L_p^{(n)}$  will mean either of the equivalent spaces  $L_p([0, \pi], \sin^{n-1} \theta d\theta / \tau_n)$ ,  $L_p([-1, 1], (1-x^2)^{(n/2)-1} dx / \tau_n)$ , clear from context.

**5. Spherical harmonics.** Examination of the identity  $(1-t^2)(1-t)^{-(n+1)} = (1+t)(1-t)^{-n}$  shows that for given integers  $n \geq 0, l \geq 0$ , the integer  $d_l^{(n)}$  is the number of solutions of

$$\begin{aligned} l &= l_0 + l_1 + \dots + l_n, \\ l_0 &\text{ restricted to } l_0 = 0 \quad \text{or} \quad l_0 = 1, \\ l_1, \dots, l_n &\geq 0 \quad \text{integers.} \end{aligned}$$

If  $l = (l_0, l_1, \dots, l_n)$  denotes such an ordered restricted partition we call  $|l| = l = l_0 + \dots + l_n$  the degree, and we define

$$\begin{aligned} m_0 &= 0, \\ m_1 &= l_0, \\ m_2 &= l_0 + l_1, \\ &\vdots \\ m_n &= l_0 + l_1 + \dots + l_{n-1}. \end{aligned}$$

The real orthonormal spherical harmonics  $Y_l^{(n)}(\omega)$  of degree  $l = |l|$  on  $S^n$ , relative to probability measure  $d\omega/\sigma_n$ , are taken to be (see [4, Chap. XI])

$$Y_l^{(n)}(\omega) = \prod_{i=0}^n \left\{ \left[ \frac{\tau_i d_i^{(i+2m_i)}}{\tau_{i+2m_i}} \right]^{1/2} \sin^{m_i} \theta_i P_{l_i}^{(i+2m_i)}(\cos \theta_i) \right\}, \quad \omega \in S^n$$

with the understanding that  $\sin^{m_0} \theta_0 = 0^0 = 1$  and  $\tau_0/\tau_0 = 1$  always. We denote by  $l$  the special partition  $(0, \dots, 0, l)$ ; all  $m$ 's vanish, and

$$Y_l^{(n)}(\omega) = \sqrt{d_l^{(n)}} P_l^{(n)}(\cos \theta_n), \quad \omega \in S^n$$

is the unique  $Y_l^{(n)}(\omega)$  of degree  $l$  which does not vanish at the standard pole  $\omega_0 = (0, \dots, 0, 1)$ . For each  $0 \leq k \leq n$  the factor  $\prod_{i=0}^k \{\cdot\}$  in  $Y_l^{(n)}(\omega)$  above is  $Y_\lambda^{(k)}(\omega_k)$  for  $\lambda = (l_0, l_1, \dots, l_k)$  of degree  $m_{k+1}$ ; and  $\omega_k \in S^k$ , the image of  $\omega$  under the mapping  $(\theta_0, \theta_1, \dots, \theta_n) \rightarrow (\theta_0, \theta_1, \dots, \theta_k)$  of  $S^n$  onto  $S^k$ .

We give the  $Y_l^{(n)}$  explicitly for small values of  $n$  or  $l$ , to make the connection with the usual treatment.

$n = 0$ :  $d_l^{(0)} = 0$  for  $l > 1$ , and only possible degrees are

$$l = l_0 = 0 \quad \text{for which } Y_0^{(0)}(\omega) = 1, \quad \omega \in S^0,$$

$$l = l_0 = 1 \quad \text{for which } Y_1^{(0)}(\omega) = \cos \theta_0 = x_0/r, \quad \omega \in S^0.$$

These are just the angular factors of the two homogeneous harmonic polynomials  $1, x_0$  on  $E^1$ .

$n = 1$ : The partitions of degree  $l$  are  $l = (l_0, l - l_0)$ :

$$l = 0 \quad \text{for which } Y_0^{(1)}(\omega) = 1,$$

$$l = (0, l) \quad \text{for } l > 0, \text{ for which}$$

$$Y_l^{(1)}(\omega) = \sqrt{\frac{\tau_1 d_l^{(1)}}{\tau_1}} P_l^{(1)}(\cos \theta_1) = \sqrt{2} \cos l\theta_1,$$

$$l = (1, l - 1) \quad \text{for } l > 0 \text{ for which}$$

$$Y_{(1, l-1)}^{(1)}(\omega) = \cos \theta_0 \sqrt{\frac{\tau_1 d_{l-1}^{(3)}}{\tau_3}} \sin \theta_1 P_{l-1}^{(3)}(\cos \theta_1) \\ = \sqrt{2} \sin l\theta_1 \cos \theta_0.$$

The usual treatment has  $e^{\pm im\phi}$  dependence on the last (hour) angle. In the present version the factor  $\cos \theta_0$  makes  $\sin l\theta_1 \cos \theta_0$  an odd function around the meridian, viz.,  $\sin l\phi$ .

$n = 2$ : For  $l = (l_0, l_1, l_2)$ , multiply  $Y_{(l_0, l_1)}^{(1)}(\theta_0, \theta_1)$  above by

$$\sqrt{\frac{\tau_2 d_{l_2}^{(2+2m_2)}}{\tau_{2+2m_2}}} \sin^{m_2} \theta_2 P_{l_2}^{(2+2m_2)}(\cos \theta_2) \\ = \sqrt{\frac{(l - m_2)!}{(l + m_2)!}} (2l + 1) P_l^{m_2}(\cos \theta_2) \quad (\text{associated Legendre function.})$$

$$l = 0: Y_0^{(n)}(\omega) = 1, \quad \omega \in S^n, \quad n \geq 0.$$

$l = 1$ : If  $(i)$  denotes the partition with  $l_i = 1 = l$ ,

$$Y_{(i)}^{(n)}(\omega) = \sqrt{n + 1} x_i / r, \quad \omega \in S^n, \quad 0 \leq i \leq n, \quad n \geq 0.$$

$l = 2$ : If  $(ij)$  denotes the partition with  $l_i = l_j = 1, l = 2, 0 \leq i < j \leq n$ ,

$$Y_{(ij)}^{(n)}(\omega) = \sqrt{(n + 1)(n + 3)} x_i x_j / r^2, \quad \omega \in S^n;$$

if (ii) denotes the partition with  $l_i = 2 = l, 1 \leq i \leq n$ ,

$$Y_{(ii)}^{(n)}(\omega) = \sqrt{\frac{(n+1)(n+3)i}{2(i+1)}} \left[ x_i^2 - (1/i) \sum_{j=0}^{i-1} x_j^2 \right] / r^2, \quad \omega \in S^n.$$

The restriction  $l_0 = 0$  or  $1$  excludes  $i = 0$  in the last; that there are  $n$  and not  $n + 1$  of them is seen from  $x_0^2 + \dots + x_n^2 = r^2$ .

Since  $A_\theta$  commutes with every  $T_\rho, \rho \in O(n+1)$ , the  $Y_l^{(n)}(\omega)$  for various  $l$  are a complete set of eigenfunctions of  $A_\theta$ ; we have invoked Schur's lemma and the Peter-Weyl theorem. Moreover, the eigenvalue of  $A_\theta$  for  $Y_l^{(n)}(\omega)$  depends only on  $l = |l|$ ; it is thus

$$(A_\theta Y_l^{(n)})(\omega) = P_l^{(n)}(\cos \theta) Y_l^{(n)}(\omega), \quad \omega \in S^n, \quad l = |l|,$$

easily obtained by evaluating  $(A_\theta Y_l^{(n)})(\omega_0)$ .

The subspace of  $L_2(S^n)$  spanned by the  $Y_l^{(n)}(\omega)$  for fixed degree  $l$  is irreducibly invariant under the  $T_\rho, \rho \in O(n+1)$ , and the projection kernel is the well-known

$$\sum_{|l|=l} Y_l^{(n)}(\omega) Y_l^{(n)}(\omega') = d_l^{(n)} P_l^{(n)}(\omega \cdot \omega'), \quad \omega, \omega' \in S^n.$$

The projection  $J_l : L_2(S^n) \rightarrow L_2(S^n)$  itself is

$$\begin{aligned} (J_l f) &= \sum_{|l|=l} Y_l^{(n)}(\omega) \int_{S^n} f(\omega') Y_l^{(n)}(\omega') d\omega' / \sigma_n \\ &= d_l^{(n)} \int_{S^n} f(\omega') P_l^{(n)}(\omega \cdot \omega') d\omega' / \sigma_n. \end{aligned}$$

In this last, let us refer the integration variable  $\omega'$  to  $\omega$  as pole:

$$\begin{aligned} \omega \cdot \omega' &= \cos \theta \\ d\omega' &= \sin^{n-1} \theta d\theta \cdot d'\omega' \end{aligned}$$

with  $d'\omega'$  the normalized surface area on the small circle at polar angle  $\theta$  from  $\omega$ , i.e.,  $\int d'\omega' = \sigma_{n-1}$ . This being the machinery for  $A_\theta$ , we see that

$$(J_l f)(\omega) = d_l^{(n)} \int_0^\pi [(A_\theta f)(\omega)] P_l^{(n)}(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n,$$

or as an operator valued integral,

$$(1) \quad J_l = d_l^{(n)} \int_0^\pi A_\theta P_l^{(n)}(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n;$$

this is the analog of [1, (11)] in the discrete case. The eigenvalues of  $A_\theta$  being as above, the analog of [1, (7)] is

$$(2) \quad A_\theta = \sum_{l=0}^\infty P_l^{(n)}(\cos \theta) J_l.$$

It follows from this that  $A_\theta$  is selfadjoint on  $L_2(S^n)$ , i.e.,  $(A_\theta f, g) = (f, A_\theta g), f, g \in L_2(S^n)$ , since each orthogonal projection  $J_l, l \geq 0$ , is selfadjoint. (This could have been proved directly, of course.)

If we substitute (2) in the formula (1) for  $J_l$ ,

$$J_l = d_l^{(n)} \int_0^\pi \left\{ \sum_{l'=0}^\infty P_{l'}^{(n)}(\cos \theta) J_{l'} \right\} P_l^{(n)}(\cos \theta) \sin^{n-1} \theta d\theta / \tau_n,$$

the orthonormality of  $\{\sqrt{d_l^{(n)}}P_l^{(n)}(\cos \theta)\}_{l=0}^\infty$  in  $L_2^{(n)}$  is apparent; completeness is via the Peter–Weyl theorem. The discrete analog is [1, (19)]; indeed, the Krawtchouk polynomials are the spherical functions of the  $n$  cube [6].

For completeness we mention the Bose–Mesner algebra, although we do not use it. Namely, the analog of [1, (4)] is

$$A_{\theta_2}A_{\theta_1} = \int_0^\pi A_n \sin^{n-2} \phi \, d\phi / \tau_{n-1},$$

$$\cos \eta = \cos \theta_1 \cos \theta_2 + \sin \theta_1 \sin \theta_2 \cos \phi;$$

this can be shown geometrically but it is immediate from (2) and the addition formula for the  $P_l^{(n)}$ .

**6. Polar functions, cap functions.** In this section we are concerned with evaluating the inner product

$$(A_\theta G, H) = \int_{S^n} H(\omega)[(A_\theta G)(\omega)] \, d\omega / \sigma_n$$

for functions  $G, H$  of a certain kind.

A polar function on  $S^n$ , with  $v \in S^n$  as pole, is a function  $W(\omega)$ ,  $\omega \in S^n$ , which depends only on the angle between  $\omega$  and  $v$ . That is, there is a function  $w(x)$ ,  $-1 \leq x \leq 1$ , such that  $W(\omega) = w(v \cdot \omega)$ ,  $\omega \in S^n$ . Using polar coordinates around  $v$ , it is easy to see that  $\|W\|_p = \|w\|_p$  in  $L_p(S^n)$ ,  $L_p^{(n)}$  respectively.

Consider  $(A_\theta W)(\omega)$ ,  $\omega \in S^n$ , where  $W(\omega) = w(v \cdot \omega)$ ,  $\omega \in S^n$ , is a polar function around  $v$ . It should be clear geometrically that  $A_\theta W$  is also a polar function with  $v$  as pole; we sketch the proof. If  $\rho \in O(n+1)$  leaves  $v$  fixed,  $\rho v = v$ , then  $T_\rho W = W$ , from  $w(v \cdot (\rho\omega)) = w((\rho^{-1}v) \cdot \omega) = w(v \cdot \omega)$ . There follows  $(A_\theta W)(\rho\omega) = (T_\rho A_\theta W)(\omega) = (A_\theta T_\rho W)(\omega) = (A_\theta W)(\omega)$ ,  $\omega \in S^n$ , for every such  $\rho$ . Since the stabilizer  $\{\rho \in O(n+1): \rho v = v\}$  acts as  $O(n)$  on each of the small circles around  $v$ ,  $(A_\theta W)(\omega)$  is constant on the small circles, i.e., a polar function with  $v$  as pole. Thus there exists a function  $(Aw)(x, y)$ ,  $-1 \leq x, y \leq 1$ , independent of  $v$  and  $\omega$ , such that

$$(A_\theta W)(\omega) = (Aw)(\cos \theta, v \cdot \omega), \quad \omega \in S^n \quad \text{when}$$

$$W(\omega) = w(v \cdot \omega), \quad \omega \in S^n.$$

To find  $(Aw)(x, y)$  explicitly we are free to choose, say,  $\omega = \omega_0$ ,  $v = (0, \dots, 0, \sin \gamma, \cos \gamma)$ , where  $\cos \gamma = v \cdot \omega$ . The polar coordinates of integration variable  $\omega'$  being  $\theta'_0, \dots, \theta'_{n-1}, \theta$ , so that  $v \cdot \omega' = \cos \theta \cos \gamma + \sin \theta \sin \gamma \cos \theta'_{n-1}$ , we obtain

$$\begin{aligned} (Aw)(\cos \theta, \cos \gamma) &= (A_\theta W)(\omega_0) \\ (3) \qquad \qquad \qquad &= \int w(v \cdot \omega') \, d'\omega' / \sigma_{n-1} \\ &= \int_0^\pi w(\cos \theta \cos \gamma + \sin \theta \sin \gamma \cos \phi) \sin^{n-2} \phi \, d\phi / \tau_{n-1}; \end{aligned}$$

the integrand is independent of coordinates  $\theta'_0, \dots, \theta'_{n-2}$  of  $\omega'$ , and we have replaced  $\theta'_{n-1}$  by  $\phi$ . Note the symmetry in  $\theta, \gamma$ .

Suppose  $w \in L_2^{(n)}$  has expansion

$$w(\cos \gamma) = \sum_{l=0}^\infty w_l P_l^{(n)}(\cos \gamma).$$

Then the spherical harmonic expansion of  $W(\omega) = w(v \cdot \omega)$  is

$$\begin{aligned} W(\omega) &= \sum_{l=0}^{\infty} w_l P_l^{(n)}(v \cdot \omega) \\ &= \sum_{l=0}^{\infty} [w_l/d_l^{(n)}] \sum_{|l|=l} Y_l^{(n)}(v) Y_l^{(n)}(\omega) \end{aligned}$$

convergent in  $L_2(S^n)$ . The spherical harmonics are eigenfunctions of  $A_\theta$ , whence

$$\begin{aligned} (A_\theta W)(\omega) &= \sum_{l=0}^{\infty} [w_l/d_l^{(n)}] P_l^{(n)}(\cos \theta) \sum_{|l|=l} Y_l^{(n)}(v) Y_l^{(n)}(\omega) \\ &= \sum_{l=0}^{\infty} w_l P_l^{(n)}(\cos \theta) P_l^{(n)}(v \cdot \omega). \end{aligned}$$

In other words,

$$\begin{aligned} (Aw)(\cos \theta, \cos \gamma) &= \sum_{l=0}^{\infty} w_l P_l^{(n)}(\cos \theta) P_l^{(n)}(\cos \gamma) \quad \text{if} \\ (4) \quad w(\cos \gamma) &= \sum_{l=0}^{\infty} w_l P_l^{(n)}(\cos \gamma), \quad w \in L_2^{(n)}. \end{aligned}$$

This is also straightforward from (3) and the addition formula for  $P_l^{(n)}$ .

Now we take up  $(A_\theta G, H)$  for the case where  $G, H$  are each polar, with poles at angle  $\chi, 0 \leq \chi \leq \pi$ . Since  $A_\theta$  commutes with each  $T_\rho$ , we may rotate coordinates so that  $G$  has pole  $\omega_0$ ; let  $H$  then have pole  $v$ . By the above, if  $G(\omega) = g(\cos \theta_n)$  has pole  $\omega_0$  then so does  $(A_\theta G)(\omega) = (Ag)(\cos \theta, \cos \theta_n)$ , so

$$(A_\theta G, H) = \int_{S^n} h(v \cdot \omega) [(Ag)(\cos \theta, \cos \theta_n)] d\omega / \sigma_n.$$

If we integrate first coordinates  $\theta_0, \dots, \theta_{n-1}$  of  $\omega$  we find

$$\begin{aligned} \int h(v \cdot \omega) \sin \theta_2 \cdots \sin^{n-2} \theta_{n-1} [d\theta_0] \cdots d\theta_{n-1} &= \sigma_{n-1} (A_{\theta_n} H)(\omega_0) \\ &= \sigma_{n-1} [(Ah)(\cos \theta_n, v \cdot \omega_0)] \\ &= \sigma_{n-1} [(Ah)(\cos \chi, \cos \theta_n)]; \end{aligned}$$

we have used  $v \cdot \omega_0 = \cos \chi$  and the symmetry of  $(Ah)(x, y)$ . Now we do the  $\theta_n$  integration, replacing  $\theta_n$  by  $\eta$ :

$$(5) \quad (A_\theta G, H) = \int_0^\pi [(Ah)(\cos \chi, \cos \eta)] [(Ag)(\cos \theta, \cos \eta)] \sin^{n-1} \eta d\eta / \tau_n,$$

where  $0 \leq \chi \leq \pi$  is the angle between the poles of polar  $G$  and  $H$ .

We can interchange  $g$  and  $h$  in the integral, using  $(A_\theta G, H) = (G, A_\theta H)$ . From this it is seen that the integral is also symmetric in  $\theta, \chi$ , although this is not geometrically obvious.

Let  $g \in L_2^{(n)}$  have expansion

$$g(\cos \theta) = \sum_{l=0}^{\infty} g_l P_l^{(n)}(\cos \theta)$$

and let  $h_l, l \geq 0$ , be the corresponding coefficients for  $h$ . From (4) and the orthogonality



of the  $P_l^{(n)}$ ,

$$(6) \quad (A_\theta G, H) = \sum_{l=0}^\infty [g_l h_l / d_l^{(n)}] P_l^{(n)}(\cos \theta) P_l^{(n)}(\cos \chi),$$

where  $0 \leq \chi \leq \pi$  is the angle between the poles of polar  $G$  and  $H$ .

The expansions of  $g, h \in L_2^{(n)}$  being as above, the spherical convolution  $(g * h)(\cos \theta)$ ,  $0 \leq \theta \leq \pi$ , can be defined by its expansion

$$(g * h)(\cos \theta) = \sum_{l=0}^\infty [g_l h_l / d_l^{(n)}] P_l^{(n)}(\cos \theta), \quad 0 \leq \theta \leq \pi.$$

Comparing with (6), this can be regarded either as  $(A_\theta G, H)$ , where polar functions  $G, H$  (from  $g, h$ ) have the same pole, or else as  $(G, H)$ , where the poles of polar  $G, H$  are at angle  $\theta$ . From (3) we have explicitly

$$\begin{aligned} (g * h)(\cos \theta) &= \int_0^\pi h(\cos \gamma) [(Ag)(\cos \theta, \cos \gamma)] \sin^{n-1} \gamma d\gamma / \tau_n \\ &= \int_0^\pi \int_0^\pi h(\cos \gamma) g(\cos \theta \cos \gamma + \sin \theta \sin \gamma \cos \phi) \\ &\quad \cdot \sin^{n-2} \phi \sin^{n-1} \gamma d\phi d\gamma / (\tau_{n-1} \tau_n) \\ &= (h * g)(\cos \theta). \end{aligned}$$

Since

$$\begin{aligned} \sum_0^\infty |g_l h_l / d_l^{(n)}| &\leq \left[ \sum_0^\infty g_l^2 / d_l^{(n)} \right]^{1/2} \left[ \sum_0^\infty h_l^2 / d_l^{(n)} \right]^{1/2} \\ &= \|g\|_2 \|h\|_2, \end{aligned}$$

it follows that the series for  $g * h$  converges uniformly to a continuous function when  $g, h \in L_2^{(n)}$ . (It is also true that

$$\|g * h\|_r \leq \|g\|_p \|h\|_q, \quad \frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1, \quad 1 \leq p, q, r \leq \infty,$$

so that  $L_1^{(n)}$  is a commutative Banach algebra under  $*$  multiplication and each  $L_p^{(n)}$  is a commutative Banach module over  $*$  operators  $L_1^{(n)}$  [7, Chap. 2], [8].) By the same argument,  $(A_\theta G, H)$  is jointly continuous in  $0 \leq \theta, \chi \leq \pi$  for any  $g, h \in L_2^{(n)}$ .

By a cap function we mean a polar function  $w(v \cdot \omega)$  such that  $w(\cos \theta) = 0$ ,  $\beta \leq \theta \leq \pi$ , for some  $0 < \beta < \pi$ . That is, the function is supported by a spherical cap of angular radius  $\beta$  around the center  $v$  of the cap. The size  $\beta$  will be mentioned when necessary. Consider  $(Aw)(\cos \theta, \cos \gamma)$  for such a cap function. From

$$\begin{aligned} \cos \theta \cos \gamma + \sin \theta \sin \gamma \cos \phi &= \cos(\theta - \gamma) - (1 - \cos \phi) \sin \theta \sin \gamma \\ &\leq \cos(\theta - \gamma) \end{aligned}$$

it is apparent in (3) that  $(Aw)(\cos \theta, \cos \gamma) = 0$  unless  $\cos(\theta - \gamma) > \cos \beta$ , i.e.,  $\theta - \beta < \gamma < \theta + \beta$ . This is a restatement of the geometrically obvious fact that for  $W(\omega) = w(v \cdot \omega)$ ,  $(A_\theta W)(\omega) = 0$  unless the support of the averaging measure, i.e., the edge of the cap of radius  $\theta$  around  $\omega$ , meets the support of  $W$ , i.e., within  $\beta$  of  $v$ .

Suppose  $G, H$  in  $(A_\theta G, H)$  are each cap functions of cap size  $\beta$ . The integral (5) vanishes when the  $\eta$  intervals  $(\chi - \beta, \chi + \beta)$  and  $(\theta - \beta, \theta + \beta)$  are disjoint, using the

property of  $(Aw)(x, y)$  just described. It follows that  $(A_\theta G, H) = 0$  unless  $\chi - 2\beta < \theta < \chi + 2\beta$ , where, again,  $\chi$  is the angle between the poles of cap functions  $G, H$ , and  $\beta$  the cap size of each. The geometrical interpretation is straightforward.

**7. Minimum distance codes.** Let  $\omega_\mu, 1 \leq \mu \leq N$ , be  $N > 1$  distinct points on  $S^n$ , at mutual angles  $0 \leq \theta_{\mu\nu} \leq \pi$  given by  $\cos \theta_{\mu\nu} = \omega_\mu \cdot \omega_\nu, 1 \leq \mu, \nu \leq N$ . If  $0 < \alpha \leq \pi/2$  is defined as  $\alpha = \frac{1}{2} \min \{\theta_{\mu\nu} : 1 \leq \mu < \nu \leq N\}$  then  $N$  disjoint open spherical caps of angular radius  $\alpha$  will fit on  $S^n$ , each centered at an  $\omega_\mu$ . Such a configuration is called a code of size  $N$  on  $S^n$ , with minimum distance  $2\alpha$ .

Let  $\omega_\mu, 1 \leq \mu \leq N$ , be a code with minimum distance  $0 < 2\alpha \leq \pi$ . With  $\beta > 0$  to be specified later, suppose  $\nu \in L_2^{(n)} \subset L_1^{(n)}$  has the properties

$$\begin{aligned} \nu(\cos \theta) &= 0, & \beta &\leq \theta \leq \pi, \\ \int_0^\beta \nu(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n &= 1. \end{aligned}$$

Then the cap function  $V_\mu(\omega) = \nu(\omega_\mu \cdot \omega), \omega \in S_n$  is a copy of  $\nu$  around  $\omega_\mu$ ; we define

$$\begin{aligned} w_{\mu\nu}(\cos \theta) &= (A_\theta V_\mu, V_\nu) \\ &= \int_0^\pi [(A\nu)(\cos \theta, \cos \eta)][(A\nu)(\cos \theta_{\mu\nu}, \cos \eta)] \sin^{n-1} \eta \, d\eta / \tau_n, \\ & \qquad \qquad \qquad 0 \leq \theta \leq \pi, \quad 1 \leq \mu, \quad \nu \leq N, \end{aligned}$$

the angle between poles  $\omega_\mu, \omega_\nu$  being  $\theta_{\mu\nu}$ . We have  $w_{\mu\nu}(\cos \theta) = 0$  unless  $\theta_{\mu\nu} - 2\beta < \theta < \theta_{\mu\nu} + 2\beta$ , by the previous discussion.

Now let  $\varepsilon$  satisfying  $0 < \varepsilon < \alpha$  be chosen and fixed. We will require henceforth that the cap size  $\beta$  of  $\nu$  satisfy  $0 < 2\beta < \varepsilon$ . Since  $\theta_{\mu\nu} - 2\beta > 2\alpha - \varepsilon, \mu \neq \nu$ , we see that

$$\begin{aligned} w_{\mu\mu}(\cos \theta) &= 0, & \varepsilon &\leq \theta \leq \pi, \\ w_{\mu\nu}(\cos \theta) &= 0, & 0 &\leq \theta \leq 2\alpha - \varepsilon, \quad \mu \neq \nu \\ & \text{provided } 0 < 2\beta < \varepsilon < 2\alpha - \varepsilon. \end{aligned}$$

There follows

$$\begin{aligned} \int_0^\varepsilon w_{\mu\nu}(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n &= \delta_{\mu\nu} \int_0^\pi w_{\mu\nu}(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \\ &= \delta_{\mu\nu} \int_0^\pi (A_\theta V_\mu, V_\mu) \sin^{n-1} \theta \, d\theta / \tau_n \\ &= \delta_{\mu\nu} (V_\mu, 1)^2 \\ &= \delta_{\mu\nu}, \quad 1 \leq \mu, \nu \leq N \quad \text{provided } 0 < 2\beta < \varepsilon < 2\alpha - \varepsilon. \end{aligned}$$

If  $\nu \in L_2^{(n)}$  above has expansion

$$\nu(\cos \theta) = \sum_{l=0}^\infty \nu_l P_l^{(n)}(\cos \theta),$$

the expansion of  $w_{\mu\nu}$  is

$$(7) \quad w_{\mu\nu}(\cos \theta) = \sum_{l=0}^\infty [\nu_l^2 / d_l^{(n)}] P_l^{(n)}(\cos \theta_{\mu\nu}) P_l^{(n)}(\cos \theta),$$

$$0 \leq \theta \leq \pi, \quad 1 \leq \mu, \quad \nu \leq N,$$

using (6); the series for  $w_{\mu\nu}(\cos \theta)$  converges uniformly on  $0 \leq \theta \leq \pi$  to a continuous function.

The function

$$V(\omega) = (1/N) \sum_{\mu=1}^N V_{\mu}(\omega), \quad \omega \in S^n$$

is a superposition of copies of  $(1/N)\nu$  around each code point  $\omega_{\mu}$ ,  $1 \leq \mu \leq N$ . Consider

$$(8) \quad \begin{aligned} w(\cos \theta) &= (A_{\theta}V, V) \\ &= (1/N^2) \sum_{\mu} \sum_{\nu} w_{\mu\nu}(\cos \theta), \quad 0 \leq \theta \leq \pi. \end{aligned}$$

This satisfies

$$\begin{aligned} \int_0^{\pi} w(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n &= (1, V)^2 \\ &= 1. \end{aligned}$$

It also satisfies

$$\begin{aligned} w(\cos \theta) &= 0, \quad \varepsilon \leq \theta \leq 2\alpha - \varepsilon, \\ \int_0^{\varepsilon} w(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n &= 1/N \quad \text{provided } 0 < 2\beta < \varepsilon < 2\alpha - \varepsilon \end{aligned}$$

using the properties given previously for the  $w_{\mu\nu}(\cos \theta)$ .

The expansion coefficients in

$$w(\cos \theta) = \sum_0^{\infty} w_l P_l^{(n)}(\cos \theta), \quad 0 \leq \theta \leq \pi$$

are seen to be

$$(9) \quad \begin{aligned} w_l &= [\nu_l^2 / d_l^{(n)}] M_l \quad \text{for} \\ M_l &= (1/N^2) \sum_{\mu} \sum_{\nu} P_l^{(n)}(\cos \theta_{\mu\nu}) \\ &= (1/d_l^{(n)}) \sum_{|l|=l} \left[ (1/N) \sum_{\mu} Y_l^{(n)}(\omega_{\mu}) \right]^2 \\ &\geq 0, \quad l \geq 0 \end{aligned}$$

using (7) for the coefficients of  $w_{\mu\nu}(\cos \theta)$ . (The property  $w_l \geq 0$  derives from (2):  $w_l = (J_l V, V) = (J_l V, J_l V) = [\|J_l V\|_2]^2 \geq 0$ ,  $l \geq 0$ ; cf. [1, Thm. 8].)

We exhibit the first few of the  $w_l$ .

$$l = 0: \quad \nu_0 = \int_0^{\pi} \nu(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n = 1,$$

$$M_0 = (1/N^2) \sum_{\mu} \sum_{\nu} 1 = 1,$$

$$w_0 = [\nu_0^2 / d_0^{(n)}] M_0 = 1$$

$$\left( = \int_0^{\pi} w(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \right)$$

$$l = 1: \quad M_1 = (1/N^2) \sum_{\mu} \sum_{\nu} \omega_{\mu} \cdot \omega_{\nu}$$

$$= \mathbf{m} \cdot \mathbf{m} \quad \text{where}$$

$$\mathbf{m} = (1/N) \sum_{\mu} \omega_{\mu} \text{ is the centroid in } E^{n+1} \text{ of the code,}$$

$$w_1 = [\nu_1^2 / (n + 1)] \mathbf{m} \cdot \mathbf{m}.$$

$$l = 2: \quad M_2 = (nN^2)^{-1} \sum_{\mu} \sum_{\nu} [(n + 1)(\omega_{\mu} \cdot \omega_{\nu})^2 - 1]$$

$$= [(n + 1)/n] \Phi : \Phi$$

$$w_2 = \frac{2(n + 1)\nu_2^2}{n^2(n + 3)} \Phi : \Phi$$

for traceless symmetric  $E^{n+1}$  dyadic  $\Phi$  defined by

$$\Phi = (1/N) \sum_{\mu} [\omega_{\mu} \omega_{\mu} - (n + 1)^{-1} I_{n+1}],$$

$I_{n+1}$  being the unit dyadic. The  $M_l$  for large  $l$  will involve a traceless symmetric  $l$ -adic in  $\omega$  averaged over the  $\omega_{\mu}$ ,  $1 \leq \mu \leq N$  (tensor form of spherical harmonics.)

The limit as  $\varepsilon \rightarrow 0$  will be of interest. From (8),  $w(\cos \theta)$  behaves as  $(1/N) \delta_+(0)$  near  $\theta = 0$ , and it can be shown that the rest of  $w(\cos \theta)$  behaves as  $(1/N^2) \sum_{\mu \neq \nu} \delta(\theta - \theta_{\mu\nu})$ . The moment squares  $M_l$  do not depend on the cap function  $\nu$ .

To get the variation of  $\nu_l$  with  $\varepsilon$  consider

$$d_l^{(n)} - \nu_l = d_l^{(n)} \int_0^{\varepsilon/2} \nu(\cos \theta) [1 - P_l^{(n)}(\cos \theta)] \sin^{n-1} \theta \, d\theta / \tau_n;$$

we have used here the fact that  $\nu$  integrates to unity and has support  $0 \leq \theta < \beta < \varepsilon/2$ . If  $x_{l1}$  denotes the largest of the zeros of  $P_l^{(n)}(x)$ , then  $P_l^{(n)}(x)$  increases monotonely from 0 to 1 on  $x_{l1} \leq x \leq 1$ . There follows

$$|d_l^{(n)} - \nu_l| \leq d_l^{(n)} \int_0^{\varepsilon/2} |\nu(\cos \theta)| [1 - P_l^{(n)}(\cos \theta)] \sin^{n-1} \theta \, d\theta / \tau_n$$

$$\leq d_l^{(n)} [1 - P_l^{(n)}(\cos(\varepsilon/2))] \|\nu\|_1 \quad \text{provided } \cos(\varepsilon/2) \geq x_{l1}.$$

If the various  $\nu$ 's all satisfy  $\nu \geq 0$  then  $\|\nu\|_1 = 1$  and  $P_l^{(n)}(\cos(\varepsilon/2)) \leq \nu_l / d_l^{(n)} \leq 1$  provided  $\cos(\varepsilon/2) \geq x_{l1}$ . Otherwise, we must make the assumption that the family of  $\nu$ 's involved is bounded in  $L_1^{(n)}$ , i.e.,  $\|\nu\|_1 \leq a$  for some fixed  $1 < a < \infty$ . We obtain

$$\lim_{\varepsilon \rightarrow 0} \nu_l = d_l^{(n)}, \quad \lim_{\varepsilon \rightarrow 0} w_l = d_l^{(n)} M_l, \quad l \geq 0.$$

We do not expect the limit to be uniform in  $l$ , since  $x_{l1} \rightarrow 1$  as  $l \rightarrow \infty$ .

**8. Linear programming bound on the code size.** The code size problem is: if a positive lower bound on the minimum distance  $2\alpha$  is specified, what is the largest  $N$  can be? The special value  $\alpha = \pi/6$  corresponds to finding the largest number of nonoverlapping unit spheres in  $E^{n+1}$  which can touch the central unit sphere  $S^n$ . Various linear programs give a bound which is the analog of the bound in [1]. A discussion of programming in linear spaces is given by Hurwicz [9, Chapter 4].

Consider the following linear program.

- (i)  $0 < \epsilon < \alpha \leq \pi/2$  are given.
- (ii) Minimize

$$(10) \quad \lambda = \int_0^\epsilon u(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n$$

over all continuous functions  $u(x) \geq 0, -1 \leq x \leq 1$ , satisfying

- (iii)  $u(\cos \theta) = 0, \epsilon \leq \theta \leq 2\alpha - \epsilon,$
- (iv)  $u_l \geq \delta_{l0}, l \geq 0,$  where

$$u_l = d_l^{(n)} \int_0^\pi u(\cos \theta) P_l^{(n)}(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n.$$

If  $\omega_\mu, 1 \leq \mu \leq N$ , is a code on  $S^n$  of minimum distance  $\geq 2\alpha$ , and if cap function  $\nu$  of § 7 satisfies  $\nu \geq 0$ , then  $u(\cos \theta) = (A_\theta V, V)$  of (8) is feasible with  $\lambda = 1/N$ . If the value  $\lambda(\alpha, \epsilon)$  is the infimum of  $\lambda$ 's in (10ii), then

$$N \leq \frac{1}{\lambda(\alpha, \epsilon)}$$

is an upper bound on the code size. This bound is the direct generalization of linear programming problem (I) of [1]. Observe that as  $\epsilon \rightarrow 0$  the constraints on  $u(\cos \theta)$  increase, so  $\lambda(\alpha, \epsilon)$  increases to a best value  $\lambda(\alpha)$  as  $\epsilon \rightarrow 0$ . In the limit, however, there is no feasible  $u$ , the  $u$ 's are trying to become  $\delta$  functions.

The dual program, the analog of linear programming problem (II) of [1], does have a limiting form as  $\epsilon \rightarrow 0$ , as follows.

- (i)  $0 < \alpha \leq \pi/2$  is given.
- (ii) Maximize

$$(11) \quad \eta_0 = \int_0^\pi \eta(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n$$

over all continuous functions  $\eta(x), -1 \leq x \leq 1$ , satisfying

- (iii)  $\eta(\cos \theta) \leq 0, 2\alpha \leq \theta \leq \pi,$
- (iv)  $\eta_l \geq 0, l \geq 0,$  where

$$\eta_l = d_l^{(n)} \int_0^\pi \eta(\cos \theta) P_l^{(n)}(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n,$$

- (v)  $\eta(1) = \sum_{l=0}^\infty \eta_l \leq 1.$

The supremum of  $\eta_0$ 's in (ii) is the limiting value  $\lambda(\alpha)$  for program (10). This version of the bound is equivalent to [10, Thm. 4.3], except that we do not require  $\eta$  to be a polynomial. However,  $\eta_l \geq 0, \sum_{l=0}^\infty \eta_l \leq 1$  imply that  $\sum_{l=0}^\infty \eta_l P_l^{(n)}$  converges uniformly from which it can be shown that the bounds are the same.

In both versions above the constraints involve function values on an interval and also the nonnegativity of the  $P_l^{(n)}$  expansion coefficients. The following result applies to the coefficient constraint.

LEMMA. An  $f(\cos \theta) = \sum_{l=0}^\infty f_l P_l^{(n)}(\cos \theta), 0 \leq \theta \leq \pi$  satisfies

$$f_l \geq 0 \text{ for all } l \geq 0,$$

$$f(1) = \sum_{l=0}^\infty f_l < \infty$$

iff it is of the form  $f = g * g$  for some  $g \in L_2^{(n)}$ ; all such  $g$  have the form

$$g(\cos \theta) = \sum_{l=0}^{\infty} \pm \sqrt{d_l^{(n)}} f_l P_l^{(n)}(\cos \theta) \quad \text{in } L_2^{(n)}.$$

*Proof.*  $f$  being given, if  $g_l = \pm \sqrt{d_l^{(n)}} f_l$ ,  $l \geq 0$ , then  $\sum_0^{\infty} g_l^2 / d_l^{(n)} = \sum_0^{\infty} f_l^2 < \infty$ , so  $\{g_l\}_0^{\infty}$  are the coefficients of some  $g \in L_2^{(n)}$ . The coefficients of  $g * g$  are then  $g_l^2 / d_l^{(n)} = f_l^2$ ,  $l \geq 0$ . The converse is equally trivial. Notice that each  $(\pm)_l$  is arbitrary, so there are  $2^a$  such  $g$ 's where  $0 \leq a \leq \aleph_0$  is the cardinal of  $\{l : f_l > 0\}$ .

Using this, we restate the program (10) as

- (i)  $0 < \varepsilon < \alpha \leq \pi/2$  are given.
- (ii) Minimize

$$(12) \quad \lambda = \int_0^{\varepsilon} (g * g)(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n$$

over all  $g \in L_2^{(n)}$  with the properties

- (iii)  $(g * g)(\cos \theta) \geq 0$ ,  $0 \leq \theta \leq \pi$ ,
- (iv)  $(g * g)(\cos \theta) = 0$ ,  $\varepsilon \leq \theta \leq 2\alpha - \varepsilon$ ,
- (v)  $g_0 = \int_0^{\pi} g(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \geq 1$ .

The infimum of the  $\lambda$  values is  $\lambda(\alpha, \varepsilon)$  as before; we have used  $(-g) * (-g) = g * g$  and  $(g * g, 1) = (\pm g, 1)^2$ .

In the same way, the program (11) can be rewritten as

- (i)  $0 < \alpha \leq \pi/2$  is given.
- (ii) Maximize

$$(13) \quad h_0 = \int_0^{\pi} h(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n$$

over all  $h \in L_2^{(n)}$  with the properties

- (iii)  $(h * h)(\cos \theta) \leq 0$ ,  $2\alpha \leq \theta \leq \pi$ ,
- (iv)  $\int_0^{\pi} h^2(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \leq 1$ .

In (iv) we have used  $(h * h)(1) = [\|h\|_2]^2$ ; the supremum of  $h_0^2$  in (ii) is the value  $\lambda(\alpha)$  of program (11).

In the reformulated programs the coefficient constraint is built in, but the values of  $f * f$  are constrained on an interval. In the first version, conditions (12ii-iv) are not simply expressible in terms of  $g$ , although (12iv) has strong implications. In the dual, (13iii) is satisfied if  $h$  has the property

$$h(\cos \theta) = 0, \quad \alpha \leq \theta \leq \pi$$

from the results of § 6. It is straightforward that under this constraint the best  $h$  is constant on  $0 \leq \theta < \alpha$ . If  $\rho(\gamma)$  denotes the fractional area of  $S^n$  covered by a cap of radius  $\gamma$ ,

$$\rho(\gamma) = \int_0^{\gamma} \sin^{n-1} \theta \, d\theta / \tau_n, \quad 0 \leq \gamma \leq \pi$$

then the following is feasible in (13):

$$\begin{aligned} h(\cos \theta) &= [\rho(\alpha)]^{-1/2}, & 0 \leq \theta < \alpha, \\ &= 0, & \alpha \leq \theta \leq \pi \end{aligned}$$

and gives the elementary bound [11]

$$N \leq \frac{1}{\rho(\alpha)}.$$

A related argument gives an upper bound for the value in (13), namely,

$$\lambda(\alpha) \leq \rho(2\alpha).$$

This is a consequence of

$$\begin{aligned} h_0^2 &= \left[ \int_0^\pi h(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \right]^2 \\ &= \int_0^{2\alpha} + \int_{2\alpha}^\pi (h * h)(\cos \theta) \sin^{n-1} \theta \, d\theta / \tau_n \\ &\leq [(h * h)(1)] \rho(2\alpha), \end{aligned}$$

where we have used (13 iii) and also  $(h * h)(\cos \theta) \leq (h * h)(1)$ ,  $0 \leq \theta \leq \pi$ . Curiously enough,

$$N \geq \left[ \frac{1}{\rho(2\alpha)} \right] \quad (\text{integer part})$$

is a known lower bound on the maximal code size, but the proof is not related to the above.

**Acknowledgments.** The author wishes to acknowledge the many stimulating conversations with colleagues N. J. A. Sloane and A. M. Odlyzko. Numerical results for the linear programming bound are given in [12] along with a simplified derivation of the linear program.

A referee points out that the associator as defined above is a mean value operator in the sense of [13, p. 435] with the spherical harmonics as eigenfunctions [13, p. 438].

#### REFERENCES

- [1] N. J. A. SLOANE, *An introduction to association schemes and coding theory*, Theory and Applications of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975, pp. 225–260.
- [2] R. R. COIFMAN AND GUIDO WEISS, *Representations of compact groups and spherical harmonics*, Enseignement Math., 14 (1968), pp. 121–173.
- [3] A. J. JAMES, *Normal multivariate analysis and the orthogonal group*, Ann. Math. Stat., (1954), pp. 40–75.
- [4] A. ERDÉLYI, *Higher Transcendental Functions*, McGraw-Hill, New York, 1953.
- [5] MILTON ABRAMOWITZ AND IRENE A. STEGUN, *Handbook of Mathematical Functions*, NBS Applied Mathematics Series 55, U.S. Government Printing Office, Washington, DC, 1964.
- [6] C. F. DUNKL, *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups*, Indiana Univ. Math. J., 25 (1976), pp. 335–358.
- [7] RICHARD ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.
- [8] GEORGE GASPER, *Banach algebras of Jacobi series and positivity of a kernel*, Ann. of Math., 95 (1972), pp. 261–280.
- [9] KENNETH J. ARROW, LEONID HURWICZ AND HIROFUMI UZAWA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford CA, 1958.
- [10] P. DELSARTE, J. M. GOETHALS AND J. J. SEIDEL, *Spherical codes and designs*, Geometriae Dedicata, 6 (1977), pp. 363–388.
- [11] R. A. RANKIN, *The closest packing of spherical caps in n dimensions*, Proc. Glasgow Math. Assoc., 2 (1955), pp. 139–144.

- [12] A. M. ODLYZKO AND N. J. A. SLOANE, *New bounds on the number of unit spheres that can touch a unit sphere in  $n$  dimensions*. J. Combinatorial Theory Ser. A, 26 (1979), pp. 210–214.
- [13] SIGURDUR HELGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.



## FOURIER EXPANSIONS OF RATIONAL FRACTIONS OF ELLIPTIC INTEGRALS AND JACOBIAN ELLIPTIC FUNCTIONS\*

R. G. LANGEBARTEL†

**Abstract.** The Fourier expansions of rational fractions with numerators consisting of various combinations of  $sn(u, k)$ ,  $cn(u, k)$ ,  $dn(u, k)$ , and the periodic parts of the elliptic integrals  $E(am u, k)$  and  $\Pi(am u, \alpha^2, k)$ , and denominators consisting of the first or second powers of  $1 \pm \beta cn u$  or  $1 - \alpha^2 sn^2 u$  are listed. The parameter ranges are  $0 < k < 1$  and  $-\infty < \alpha^2 < 0$ .

**Introduction.** Recent work in earth satellite orbit theory and earth-moon trajectory theory [2 vol. II, p. 713], [5], and close binary star systems [4] has brought about renewed interest in the two-fixed-centers problem, the two-center orbit serving as an intermediate orbit. Elliptic functions arise naturally in such problems and treatment of perturbations involves Fourier expansions of certain combinations of elliptic functions and elliptic integrals. These combinations of functions occur as rational fractions with products of elliptic integrals and elliptic functions in the numerator, and  $1 \pm \beta cn u$  or  $1 - \alpha^2 sn^2 u$  (or their powers) in the denominator. The availability of these expansions eliminates or greatly reduces the number of Fourier series multiplications that otherwise appear in orbit theory. Where elliptic integrals of the third kind appear, it is the circular case,  $\alpha^2$  negative, that is of interest and the pertinent expansions are derived for this case.

**1. Definitions.** The notation and definitions for the standard functions follow the convention in the book of Byrd and Friedman [1].

The elliptic integrals are not periodic, consisting, as they do, of a linear term plus a periodic term. For the use to which the Fourier expansions are to be put, it is convenient to have expansions for expressions involving only the periodic parts of the elliptic integrals. The expansions listed below are of this type. Just as the Jacobian zeta function,  $Z(u, k)$ , is the periodic part of the elliptic integral of the second kind,  $E(am u, k)$ , so we define  $\Omega(u, \alpha^2, k)$  to be the periodic part of the elliptic integral of the third kind,  $\Pi(am u, \alpha^2, k)$ :

$$(1.1) \quad \begin{aligned} Z(u, k) &= E(am u, k) - Eu/K, \\ \Omega(u, \alpha^2, k) &= \Pi(am u, \alpha^2, k) - \Pi u/K. \end{aligned}$$

(The periodic part of the elliptic integral of the first kind,  $F(am u, k) = u$ , is evidently zero.)

The elliptic functions and integrals depend on the fundamental parameters  $k$  and  $\alpha^2$ . We introduce the related parameters:

$$(1.2) \quad \begin{aligned} k' &= \sqrt{1 - k^2}, & \beta' &= \sqrt{1 - \beta^2} = (\alpha')^{-1}, \\ \alpha' &= \sqrt{1 - \alpha^2}, & \eta &= \sqrt{k^2 + \beta^2 k'^2}, \\ \beta &= \sqrt{-\alpha^2 / \alpha'^2}, & \lambda &= \beta(\alpha' \eta)^{-1}. \end{aligned}$$

We define  $v_0$  by

$$(1.3) \quad cn(v_0, k') = \beta, \quad 0 < v_0 < K(k'),$$

\* Received by the editors May 22, 1978.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

and the parameters  $w, w_0$  by

$$(1.4) \quad w = \frac{\pi[K(k') - v_0]}{2K(k)}, \quad w_0 = \frac{\pi K(k')}{2K(k)}.$$

The parameter ranges considered are

$$(1.5) \quad \begin{aligned} 0 < k < 1, & \quad 1 < \alpha' < \infty, \\ 0 < k' < 1, & \quad 0 < \beta < 1, \\ 0 < -\alpha^2 < \infty, & \quad 0 < \beta' < 1. \end{aligned}$$

There is the further notation:

$$(1.6) \quad \begin{aligned} K &= K(k), & E &= E(k) = E(am K, k), & \Pi &= \Pi(am K, \alpha^2, k), \\ K' &= K(k'), & E' &= E(k') = E(am K', k'), & \Pi' &= \Pi(am K', \alpha'^2, k'); \end{aligned}$$

$$(1.7) \quad A(z; n) = e^{nz} + (-1)^n e^{-nz}, \quad B(z; n) = e^{nz} + (-1)^{n+1} e^{-nz}.$$

**2. Expansions.** The expansions (2.1), (2.2), (2.3) for  $Z(u, k)$ ,  $\Omega(u, \alpha^2, k)$  and  $(1 - \alpha^2 sn^2 u)^{-1}$  are not new. The first two go back at least as far as Jacobi [3] and the third is immediately obtained from Jacobi's expansion for  $\Pi(am u, \alpha^2, k)$  by differentiation. A residue theory derivation of the expansion for  $(1 - \alpha^2 sn^2 u)^{-1}$  has been given by I. G. Izsak [2, vol. I, p. 209]. These three expansions are listed for easy reference.

Where the ambiguous sign,  $\pm$ , occurs, the top signs are to be taken together throughout the formula and similarly for the bottom signs.

$$(2.1) \quad Z(u, k) = \frac{\pi}{K} \sum_{n=1}^{\infty} \frac{\sin(n\pi u/K)}{\sinh 2nw_0},$$

$$(2.2) \quad \Omega(u, \alpha^2, k) = \lambda \sum_{n=1}^{\infty} \frac{\sinh 2nw \sin(n\pi u/K)}{n \sinh 2nw_0},$$

$$(2.3) \quad \frac{1}{1 - \alpha^2 sn^2 u} = \frac{\Pi}{K} + \frac{\pi\lambda}{K} \sum_{n=1}^{\infty} \frac{\sinh 2nw \cos(n\pi u/K)}{\sinh 2nw_0},$$

$$(2.4) \quad \frac{1}{1 \pm \beta cn u} = \frac{\alpha'^2 \Pi}{K} \mp \frac{\alpha'^2 \pi\lambda}{K} \sum_{n=1}^{\infty} \frac{B(w; \mp n) \cos(n\pi u/2K)}{B(w_0; n)},$$

$$(2.5) \quad \frac{sn u}{1 \pm \beta cn u} = \frac{\pi}{\eta K} \sum_{n=1}^{\infty} \frac{B(w; \mp n) \sin(n\pi u/2K)}{A(w_0; n)},$$

$$(2.6) \quad \frac{sn u \, dn u}{1 \pm \beta cn u} = \mp \frac{\pi}{\beta K} \sum_{n=1}^{\infty} \frac{A(w; \mp n) \sin(n\pi u/2K)}{B(w_0; n)} \pm \frac{1}{\beta} Z(u, k),$$

$$(2.7) \quad \frac{dn u}{1 \pm \beta cn u} = \frac{\pi\beta'}{2K} + \frac{\pi\beta'}{K} \sum_{n=1}^{\infty} \frac{A(w; \mp n) \cos(n\pi u/2K)}{A(w_0; n)},$$

$$(2.8) \quad \begin{aligned} \frac{1}{(1 \pm \beta cn u)^2} &= \alpha'^2 K^{-1} [(2\alpha'^2 - \alpha'^4 \lambda^2) \Pi - K + \alpha'^2 \lambda^2 E] \\ &+ \frac{\alpha'^4 \pi}{K} \sum_{n=1}^{\infty} [n\pi \lambda^2 (2K)^{-1} A(w; \mp n) \mp \lambda (2 - \alpha'^2 \lambda^2) B(w; \mp n)] \\ &\cdot \frac{\cos(n\pi u/2K)}{B(w_0; n)}, \end{aligned}$$

$$(2.9) \quad \left(\frac{\beta \pm cn u}{1 \pm \beta cn u}\right)^2 = \frac{1}{\eta^2 K} [\eta^2 K + \beta'^2 E - \Pi] + \frac{\pi \lambda^2}{\beta^2 K} \cdot \sum_{n=1}^{\infty} [n\pi(2K)^{-1}A(w; \mp n) \pm \alpha'^2 \lambda B(w; \mp n)] \frac{\cos(n\pi u/2K)}{B(w_0; n)},$$

$$(2.10) \quad \frac{(\pm\beta + cn u) dn u}{(1 \pm \beta cn u)^2} = \frac{\pi^2}{2\eta K^2} \sum_{n=1}^{\infty} \frac{nB(w; \mp n) \cos(n\pi u/2K)}{A(w_0; n)},$$

$$(2.11) \quad \frac{sn u dn u}{(1 \pm \beta cn u)^2} = \frac{\pi^2 \alpha'^2 \lambda}{2\beta K^2} \sum_{n=1}^{\infty} \frac{nB(w; \mp n) \sin(n\pi u/2K)}{B(w_0; n)},$$

$$(2.12) \quad \frac{[\pm 2\beta + (1 + \beta^2)cn u]sn u dn u}{(1 \pm \beta cn u)^2} = \frac{\pi}{K} \sum_{n=1}^{\infty} \left[ \mp \frac{n\pi \lambda B(w; \mp n)}{2\beta^2 K} - \frac{(1 + \beta^2)A(w; \mp n)}{\beta^2} \right] \frac{\sin(n\pi u/2K)}{B(w_0; n)} + \frac{1 + \beta^2}{\beta^2} Z(u, k),$$

$$(2.13) \quad \frac{sn^2 u}{(1 \pm \beta cn u)^2} = \frac{\alpha'^2 \Pi - E}{\eta^2 K} + \frac{\alpha' \pi}{\eta^3 K} \sum_{n=1}^{\infty} [\mp \beta B(w; \mp n) - n\pi \beta' \eta (2K)^{-1} A(w; \mp n)] \cdot \frac{\cos(n\pi u/2K)}{B(w_0; n)},$$

$$(2.14) \quad \frac{cn u}{1 - \alpha^2 sn^2 u} = \frac{\pi \lambda}{\beta K} \sum_{n=1}^{\infty} \frac{\cosh(2n - 1)w \cos[(2n - 1)\pi u/2K]}{\cosh(2n - 1)w_0},$$

$$(2.15) \quad \frac{sn u dn u}{1 - \alpha^2 sn^2 u} = \frac{\pi \beta'^2}{\beta K} \sum_{n=1}^{\infty} \frac{\sinh(2n - 1)w \sin[(2n - 1)\pi u/2K]}{\cosh(2n - 1)w_0},$$

$$(2.16) \quad \frac{1}{(1 - \alpha^2 sn^2 u)^2} = -(2\alpha'^4 \eta^2 K)^{-1} \{ \alpha^2 E + \alpha'^2 \eta^2 K - \alpha'^2 [k^2 + (1 + \alpha'^2) \eta^2] \Pi \} + \pi (2\alpha'^2 \eta^3 K)^{-1} \sum_{n=1}^{\infty} \{ n\pi \beta^2 \eta K^{-1} \cosh 2nw + \beta \beta' [k^2 + (1 + \alpha'^2) \eta^2] \sinh 2nw \} \frac{\cos(n\pi u/K)}{\sinh 2nw_0},$$

$$(2.17) \quad \frac{Z(u, k)}{1 \pm \beta cn u} = \sum_{n=1}^{\infty} \left\{ \pi (KK')^{-1} [K' + \alpha'^2 \beta^2 \Pi' + \alpha'^2 \lambda (w \pm w_0)] A(w; \mp n) + \frac{2(-1)^{n+1} \pi^2 \alpha'^2 \lambda \sinh n(w \pm w_0)}{K^2 B(w_0; n)} \right\} \frac{\sin(n\pi u/2K)}{B(w_0; n)},$$

$$(2.18) \quad \frac{Z(u, k)}{1 - \alpha^2 sn^2 u} = \pi \lambda K^{-1} \cdot \sum_{n=1}^{\infty} [-\pi (2K)^{-1} \coth 2nw_0 \sinh 2nw + \Pi (\lambda K)^{-1} \cosh 2nw] \cdot \frac{\sin(n\pi u/K)}{\sinh 2nw_0},$$

$$\begin{aligned}
 & \frac{(1 + \beta^2 \pm 2\beta cn u)Z(u, k)}{(1 \pm \beta cn u)^2} \\
 &= \sum_{n=1}^{\infty} \left\{ \frac{2\pi^2 \alpha' \beta^2 (-1)^{n+1} [\beta \sinh n(w \pm w_0) - n\pi\beta' \eta (2K)^{-1} \cosh n(w \pm w_0)]}{\eta^3 K^2 B(w_0; n)} \right. \\
 (2.19) \quad & \left. + \pi\beta^2 (\eta^2 K)^{-1} \left[ \frac{\eta^2}{\beta^2} + \frac{K-E}{K} + \alpha'^2 \beta^2 \frac{\Pi'}{K'} + \frac{\alpha'^2 \lambda (w \pm w_0)}{K'} \right] A(w; \mp n) \right. \\
 & \left. \pm n\pi\beta^2 (\eta^2 K)^{-1} \left[ \frac{\pi\beta' \eta}{2\beta K} + \frac{\pi\alpha' \beta \eta \Pi'}{2KK'} \right. \right. \\
 & \left. \left. + \frac{\pi(w \pm w_0)}{2KK'} \right] B(w; \mp n) \right\} \frac{\sin n\pi u / 2K}{B(w_0; n)},
 \end{aligned}$$

$$\begin{aligned}
 \frac{Z(u, k)}{(1 - \alpha'^2 sn^2 u)^2} &= \sum_{n=1}^{\infty} \left\{ \left[ \frac{\pi^2 \beta' \beta (\beta^2 - 3\eta^2 + \eta^2 \beta^2)}{4\eta^3 K^2} \sinh 2nw \right. \right. \\
 & \left. \left. - \frac{n\pi^3 \beta^2 \beta'^2}{4\eta^2 K^3} \cosh 2nw \right] \coth 2nw_0 \right. \\
 (2.20) \quad & \left. + \left[ -\frac{\pi\beta^2 \beta'^2}{2\eta^2 K} \left( \frac{n^2}{\beta^2} + \frac{K-E}{K} + \frac{\alpha'^2 \beta^2 \Pi'}{K'} + \frac{\alpha'^2 \lambda w}{K'} \right) \right. \right. \\
 & \left. \left. + \frac{\pi(3 - \beta^2) \Pi}{2K^2} \right] \cosh 2nw + \frac{n\pi^2 \beta^2 \beta'^2}{2\eta^2 K} \right. \\
 & \left. \cdot \left[ \frac{\beta' \eta}{\beta K} + \frac{\alpha' \beta \eta \Pi'}{KK'} + \frac{w}{KK'} \right] \sinh 2nw \right\} \frac{\sin(n\pi u / K)}{\sinh 2nw_0},
 \end{aligned}$$

$$\begin{aligned}
 \frac{\Omega(u, \alpha^2, k)}{1 \pm \beta cn u} &= \mp \sum_{n=1}^{\infty} \left\{ \frac{\mp 2\alpha'^2 \lambda^2 w [(-1)^{n+1} (\mp 2Kw + \pi K') \sinh n(w \pm w_0) \right.}{KK' B(w_0; n)} \\
 & \left. + (\mp 2Kw - \pi K') \sinh n(w \mp w_0)] \right. \\
 (2.21) \quad & \left. + (2\alpha'^2 \lambda^2 w K - 2\alpha'^2 \lambda \Pi K') (nKK')^{-1} \right. \\
 & \left. \cdot [1 + (-1)^{n+1} - B(w; \mp n)] \right. \\
 & \left. + \lambda n^{-1} [1 + (-1)^{n+1}] [2 - B(w; \mp n)] - n\alpha^2 \pi^2 \lambda (2K^2)^{-1} \right. \\
 & \left. \cdot \int_{v_0}^{K'} B(\pi(K' - s) / 2K; \mp n) \Omega(s, \alpha'^2, k') ds \right. \\
 & \left. + [1 + (-1)^{n+1}] \beta \lambda \pi K^{-1} \right. \\
 & \left. \cdot \int_{v_0}^{K'} \frac{\sinh [n\pi(K' - s) / 2K] ds}{\beta + cn(s, k')} \right\} \frac{\sin(n\pi u / 2K)}{B(w_0; n)},
 \end{aligned}$$

$$\begin{aligned}
 \frac{\Omega(u, \alpha^2, k)}{1 - \alpha'^2 sn^2 u} &= 2\pi\lambda K^{-1} \sum_{n=1}^{\infty} \left\{ \lambda w^2 K (\pi K')^{-1} \cosh 2nw - [(\lambda w K - K' \Pi) \right. \\
 (2.22) \quad & \left. \cdot (2n\pi K')^{-1} + \frac{1}{2} \lambda w \coth 2nw_0] \sinh 2nw \right. \\
 & \left. + n\pi\beta^2 (2K)^{-1} \int_{v_0}^{K'} \sinh [n\pi(K' - s) / K] \right. \\
 & \left. \cdot \Omega(s, \alpha'^2, k') ds \right\} \frac{\sin(n\pi u / K)}{\sinh 2nw_0}.
 \end{aligned}$$

The expansion (2.1) is valid for  $|\mathbf{I}(u)| < K'$ , the expansions (2.2)–(2.22) for  $|\mathbf{I}(u)| < v_0$ .

**3. Method.** The fundamental procedure used was the evaluation of the Fourier coefficients by contour integration. Some of the formulae can be obtained from others by algebraic or calculus-type operations. There is, for example, the relation

$$(3.1) \quad \frac{1}{1 - \alpha^2 \operatorname{sn}^2 u} = \frac{1}{2\alpha'^2} \left( \frac{1}{1 - \beta \operatorname{cn} u} + \frac{1}{1 + \beta \operatorname{cn} u} \right).$$

Since the functions  $\operatorname{sn} u$ ,  $\operatorname{cn} u$ ,  $\operatorname{dn} u$ ,  $Z(u)$ ,  $\Omega(u)$  all have  $4K$  as a period, the Fourier coefficients of a rational fraction combination,  $R(u)$ , of these functions are

$$c_n \int_{-2K}^{2K} e^{in\pi u/(2K)} R(u) du,$$

where  $c_n = (2iK)^{-1}$  if  $R(u)$  is odd, and  $c_0 = (4K)^{-1}$ ,  $c_n = (2K)^{-1}$ ,  $n \geq 1$ , if  $R(u)$  is even. The contour integral of  $e^{in\pi u/(2K)} R(u)$  is taken around the fundamental parallelogram with vertices  $-2K$ ,  $2K$ ,  $4K + 2iK'$ ,  $2iK'$ . Inside this parallelogram the functions  $\operatorname{sn} u$ ,  $\operatorname{cn} u$ ,  $\operatorname{dn} u$ , and  $Z(u)$  are all single-valued and have simple poles at  $iK'$  and  $2K + iK'$ . The function  $(1 - \beta \operatorname{cn} u)^{-1}$  has simple poles inside the parallelogram at  $z_1 = iv_0$  and  $z_3 = 2K + 2iK' - iv_0$ , with  $v_0$  given by (1.3); the poles of  $(1 + \beta \operatorname{cn} u)^{-1}$  come at  $z_2 = 2K + iv_0$  and  $z_4 = 2iK' - iv_0$ . Evidently,  $(1 - \alpha^2 \operatorname{sn}^2 u)^{-1}$  has simple poles at  $z_1$ ,  $z_2$ ,  $z_3$ ,  $z_4$ . The double periodicity of the elliptic functions enables us to express the integral around the parallelogram in terms of the integral from  $-2K$  to  $2K$  for those formulae not involving the elliptic integrals. Consequently, in these cases the Fourier coefficients can be obtained without difficulty by the theory of residues. The function  $Z(u)$  is only singly periodic and in those formulae involving it, recourse must be made to the relation

$$(3.2) \quad Z(u + 2iK') = Z(u) - i\pi/K,$$

in order to express the integral from  $4K + 2iK'$  to  $2iK'$  in terms of the one from  $-2K$  to  $2K$ . When the symmetry considerations do not show it to be zero the constant term can be obtained either by an appeal to known elliptic integration formulae [1] or by reducing it to a known trigonometric definite integral upon letting  $am u = \varphi$ .

Further considerations are necessary when  $\Omega(u, \alpha^2, k)$  is involved because of its multi-valued character. Its singularities inside the parallelogram are branch points at  $z_1$ ,  $z_2$ ,  $z_3$ ,  $z_4$ , where it is logarithmically infinite. The function can be made single-valued by introducing two branch cuts, one connecting  $z_1$  and  $z_4$  and the other connecting  $z_2$  and  $z_3$ . In the neighborhood of  $z_j$ ,

$$(3.3) \quad \Omega(u, \alpha^2, k) = \varepsilon_j \frac{i\beta}{2\alpha'\eta} \ln(u - z_j) + \dots,$$

where  $\varepsilon_j = -1$  for  $j = 1, 2$ , and  $\varepsilon_j = 1$  for  $j = 3, 4$ . Therefore, if  $u$  traces a closed loop around  $z_1, z_4$  without crossing the cut  $\Omega(u, \alpha^2, k)$  will return to its original value, and the same holds for  $z_2, z_3$ . Consequently, the process of finding the residues at  $z_j$  which was fundamental in deriving the previous formulae is now replaced by determining the values of the integrals around these two loops. As is customary in such cases we shrink each loop, traced in the positive direction, to two small circles around the  $z_j$  connected by two straight line segments, one on each side of the cut, and ultimately send the radii of the circles to zero.

The comment concerning the single periodicity of  $Z(u)$  also applies to  $\Omega(u)$ , the corresponding formula being

$$(3.4) \quad \Omega(u + 2iK') = \Omega(u) - 2i\lambda w.$$

This formula is to be understood as applying to the cut plane. It may be derived by first noting that

$$(3.5) \quad \begin{aligned} \Pi(am(u + 2iK'), \alpha^2, k) &= \int_0^{u+2iK'} \frac{dz}{1 - \alpha^2 sn^2 z} = \int_0^{2iK'} \frac{dz}{1 - \alpha^2 sn^2 z} + \int_{2iK'}^{2iK'+u} \frac{dz}{1 - \alpha^2 sn^2 z} \\ &= \Pi(am2iK', \alpha^2, k) + \int_0^u \frac{dz}{1 - \alpha^2 sn^2 z} \\ &= \Pi(am2iK', \alpha^2, k) + \Pi(am u, \alpha^2, k). \end{aligned}$$

Whether one obtains the value of  $\Pi(am2iK', \alpha^2, k)$  from its defining integral by going up the left side of the imaginary axis from 0 to  $2iK'$  with indentations to the left at the singularities  $z_1$  and  $z_4$ , or up the right side is immaterial since the result is the same as long as the cut is observed. Carrying out this process, letting  $z = i\zeta$ , using the imaginary argument transformation for  $sn(i\zeta, k)$  [1, p. 38], and letting the radii of the indentations go to zero gives

$$(3.6) \quad \Pi(am2iK', \alpha^2, k) = 2i\beta'^2 K' + 2i\beta^2 \Pi'.$$

It should be pointed out that this automatically defines  $\Pi'$  as the Cauchy principal value of the integral

$$\int_0^{2K'} \frac{d\zeta}{1 - \alpha'^2 sn^2(\zeta, k')}$$

with respect to the singularities at  $v_0$  and  $2K' - v_0$ . Combining standard complete elliptic integral formulae [1, pp. 230, 226] leads to

$$(3.7) \quad \alpha'^2 K' \Pi - K K' + \alpha^2 K \Pi' = \alpha'^2 \lambda w K,$$

a result valid for the parameter ranges (1.5). An appeal to the definition of  $\Omega(u, \alpha^2, k)$ , together with (3.5), (3.6), and (3.7) immediately validates (3.4).

We illustrate the details of evaluating the loop integrals by sketching the work for the expansion of  $\Omega(u) \cdot (1 - \alpha^2 sn^2 u)^{-1}$ . We consider first the integral

$$(3.8) \quad I(z_i; n) = \int_{c_1} \frac{e^{in\pi u/2K} \Omega(u) du}{1 - \alpha^2 sn^2 u}$$

taken along the lower half of the loop around  $z_1, z_4$ , i.e., the half-loop starting at  $iK'$ , traveling down the left side of the imaginary axis to just above  $z_1$ , swinging about  $z_1$  on a circular arc, and finally moving back up the right side of the imaginary axis to  $iK'$ . On the first line segment  $u = z_1 + r e^{-i3\pi/2}$ , on the second  $u = z_1 + r e^{i\pi/2}$ , and on the circle  $u = z_1 + \rho e^{i\theta}$ . If we use

$$(3.9) \quad \frac{\Omega(u)}{1 - \alpha^2 sn^2 u} = \frac{1}{2} \frac{d\Omega^2(u)}{du} + \frac{\Pi}{K} \Omega(u),$$

followed by an integration by parts, the integral taken along the first line segment

becomes

$$-i e^{-n\pi v_0/2K} \left\{ -\frac{i}{2} \left[ e^{-n\pi(K'-v_0)/2K} \Omega^2(z_1 + e^{-i3\pi/2}(K'-v_0)) - e^{-n\pi\rho/2K} \Omega^2(z_1 + e^{-i3\pi/2}\rho) \right. \right. \\ \left. \left. + \frac{n\pi}{2K} \int_{\rho}^{K'-v_0} e^{-n\pi r/2K} \Omega^2(z_1 + r e^{-i3\pi/2}) dr \right] \right. \\ \left. + \frac{\Pi}{K} \int_{\rho}^{K'-v_0} e^{-n\pi r/2K} \Omega(z_1 + r e^{-i3\pi/2}) dr \right\}.$$

A similar result is obtained for the second line segment and we find the half-loop integral from  $iK'$  back to  $iK'$  can be written as

$$I(z_1; n) = i e^{-n\pi v_0/2K} \left\{ -\frac{i}{2} e^{-n\pi(K'-v_0)/2K} [\Omega^2(z_1 + e^{i\pi/2}(K'-v_0)) \right. \\ \left. - \Omega^2(z_1 + e^{-i3\pi/2}(K'-v_0))] \right. \\ \left. + \frac{i}{2} e^{-n\pi\rho/2K} [\Omega^2(z_1 + e^{i\pi/2}\rho) - \Omega^2(z_1 + e^{-i3\pi/2}\rho)] \right. \\ (3.10) \quad \left. - \frac{i n\pi}{4K} \int_{\rho}^{K'-v_0} e^{-n\pi r/2K} [\Omega^2(z_1 + e^{i\pi/2}r) - \Omega^2(z_1 + e^{-i3\pi/2}r)] dr \right. \\ \left. + \frac{\Pi}{K} \int_{\rho}^{K'-v_0} e^{-n\pi r/2K} [\Omega(z_1 + e^{i\pi/2}r) - \Omega(z_1 + e^{-i3\pi/2}r)] dr \right. \\ \left. + \int_{-3\pi/2}^{\pi/2} \frac{e^{in\pi\rho(\cos\theta + i\sin\theta)/2K} \Omega(z_1 + \rho e^{i\theta}) \rho e^{i\theta} d\theta}{1 - \alpha^2 sn^2(z_1 + \rho e^{i\theta})} \right\}.$$

The sort of analysis indicated above to determine  $\Pi(am2iK')$  also suffices to determine  $\Omega(z)$  on each side of the cut. We find

$$\Omega(z_1 + e^{i\pi/2}r, \alpha^2, k) = -i\lambda w(v_0 + r)/K' \\ + i\beta^2 \Omega(v_0 + r, \alpha'^2, k') + \lambda\pi/2, \\ (3.11) \quad \Omega(z_1 + e^{-i3\pi/2}r, \alpha^2, k) = -i\lambda w(v_0 + r)/K' \\ + i\beta^2 \Omega(v_0 + r, \alpha'^2, k') - \lambda\pi/2, \quad (0 < r < K' - v_0),$$

where, as for  $\Pi'$ ,  $\Omega(v_0 + r, \alpha'^2, k')$  is to be interpreted as the Cauchy principal value of the defining integral. Use of (3.11) enables us to write the half-loop integral as

$$I(z_1; n) = i \left\{ \left[ \frac{2\lambda^2 wK}{nK'} - \frac{2\lambda\Pi}{n} \right] e^{-n\pi K'/2K} \right. \\ \left. + \left[ -\frac{\pi\lambda^2 w(v_0 + \rho)}{K'} - \frac{2\lambda^2 wK}{nK'} + \frac{2\lambda\Pi}{n} \right] e^{-n\pi(v_0 + \rho)/2K} \right. \\ (3.12) \quad \left. + \frac{i\lambda\pi}{2} [\Omega(z_1 + \rho e^{i\pi/2}) + \Omega(z_1 + \rho e^{-i3\pi/2})] e^{-n\pi(v_0 + \rho)/2K} \right. \\ \left. + e^{-n\pi v_0/2K} \int_{-3\pi/2}^{\pi/2} \frac{\exp(in\pi\rho e^{i\theta}/2K) \Omega(z_1 + \rho e^{i\theta}) \rho e^{i\theta} d\theta}{1 - \alpha^2 sn^2(z_1 + \rho e^{i\theta})} \right. \\ \left. + \frac{n\pi^2 \lambda \beta^2}{2K} \int_{\rho+v_0}^{K'} e^{-n\pi s/2K} \Omega(s, \alpha'^2, k') ds \right\}.$$

For small  $\rho$ ,

$$\begin{aligned}
 & \frac{i\lambda\pi}{2} e^{-n\pi\rho/2K} \Omega(z_1 + \rho e^{i\pi/2}) + \int_{-\pi/2}^{\pi/2} \frac{\exp(in\pi\rho e^{i\theta}/2K) \Omega(z_1 + \rho e^{i\theta}) \rho e^{i\theta} d\theta}{1 - \alpha^2 sn^2(z_1 + \rho e^{i\theta})} \\
 &= \frac{i\lambda\pi}{2} e^{-n\pi\rho/2K} \Omega(z_1 + \rho e^{i\pi/2}) \\
 &+ \int_{-\pi/2}^{\pi/2} \exp(in\pi\rho e^{i\theta}/2K) \Omega(z_1 + \rho e^{i\theta}) \rho \\
 &\cdot e^{i\theta} \left[ -\frac{i\lambda}{2} \cdot \frac{1}{\rho e^{i\theta}} + O(1) \right] d\theta, \\
 &= -\frac{i\lambda}{2} \int_{-\pi/2}^{\pi/2} [\exp(in\pi\rho e^{i\theta}/2K) \Omega(z_1 + \rho e^{i\theta}) \\
 &\quad - \exp(-n\pi\rho/2K) \Omega(z_1 + \rho e^{i\pi/2})] d\theta + o(1), \\
 &= -\frac{\lambda^2}{2} \int_{-\pi/2}^{\pi/2} [\ln(\rho e^{i\theta}) - \ln(\rho e^{i\pi/2})] d\theta + o(1) = i\lambda^2 \pi^2/8 + o(1).
 \end{aligned}$$

The pair of terms in (3.12) connected with the angle  $-3\pi/2$  can be treated accordingly and their sum is found to be equal to  $-i\lambda^2 \pi^2/8 + o(1)$ . Consequently, sending  $\rho \rightarrow 0$  gives

$$\begin{aligned}
 (3.13) \quad I(z_1; n) = i \left\{ \left[ \frac{2\lambda^2 wK}{nK'} - \frac{2\lambda\Pi}{n} \right] e^{-n\pi K'/2K} + \left[ -\frac{\pi\lambda^2 wv_0}{K'} - \frac{2\lambda^2 wK}{nK'} + \frac{2\lambda\Pi}{n} \right] e^{-n\pi v_0/2K} \right. \\
 \left. + \frac{n\pi^2 \lambda \beta^2}{2K} \int_{v_0}^{K'} e^{-n\pi s/2K} \Omega(s, \alpha'^2, k') ds \right\}.
 \end{aligned}$$

The upper half-loop integral (around  $z_4$ ),  $I(z_4; n)$ , becomes simply expressed in terms of  $I(z_1; n)$  under the transformation  $u = 2iK' - v$ . In fact, using (3.4) and the single-valuedness of  $sn v$  results in

$$\begin{aligned}
 I(z_4; n) &= \int_{C_4} \frac{e^{in\pi u/2K} \Omega(u) du}{1 - \alpha^2 sn^2 u} \\
 &= e^{-n\pi K'/K} I(z_1; -n) + 2i\lambda w e^{-n\pi K'/K} \oint_{|v-z_1|=\rho} \frac{e^{-in\pi v/2K} dv}{1 - \alpha^2 sn^2 v}.
 \end{aligned}$$

The last integral on the right is easily computed by residues.

Similarly, the closed-loop integral around  $z_2, z_3$  is transformed by  $u = 2K + 2iK' - v$  into a simple expression involving the closed-loop integral around  $z_1, z_4$ .

#### REFERENCES

- [1] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Physicists*, Springer, Berlin, 1954.
- [2] Y. HAGIHARA, *Celestial Mechanics*, M.I.T. Press, Cambridge, MA, 1970.
- [3] C. G. JACOBI, *Fundamenta Nova Theoriae Functionum Ellipticarum*, Königsberg, 1829.
- [4] Z. KOPAL, *The Roche model and its applications to close binary systems*, Advances in Astronomy and Astrophysics, Vol. 9, Z. Kopal, ed., Academic Press, New York, 1972, pp. 1-64.
- [5] R. G. LANGEBARTEL, *Two-center problem orbits as intermediate orbits for the restricted three-body problem*, NASA TN D-2939, National Aeronautics and Space Administration, Washington, DC, 1965, pp. 1-19.



## LIMIT ANALYSIS FOR PLASTIC PLATES\*

EDMUND CHRISTIANSEN†

**Abstract.** The collapse problem for plate bending is considered as an infinite dimensional mathematical programming problem. The duality between the static and kinematic formulations of limit analysis is proved, and it is shown that limit fields for bending moments and displacement rates exist. Finally we analyze the approximation of the continuous problem by finite-dimensional convex programming problems using the finite element method.

**1. Introduction.** In [4] the collapse problem for a 3-dimensional plastic continuum is analyzed. Here we shall attack the considerably modified problem, which arises when the plate bending approximation is applicable (see [6]).

For a solid with volume  $V$  the notation is:

$$\mathbf{u} = \frac{d}{dt} \mathbf{v} \quad \text{the displacement rate vector;}$$

$$\varepsilon_{ij} = \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad \text{the symmetric } 3 \times 3 \text{ strain tensor;}$$

$$\boldsymbol{\sigma} = (\sigma_{ij}) \quad \text{the symmetric } 3 \times 3 \text{ stress tensor.}$$

The internal work rate for the pair  $(\boldsymbol{\sigma}, \mathbf{u})$  is then given by (see [4])

$$a(\boldsymbol{\sigma}, \mathbf{u}) = \sum_{i,j} \int_V \sigma_{ij} \frac{\partial u_j}{\partial x_i} dv.$$

Now consider the case where the solid is a plate occupying the area  $\Omega$  in the  $x_1 - x_2$  plane. Let  $u = u_3$  be the transversal displacement rate. Then at distance  $x_3$ , measured with sign, from the mid-plane of the plate we have ([6]):

$$u_i = -x_3 \frac{\partial u}{\partial x_i}, \quad i = 1, 2$$

and hence

$$(1.1) \quad \frac{d}{dt} \varepsilon_{ij} = -x_3 \frac{\partial^2 u}{\partial x_i \partial x_j}, \quad i, j = 1, 2.$$

All other components of  $\varepsilon$  vanish in the plate bending approximation.

The internal work rate may now be written

$$(1.2) \quad \begin{aligned} a(\boldsymbol{\sigma}, \mathbf{u}) &= - \sum_{i,j=1}^2 \int_V \sigma_{ij} x_3 \frac{\partial^2 u}{\partial x_i \partial x_j} dv \\ &= - \sum_{i,j} \int_{\Omega} m_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} dA \equiv b(\mathbf{m}, u), \end{aligned}$$

where

$$(1.3) \quad m_{ij} = \int_{-(1/2)H}^{(1/2)H} x_3 \sigma_{ij} dx_3, \quad i, j = 1, 2$$

\* Received by the editors January 30, 1979, and in revised form July 10, 1979.

† Department of Mathematics, University of Odense, DK-5230, Odense M, Denmark.

are the bending moments.  $H$  is the thickness of the plate and may depend on  $(x_1, x_2)$ .

Let  $f = f(x_1, x_2)$  be the transversal external load and define for convenience

$$(1.4) \quad F(u) = \int_{\Omega} f \cdot u \, dA.$$

Finally the set  $B$  of admissible bending moments is determined by the admissible stresses for the material and (1.3). The 2 most important yield conditions, the Mises and the Tresca conditions, are for the case of plate bending given by ([6]),

$$B_M = \{\mathbf{m} | m_{xx}^2 - m_{xx}m_{yy} + m_{yy}^2 + 3m_{xy}^2 \leq M_0\},$$

$$B_T = \{\mathbf{m} | \max(|m_1|, |m_2|, |m_1 - m_2|) \leq M_0\},$$

$m_1$  and  $m_2$  being the principal moments.

The problem of limit analysis is to find the maximal (or rather limit) multiple  $\lambda$  of  $f$ , which the plate can carry without collapsing. This problem may now be stated formally as in [4].

*Static formulation.*

$$(1.5) \quad \lambda = \sup \{ \lambda | \exists \mathbf{m} \in B : B(\mathbf{m}, u) = \lambda F(u) \forall u \}$$

$$= \sup_{\mathbf{m} \in B} \inf_{F(u)=1} b(\mathbf{m}, u).$$

The interpretation is, that  $b(\mathbf{m}, u) = F(u)$  is the weak form of the equilibrium equations for  $\mathbf{m}$  with the load  $f$ .

*Kinematic formulation* (the dual problem).

$$(1.6) \quad \lambda = \inf_{F(u)=1} \sup_{\mathbf{m} \in B} b(\mathbf{m}, u)$$

$$= \inf_{F(u)=1} D(u),$$

where

$$(1.7) \quad D(u) = \sup_{\mathbf{m} \in B} b(\mathbf{m}, u)$$

is the energy dissipation rate associated with  $u$ . Also (1.6) has a natural physical interpretation (see [4]).

Our present aim is to formalize this approach choosing adequate spaces for  $\mathbf{m}$ ,  $u$  and  $f$ , and to prove the duality theorem: (1.5) and (1.6) give the same collapse multiplier  $\lambda$ , and limit fields for  $\mathbf{m}$  and  $u$  exist. This generalizes the analysis in [2] and [3] to the continuous case and falls in line with several recent contributions in limit analysis among which [5] and [9] should be mentioned.

**2. Prerequisites.** The following theorem, which is proved in [4] is essential in our approach.

**THEOREM 2.1.**  *$X$  and  $Y$  are normed real vector spaces with a continuous bilinear form  $a(\cdot, \cdot)$  on  $X \times Y$ . Let  $B \subset X$  and  $C \subset Y$  be convex sets such that  $B$  has nonempty interior, and  $C$  is closed. Assume the following "reflexivity" condition*

(i) *If  $y^* \in Y^*$  satisfies  $\sup (y^*(y) | y \in C) < \infty$ , then there exists  $x_0 \in X$  such that  $y^*(y) = a(x_0, y) \forall y \in Y$ .*

(ii) *If  $x^* \in X^*$  satisfies (a):  $\inf (x^*(x) | x \in B) > -\infty$ , and (b): (if  $a(x, y) = 0 \forall y \in Y$ , then  $x^*(x) = 0$ ) then there exists  $y_0 \in Y$  such that  $x^*(x) = a(x, y_0) \forall x \in X$ .*

*Under these conditions*

$$(2.1) \quad \inf_{x \in B} \sup_{y \in C} a(x, y) = \max_{y \in C} \inf_{x \in B} a(x, y)$$

*provided the left-hand side is finite.*

For  $\Omega \subset \mathbb{R}^n$  let  $W^{m,p}(\Omega)$  denote the standard Sobolev spaces of functions with generalized derivatives up to order  $m$  in  $L^p$  for  $m \geq 0$  integer. For  $m \leq 0$ ,  $W^{m,p}$  is the dual of  $W_0^{-m,q}$  where  $1/p + 1/q = 1$ . A quick review can be found in for example [1].

We need the following theorem which is basically proved in [7].

**THEOREM 2.2.** *Let  $\Omega \subseteq \mathbb{R}^n$  be a bounded domain with smooth boundary, and let  $p > 1$ . For every pair*

$$f \in W^{-1,p}(\Omega) \quad \text{and} \quad g \in W^{1-1/p,p}(\partial\Omega)$$

*there is a solution  $u \in W^{1,p}(\Omega)$  to the problem*

$$\begin{aligned} \Delta u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned}$$

*Proof.* There exist  $f_i \in L^p(\Omega)$ ,  $i = 1, \dots, n$ , such that

$$f = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}.$$

By Theorem 4.2 in [7] we may find  $\phi_i \in W^{2,p}(\Omega)$  such that

$$\Delta \phi_i = f_i \quad \text{in } \Omega.$$

Let

$$u_1 = \sum_{i=1}^n \frac{\partial \phi_i}{\partial x_i} \in W^{1,p}(\Omega).$$

Then

$$\Delta u_1 = \sum_{i=1}^n \frac{\partial}{\partial x_i} f_i = f \quad \text{in } \Omega$$

and  $u_1$  has a trace in  $W^{1-(1/p),p}(\partial\Omega)$ . Using again Theorem 4.2 in [7] we may solve

$$\begin{aligned} \Delta u_2 &= 0 \quad \text{in } \Omega, \\ u_2 &= g - u_1 \quad \text{on } \partial\Omega \end{aligned}$$

with  $u_2 \in W^{1,p}(\Omega)$ . Now  $u_1 + u_2$  is the solution to the original problem. Q.E.D.

*Remark.* In applications in solid mechanics  $\Omega$  is frequently a domain with Lipschitz continuous boundary satisfying the cone property as for example a cube. In [8] the Dirichlet problem is proved to have solutions in such domains, but in less generality than in Theorem 2.2. We believe this difficulty to be purely technical, however, and shall not hesitate to apply the results of this paper to such domains.

**3. The theorem of limit analysis.** Let  $\Omega \subseteq \mathbb{R}^2$  be the domain of a plate with regular boundary (see remark following Theorem 2.2). The plate is fixed along its boundary (and only there)

$$(3.1) \quad u = 0 \quad \text{on } \partial\Omega.$$

The plate may be “clamped” along part of its boundary,

$$(3.2) \quad \frac{\partial u}{\partial n} = 0 \quad \text{on } S \subseteq \partial\Omega.$$

If  $f$  is the load distribution per unit area of the plate, the classical form of the equilibrium equation for the bending moment tensor  $\mathbf{m}$  is

$$(3.3a) \quad -\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} m_{ij} = f \quad \text{in } \Omega$$

or

$$(3.3b) \quad -\nabla \cdot (\nabla \cdot \mathbf{m}) = f \quad \text{in } \Omega$$

with the boundary condition

$$(3.4) \quad m_n \equiv \mathbf{n} \cdot \mathbf{m} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \setminus S.$$

Assuming sufficient regularity the internal work rate (1.2) may be written as

$$(3.5a) \quad \begin{aligned} b(\mathbf{m}, u) &= -\sum_{i,j} \int_{\Omega} m_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} dA \\ &= -\sum_{j=1}^2 \int_{\Omega} \mathbf{m}_j \cdot \nabla \left( \frac{\partial u}{\partial x_j} \right) dA \\ &= \sum_j \int_{\Omega} (\nabla \cdot \mathbf{m}_j) \frac{\partial u}{\partial x_j} dA - \sum_j \int_{\partial\Omega} (\mathbf{n} \cdot \mathbf{m}_j) \frac{\partial u}{\partial x_j} ds \\ &= \int_{\Omega} (\nabla \cdot \mathbf{m}) \cdot \nabla u dA - \int_{\partial\Omega} (\mathbf{n} \cdot \mathbf{m}) \cdot \nabla u ds, \end{aligned}$$

where  $\mathbf{m}_j$  is the  $j$ th column of  $\mathbf{m}$ .

Using that  $\nabla u = (\partial u / \partial n)\mathbf{n} + (\partial u / \partial \tau)\boldsymbol{\tau}$ , where  $\mathbf{n}$  and  $\boldsymbol{\tau}$  denote the normal resp. the tangent of  $\partial\Omega$ , and that (3.2) and (3.4) hold we get

$$(3.5b) \quad \begin{aligned} b(\mathbf{m}, u) &= \int_{\Omega} (\nabla \cdot \mathbf{m}) \cdot \nabla u dA - \int_{\partial\Omega} (\mathbf{n} \cdot \mathbf{m}) \cdot \boldsymbol{\tau} \frac{\partial u}{\partial \tau} ds \\ &= \int_{\Omega} (\nabla \cdot \mathbf{m}) \cdot \nabla u dA + \int_{\partial\Omega} \frac{\partial}{\partial \tau} (\mathbf{n} \cdot \mathbf{m} \cdot \boldsymbol{\tau}) u ds \\ (3.5c) \quad &= -\int_{\Omega} \nabla \cdot (\nabla \cdot \mathbf{m}) u dA + \int_{\partial\Omega} (\mathbf{n} \cdot (\nabla \cdot \mathbf{m}) + \frac{\partial}{\partial \tau} (\mathbf{n} \cdot \mathbf{m} \cdot \boldsymbol{\tau})) u ds \\ &= -\int_{\Omega} \nabla \cdot (\nabla \cdot \mathbf{m}) u dA. \end{aligned}$$

*Remarks.* The equality between (3.5a) and (3.5b) should be imposed as the natural boundary condition (3.2). The equalities (3.5) are also regularity conditions on the pair  $(\mathbf{m}, u)$ . The boundary conditions (3.1) and (3.4) should be imposed as essential boundary conditions.

As usual the equilibrium equation (3.3) must be interpreted in a weak sense. From (3.2), (3.4) and (3.5) we get the weak form

$$(3.6) \quad b(\mathbf{m}, u) = F(u) \quad \text{for all } u,$$

where we have used the convenient notation

$$(3.7) \quad F(u) = \int_{\Omega} fu \, dA;$$

(3.6) must hold for all  $u$  satisfying (3.1).

We are now ready to define the admissible spaces for  $\mathbf{m}$  and  $u$ .

Let  $p > 2$ , such that  $W^{1,p}(\Omega)$  is continuously imbedded in  $C(\bar{\Omega})$ , and let  $1/p + 1/q = 1$ .

$$(3.8a) \quad X = \{\mathbf{m} = (m_{ij}) \mid m_{ij} \in W^{1,p}(\Omega), m_{ij} = m_{ji}, m_n = 0 \text{ on } \partial\Omega \setminus S\},$$

where  $S$  is the clamped part of the boundary, and  $m_n$  is given by (3.4).

$$Y = \left\{ u \in W_0^{1,q}(\Omega) \mid \text{there exist } \mu_{ij} \in C^*(\bar{\Omega}) \text{ such that} \right.$$

$$(3.8b) \quad \mu_{ij} = \mu_{ji} \text{ and for all } \mathbf{m} \in X:$$

$$\sum_{i,j} \langle \mu_{ij}, m_{ij} \rangle_{C^* \times C} = \langle \nabla \cdot (\nabla \cdot \mathbf{m}), u \rangle_{W^{-1,p} \times W_0^{1,q}}.$$

Equation (3.2) is imposed as a natural boundary condition by the definition of  $Y$ .

We now interpret the integrals in (3.5) as formal dualities and define the bilinear form  $b(\cdot, \cdot)$  on  $X \times Y$  by (3.5a) or (3.5c).

LEMMA 3.1. *For every  $f \in W^{-1,p}(\Omega)$  there exists  $\mathbf{m} \in X$ , such that*

$$-\nabla \cdot (\nabla \cdot \mathbf{m}) = f \quad \text{in } \Omega.$$

*Proof.* By Theorem 2.2 we may find  $\phi \in W^{1,p}(\Omega)$  such that

$$-\Delta\phi = f \quad \text{in } \Omega,$$

$$\phi = 0 \quad \text{on } \partial\Omega.$$

Let  $u_1, u_2 \in C^3(\bar{\Omega})$  such that  $u_i$  is constant on  $\partial\Omega$ . Define

$$m_{11} = \frac{\partial u_2}{\partial x_2} + \phi, \quad m_{22} = \frac{\partial u_1}{\partial x_1} + \phi, \quad m_{12} = -\frac{1}{2} \left( \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right).$$

Then

$$\nabla \cdot (\nabla \cdot \mathbf{m}) = \frac{\partial^2 m_{11}}{\partial x_1^2} + 2 \frac{\partial^2 m_{12}}{\partial x_1 \partial x_2} + \frac{\partial^2 m_{22}}{\partial x_2^2} = \Delta\phi = -f$$

and

$$\begin{aligned} m_n &= n_1^2 m_{11} + 2n_1 n_2 m_{12} + n_2^2 m_{22} \\ &= \phi + n_1 \left( n_1 \frac{\partial u_2}{\partial x_2} - n_2 \frac{\partial u_2}{\partial x_1} \right) + n_2 \left( n_2 \frac{\partial u_1}{\partial x_1} - n_1 \frac{\partial u_1}{\partial x_2} \right) \\ &= \phi + n_1 \frac{\partial u_2}{\partial \tau} - n_2 \frac{\partial u_1}{\partial \tau} = 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where  $\partial/\partial\tau$  denotes the tangential derivative. Q.E.D.

Lemma 3.1 implies that for any load  $f \in W^{-1,p}(\Omega)$  there is  $\mathbf{N} \in X$ , such that

$$(3.9) \quad b(\mathbf{N}, u) = F(u) \equiv \langle f, u \rangle_{W^{-1,p} \times W_0^{1,q}} \quad \forall u \in Y,$$

i.e. there is  $\mathbf{N} \in X$ , which is in equilibrium with  $f$ . Otherwise the collapse multiplier for  $f$  obviously would be zero.

**THEOREM 3.2.** *Let  $X, Y$  and  $b(\cdot, \cdot)$  be as above, let  $f \in W^{-1,p}(\Omega)$  and let  $\mathbf{N} \in X$  satisfy (3.9). Let  $B \subseteq X$  be convex, such that with respect to the maximum-norm on  $X$ ,  $B$  is bounded, closed and contains 0 in its interior. Finally let*

$$C = \{u \in Y \mid F(u) \equiv \langle f, u \rangle_{W^{-1,p} \times W_0^{1,q}} = 1\}.$$

*Then the following problems all give the same value:*

$$\begin{aligned} & \text{(a)} \\ \lambda & \equiv \inf_{u \in C} \sup_{\mathbf{m} \in B} b(\mathbf{m}, u) = \sup_{\mathbf{m} \in B} \inf_{u \in C} b(\mathbf{m}, u) \\ & \quad \parallel \text{(b)} \qquad \qquad \parallel \text{(c)} \\ \text{(D): } & \min_{\substack{u \in Y \\ F(u)=1}} D(u) \qquad \sup_{\substack{\mathbf{m} \in B \\ b(\mathbf{m}, u) = \lambda F(u) \quad \forall u \in Y}} \lambda \\ & \qquad \qquad \qquad \parallel \text{(d)} \\ \text{(P): } & \max_{\substack{\mathbf{m} \in \tilde{B} \\ b(\mathbf{m}, u) = \lambda F(u) \quad \forall u \in Y}} \lambda \end{aligned}$$

where

$$D(u) = \sup_{\mathbf{m} \in B} b(\mathbf{m}, u)$$

and

$$\begin{aligned} \tilde{B} &= \{\mathbf{m} \in L^\infty(\Omega) \mid \exists \{\mathbf{m}^{(k)}\} \subseteq B : \mathbf{m}^{(k)} \rightharpoonup \mathbf{m} \text{ weak-}^* \text{ in } L^\infty \\ & \text{and } b(\mathbf{m}^{(k)}, u) = \lambda_k F(u) \forall u \in Y \text{ for some converging sequence } \{\lambda_k\} \text{ in } \mathbb{R}.\} \end{aligned}$$

Clearly  $b(\mathbf{m}, u)$  extends to  $\tilde{B} \times Y$  by the limit of the sequence  $\{\lambda_k\}$ .

*Proof.* (a) and (b): We apply Theorem 2.1 to  $X, Y, -b(\mathbf{m}, u), B$  and  $C$ .

ad(i) If  $y^* \in Y^*$  is bounded on the affine hyperspace  $C$ , then  $y^*(u) = \lambda F(u) = \lambda b(\mathbf{N}, u), \forall u \in Y$  for some  $\lambda$ .

ad(ii) Let  $x^* \in X^*$  be bounded (from below) on  $B$ . Since  $B$  contains a ball for the  $L^\infty$ -norm there exist measures  $\mu_{ij} \in C^*(\bar{\Omega})$ , such that  $\mu_{ij} = \mu_{ji}$  and

$$(3.10) \qquad x^*(\mathbf{m}) = \sum_{i,j} \langle \mu_{ij}, m_{ij} \rangle_{C^* \times C}.$$

We may additionally assume that  $x^*$  satisfies:

$$(3.11) \qquad b(\mathbf{m}, u) = 0 \quad \forall u \in Y \Rightarrow x^*(\mathbf{m}) = 0.$$

By Lemma 3.1 the map

$$T : \mathbf{m} \rightarrow \nabla \cdot (\nabla \cdot \mathbf{m})$$

maps  $X$  onto  $W^{-1,p}(\Omega)$  and is thus open. By (3.11) we have

$$\text{kernel}(T) \subseteq \text{kernel}(x^*)$$

and hence  $x^*$  may be factorized over the kernel of  $T$ :

$$x^*(\mathbf{m}) = u(T(\mathbf{m})) \quad \forall \mathbf{m} \in X$$

for some  $u \in (W^{-1,p}(\Omega))^* = W_0^{1,q}(\Omega)$ .

Comparing with (3.10) we see that  $u \in Y$ . Hence (ii) is satisfied, and by Theorem 2.1 (a) and (b) are proved.

(c) Since  $C$  is an affine hyperplane the inner infimum equals  $-\infty$ , unless  $b(\mathbf{m}, u) = \lambda F(u) \forall u \in Y$  for some real  $\lambda$ . In this case  $b(\mathbf{m}, u) = \lambda$  for  $u \in C$ .

(d) Let  $b(\mathbf{m}^{(k)}, u) = \lambda_k F(u) = \lambda_k b(\mathbf{N}, u)$  for all  $u \in Y$ , and assume that  $\lambda_k$  converges to the supremum  $\lambda$ . By weak-\* compactness of  $B$  in  $L^\infty$  there is  $\mathbf{m}^0 \in \tilde{B}$  such that

$$b(\mathbf{m}^0, u) = \lambda F(u) \quad \forall u \in Y. \quad \text{Q.E.D.}$$

It is now easy to see that the limit fields are in fact a saddle point for  $b$  on  $B \times C$ .

**THEOREM 3.3.** *If  $(\mathbf{m}^0, u^0)$  solve (P) and (D) in Theorem 3.2, then*

$$(3.12) \quad b(\mathbf{m}, u^0) \leq \lambda = b(\mathbf{m}^0, u^0) = b(\mathbf{m}^0, u) \quad \forall \mathbf{m} \in B, \quad \forall u \in C.$$

*Proof.* For  $\mathbf{m} \in B$  and  $u \in C$  we have

$$b(\mathbf{m}, u^0) \leq D(u^0) = \lambda = \lambda F(u^0) = b(\mathbf{m}^0, u). \quad \text{Q.E.D.}$$

**4. Approximate solution.** It is now clear how the solutions to the primal and dual problems of limit analysis can be approximated: Replace  $X$  and  $Y$  by finite dimensional subspaces  $X_h$  and  $Y_h$  and solve the finite dimensional min-max problem. We always choose  $X_h$  and  $Y_h$  such that the equalities (3.5) hold in classical sense, so that the bilinear form may be computed by the symmetric expression

$$b(\mathbf{m}_h, u_h) = \int_{\Omega} (\nabla \cdot \mathbf{m}_h) \cdot \nabla u_h \, dA \quad \forall (\mathbf{m}_h, u_h) \in X_h \times Y_h.$$

This way the mixed approximation can be based on subspaces of functions of less regularity than the primal or dual problem. Another advantage of the mixed method is that bending moments and displacements in the collapse state are approximated simultaneously.

Let

$$X_h \subseteq X, \quad Y_h \subseteq Y, \quad B_h \subseteq X_h \cap B,$$

and define

$$(4.1) \quad b_h(\mathbf{m}_h, u_h) = b(\mathbf{m}_h, u_h) \quad \forall (\mathbf{m}_h, u_h) \in X_h \times Y_h,$$

$$(4.2) \quad F_h(u_h) = F(u_h) \quad \forall u_h \in Y_h.$$

We identify  $X_h$  and  $Y_h$  with copies of  $R^m$  and  $R^n$  respectively through a fixed choice of bases and identify each space with its dual in canonical way. Then there is a linear map, i.e., a matrix  $A: X_h \rightarrow Y_h$ , such that for all  $(\mathbf{m}_h, u_h) \in X_h \times Y_h$

$$(4.3) \quad b(\mathbf{m}_h, u_h) = \langle A\mathbf{m}_h, u_h \rangle_n = \langle \mathbf{m}_h, A^t u_h \rangle_m,$$

where  $\langle \cdot, \cdot \rangle_n$  denotes the Euclidean inner product on  $R^n$ , and  $A^t$  is the transposed of  $A$ . Also  $F_h$  may be identified with an element of  $Y_h$ ,

$$(4.4) \quad F(u_h) = \langle F_h, u_h \rangle_n \quad \forall u_h \in Y_h.$$

We now have the discrete analogue of Theorem 3.2.

**THEOREM 4.1.** *Notation as above. Assume  $F_h \in A(X_h)$  and let  $B_h \subseteq X_h$  be convex*

and compact with zero in its interior. Then

$$0 < \lambda_h \equiv \min_{F_h(u_h)=1} \max_{\mathbf{m}_h \in B_h} \langle \mathbf{m}_h, A^t u_h \rangle = \max_{\mathbf{m}_h \in B_h} \min_{F_h(u_h)=1} \langle A \mathbf{m}_h, u_h \rangle$$

$$\| \qquad \qquad \qquad \|$$

$$(D_h): \quad \min_{F_h(u_h)=1} D_h(u_h) \qquad (P_h): \quad \max_{\substack{\mathbf{m}_h \in B_h \\ A \mathbf{m}_h = \lambda F_h}} \lambda$$

where

$$(4.5) \qquad D_h(u_h) = \max_{\mathbf{m}_h \in B_h} \langle \mathbf{m}_h, A^t u_h \rangle.$$

This theorem can be proved using classical convex programming results or by repeating the arguments (now simplified) from Theorem 3.2. Since it is a special case of Theorem 5.1 in [4] we shall omit the details here.

*Remark.* The condition  $F_h \in A(X_h)$  is the discrete analogue of (3.9). It is easy to see, that if  $F_h \notin A(X_h)$  then the duality in Theorem 4.1 still holds, but  $\lambda_h = 0$ . Of course we may in that case replace  $F_h$  by its orthogonal projection on  $A(X_h)$ , but simple examples show that this process is inconsistent so that  $\lambda_h$  diverges. In order to allow general forces we shall always require

$$(4.6) \qquad A(X_h) = Y_h.$$

**COROLLARY 4.2.** *If  $\mathbf{m}_h^0$  and  $u_h^0$  solve  $(P_h)$ , respectively  $(D_h)$ , then  $(\mathbf{m}_h^0, u_h^0)$  is a saddle point for  $b_h$ : For all  $\mathbf{m}_h \in B_h$  and  $u_h \in C_h$*

$$(4.7) \qquad b_h(\mathbf{m}_h, u_h^0) \leq \lambda_h = b_h(\mathbf{m}_h^0, u_h^0) = b_h(\mathbf{m}_h^0, u_h).$$

The proof is identical to the proof of Theorem 3.3.

**THEOREM 4.3.** *Let  $(\lambda, \mathbf{m}^0, u^0)$  and  $(\lambda_h, \mathbf{m}_h^0, u_h^0)$  solve the continuous and discrete problem respectively. Then we have for all  $\mathbf{m}_h \in B_h$  and  $u_h \in C_h$ ;*

$$(4.8) \qquad b(\mathbf{m}_h - \mathbf{m}^0, u_h^0) \leq \lambda_h - \lambda \leq b(\mathbf{m}_h^0, u_h - u^0).$$

*Proof.* Subtract (3.12) from (4.7) and put  $u = u_h^0$  and  $\mathbf{m} = \mathbf{m}^0$ . Q.E.D.

Theorem 4.3 is the main inequality for convergence results for  $\lambda$ . Let  $\|\mathbf{m}\|_1$  and  $\|u\|_2$  be norms on  $X$  and  $Y$  respectively such that

$$(4.9) \qquad |b(\mathbf{m}, u)| \leq C \|\mathbf{m}\|_1 \|u\|_2.$$

There are several useful choices for these norms.

Assume the *stability conditions*:

$$(4.10) \qquad \|u_h^0\|_2 \leq \text{constant}, \quad \|\mathbf{m}_h^0\|_1 \leq \text{constant},$$

where the constants are independent of  $h$ , and the *consistency conditions*

$$(4.11a) \qquad \min_{\mathbf{m}_h \in B_h} \|\mathbf{m}_h - \mathbf{m}^0\|_1 \rightarrow 0 \quad \text{as } h \rightarrow 0,$$

$$(4.11b) \qquad \min_{u_h \in C_h} \|u_h - u^0\|_2 \rightarrow 0 \quad \text{as } h \rightarrow 0;$$

then Theorem 4.3 immediately implies the following convergence results.

**COROLLARY 4.4.** *If we can find norms such that (4.9), (4.10) and (4.11) hold, then  $\lambda_h \rightarrow \lambda$ .*



The important question of convergence rate requires analysis of the approximation used.

The stability condition (4.10) depends on the domain as well as the forces. No general a priori results in this direction are known to the author, but in specific cases the situation is usually better.

The consistency condition depends on the regularity of the solutions  $\mathbf{m}^0$  and  $u^0$ . For approximation by the finite element method the question of appropriate norms and corresponding convergence rates is well analyzed, once the regularity of the solutions is known. However the regularity problem for saddle point problems does not seem to be solved in any generality yet.

In a forthcoming publication we shall apply the method developed here in combination with the finite element method to a classical problem of plate bending.

#### REFERENCES

- [1] M. BERGER, *Nonlinearity and functional analysis*. Academic Press, New York, 1977.
- [2] A. CHARNES AND H. J. GREENBERG, *Plastic collapse and linear programming* (abstract), Bull. Amer. Math. Soc., 57 (1951), p. 480.
- [3] A. CHARNES, C. E. LEMKE AND O. C. ZIENKIEWICS, *Virtual work, linear programming and plastic limit analysis*, Proc. Roy. Soc. London Ser. A, 251 (1959), pp. 110–116.
- [4] E. CHRISTIANSEN, *Limit analysis in plasticity as a mathematical programming problem*, Calcolo, to appear.
- [5] E. CHRISTIANSEN, H. MATTHIES AND G. STRANG, *The saddle point of a differential program*, Energy Methods in Finite Element Analysis, R. Glowinski, E. Y. Rodin, O. C. Zienkiewics, eds., John Wiley, New York, 1979.
- [6] P. G. HODGE, *Plastic Analysis of Structures*. McGraw-Hill, New York, 1959.
- [7] J. L. LIONS AND E. MAGENES, *Problemi ai limiti non omogenei (V)*, Ann. Scuola Norm. Sup. Pisa Sci. Fiz. Mat., 16 (1962), pp. 1–44.
- [8] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.
- [9] G. STRANG AND R. TEMAM, *Functions of bounded deformation*, Arch. Rational Mech. Anal., to appear.

## SUMMATION THEOREMS FOR HYPERGEOMETRIC SERIES IN $U(n)^*$

WAYNE J. HOLMAN III†

**Abstract.** Explicit expressions for the matrix elements of multiplicity-free Wigner and Racah coefficients in  $U(n)$  are used to establish multidimensional analogs of known hypergeometric summation theorems. Infinite sequences of such analogs are found for the Gauss theorem, Saalschütz' theorem, and the summation theorem for well-poised  ${}_5F_4(1)$ . For the sake of completeness, results published earlier are also presented: an infinite sequence of summation theorems analogous to that for well-poised  ${}_4F_3(-1)$  and a single analog to Whipple's theorem which arises from the construction of tensor operators in  $U(3)$ .

In a previous publication [4] Whipple's notion of a well-poised hypergeometric series [6] was generalized to multidimensional cases. The basis for the generalization was analogy of structural function in the representation theory of  $U(n)$ . For example, the summation theorem for well-poised  ${}_4F_3(-1)$  is realized in the representation theory of  $U(2)$ ; it is expressed by the orthonormality relation for degenerate Wigner coefficients in  $U(2)$ . Hence we can form the analogous relations for degenerate multiplicity-free Wigner coefficients in  $U(n)$  and interpret the resulting expressions as a sequence of summation theorems for multidimensional series (in the cases  $n > 2$ ) which form analogs to the summation theorem for well-poised  ${}_4F_3(-1)$  which is realized in the case  $n = 2$ .

This method can be extended further, and it is the purpose of the present note to state multidimensional analogs for the Gauss summation theorem for  ${}_2F_1(1)$ , Saalschütz' theorem for certain cases of  ${}_3F_2(1)$ , and the summation theorem for well-poised  ${}_5F_4(1)$ . These analogs can all be obtained from known relationships between those matrix elements of tensor or Racah operators in  $U(n)$  which have been explicitly constructed. For the sake of completeness we shall state the summation theorems presented in [4] as well: a sequence of analogs of the summation theorem for well-poised  ${}_4F_3(-1)$  and a single analog of Whipple's theorem [7], which relates well-poised  ${}_7F_6(1)$  and Saalschützian  ${}_4F_3(1)$ , from the representation theory of  $U(3)$ .

We must first establish our notation and some preliminary results. Following Chacón, Ciftan, and Biedenharn [3] we denote

$$(1) \quad [m]_n = [m_1, m_2, \dots, m_n]$$

and

$$(2) \quad S_{nm}([h]_n; [q]_m) = S_{nm}([h]; [q]) = \left[ \frac{\prod_{k=1}^m \prod_{s=1}^k \Gamma(h_s - q_k + k - s + 1)}{\prod_{k=1}^{n-1} \prod_{s=k+1}^n \Gamma(q_k - h_s + s - k)} \right]^{1/2},$$

where  $n$  and  $m$  are positive integers,  $n \geq m$ . Also, we shall denote the omission of all factors containing  $q_i$  from (2) above in the following manner:

$$(3) \quad S_{nm}([h]; [q]) = \left[ \frac{\prod_{\ell=i+1}^n \Gamma(q_i - h_\ell + \ell - i)}{\prod_{\ell=1}^i \Gamma(h_\ell - q_i + i - \ell + 1)} \right]^{1/2} S_{nm}([h]; [q]);$$

and similarly

$$(4) \quad S_{nm}([h]; [q]) = \left[ \frac{\prod_{\ell=1}^{i-1} \Gamma(q_\ell - h_i + i - \ell)}{\prod_{\ell=i}^m \Gamma(h_i - q_\ell + \ell - i + 1)} \right]^{1/2} S_{nm}([h]; [q])$$

\* Received by the editors April 3, 1979.

† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514.

and

$$(5a) \quad S_{nm}^i([h]; [q]) = \left[ \Gamma(h_i - q_j + j - i + 1) \frac{\prod_{\ell=1}^{i-1} \Gamma(q_\ell - h_i + i - \ell)}{\prod_{\ell=i}^m \Gamma(h_i - q_\ell + \ell - i + 1)} \right]^{1/2} S_{nm}^j([h]; [q])$$

for  $i \leq j$ ;

$$(5b) \quad S_{nm}^i([h]; [q]) = \left[ \frac{\prod_{\ell=1}^{i-1} \Gamma(q_\ell - h_i + i - \ell)}{\Gamma(q_j - h_i + i - j) \prod_{\ell=i}^m \Gamma(h_i - q_\ell + \ell - i + 1)} \right]^{1/2} S_{nm}^j([h]; [q])$$

for  $i > j$ .

In terms of these quantities we can express the reduced matrix elements of totally symmetric tensor operators in the following two distinct ways. From the calculation of Chacón, Ciftan, and Biedenharn [3] we have

$$(6) \quad \left\langle \left( \begin{array}{cccc} h_1 & h_2 & \cdots & h_n \\ q_1 & q_2 & \cdots & q_{n-1} \end{array} \right) \left| \begin{array}{cccc} p & 0 & \cdots & 0 \\ q & 0 & \cdots & 0 \end{array} \right| \left( \begin{array}{cccc} h'_1 & h'_2 & \cdots & h'_n \\ q'_1 & q'_2 & \cdots & q'_{n-1} \end{array} \right) \right\rangle$$

$$= \left\langle \left( \begin{array}{c} [h]_n \\ (q)_{n-1} \end{array} \right) \left| \begin{array}{cc} p & \dot{0} \\ q & \dot{0} \end{array} \right| \left( \begin{array}{c} [h']_n \\ (q')_{n-1} \end{array} \right) \right\rangle = \delta_{\sum_{i=1}^n h'_i + p, \sum_{i=1}^n h_i} \delta_{\sum_{i=1}^{n-1} q'_i + q, \sum_{i=1}^{n-1} q_i}$$

$$\cdot \sqrt{(p-q)!} \frac{S_{nn}([h]; [h]) S_{nn-1}([h']; [q']) S_{n-1n-1}([q]; [q'])}{S_{nn}([h]; [h']) S_{nn-1}([h]; [q])}$$

$$\cdot S_{n-1n-1}([q']; [q']) \sum_{\rho_1, \dots, \rho_{n-1}} (-1)^{\rho_1 + \dots + \rho_{n-1}} \left[ \frac{S_{n-1n-1}([\bar{q}]; [\bar{q}])}{S_{nn-1}([h']; [\bar{q}]) S_{n-1n-1}([q]; [\bar{q}])} \right]^2$$

$$\cdot \left[ \frac{S_{nn-1}([h]; [\bar{q}])}{S_{n-1n-1}([\bar{q}]; [q'])} \right]^2 \quad \text{where } \bar{q}_i = q'_i + \rho_i,$$

and the expression on the right will be called the CCB form of the reduced matrix element of a totally symmetric tensor operator. From the calculation of Ališauskas, Jucys and Jucys [1] we find

$$(7) \quad \left\langle \left( \begin{array}{c} [h]_n \\ (q)_{n-1} \end{array} \right) \left| \begin{array}{cc} p & \dot{0} \\ q & \dot{0} \end{array} \right| \left( \begin{array}{c} [h']_n \\ (q')_{n-1} \end{array} \right) \right\rangle = \delta_{\sum_{i=1}^n h'_i + p, \sum_{i=1}^n h_i} \delta_{\sum_{i=1}^{n-1} q'_i + q, \sum_{i=1}^{n-1} q_i}$$

$$\cdot \frac{1}{\sqrt{(p-q)!}} S_{nn}([h]; [h]) S_{n-1n-1}([q']; [q']) S_{nn}([h]; [h'])$$

$$\cdot \frac{S_{nn-1}([h]; [q])}{S_{n-1n-1}([q]; [q']) S_{nn-1}([h']; [q'])} \sum_{r_j, j \neq i} (-1)^{\varphi_i} [S_{ni}^{nn}([r]; [r])]^2$$

$$\cdot \left[ \frac{S_{ni}^{nn-1}([r]; [q'])}{S_{nn}([h]; [r]) S_{nn}([r]; [h']) S_{ni}^{nn-1}([r]; [q])} \right]^2,$$

where

$$(8) \quad \varphi_1 = \sum_{j=2}^n (h_j - r_j),$$

$$\varphi_i = \sum_{j=1}^{i-1} (q_j - q'_j + h'_j) + \sum_{j=i+1}^n h_j - \sum_{j \neq i}^n r_j, \quad 2 \leq i \leq n-1,$$

$$\varphi_n = \sum_{j=1}^{n-1} (q_j - q'_j + h'_j) - \sum_{j=1}^{n-1} r_j.$$

The expression on the right of (7) will be called the AJJ form of the reduced matrix element of a totally symmetric tensor operator. In both (6) and (7) the quantities  $h'_i, h_i, q'_i, q_i, p, q$  are real integers which satisfy the usual “betweenness” conditions of Gel’fand labels.

LEMMA. *The CCB form is equal to the AJJ form.*

*Proof.* The Wigner coefficients of  $U(n)$  are determined up to an invariant phase. Hence the CCB form and the AJJ form can differ at most by a phase, and we can prove the lemma by induction. We must first establish it for  $n = 2$ . The relation to be proved is then

$$\begin{aligned}
 & \left[ \frac{(h_1 - h_2 + 1)(h_1 - h'_1)!(h_1 - h'_2 + 1)(h_2 - h'_2)!(h_1 - q_1)!(q'_1 - h'_2)!}{(p - q)!(q_1 - q'_1)!(h'_1 - h_2)!(q_1 - h_2)!(h'_1 - q'_1)!} \right]^{1/2} \\
 & \cdot \sum_{r_2} (-1)^{h_2 - r_2} \frac{(q_1 - r_2)!(h'_1 - r_2)!}{(h_1 - r_2 + 1)!(h_2 - r_2)!(r_2 - h'_2)!(q'_1 - r_2)!} \\
 (9) \quad & = \left[ \frac{(h_1 - h_2 + 1)(p - q)!(h'_1 - q'_1)!(h'_1 - h_2)!(q_1 - h_2)!(q_1 - q'_1)!}{(q'_1 - h'_2)!(h_1 - h'_1)!(h_1 - h'_2 + 1)!(h_2 - h'_2)!(h_1 - q_1)!} \right]^{1/2} \\
 & \cdot \sum_{\rho_1} (-1)^{\rho_1} \frac{(h_1 - q'_1 - \rho_1)!(q'_1 + \rho_1 - h'_2)!}{(q'_1 + \rho_1 - h_2)!(h'_1 - q'_1 - \rho_1)!(q_1 - q'_1 - \rho_1)!(\rho_1)!}.
 \end{aligned}$$

On the left side we have written down the AJJ form for  $n = 2$  with  $i = 1$ . The identity of the AJJ form for different values of  $i$  is demonstrated in [1]. We need demonstrate only the equality of the CCB form and the AJJ form for a particular choice of  $i$ . The two sides of (9) are easily recognized as different forms of the  $U(2)$  Wigner coefficient with the standard (Condon–Shortley) phase convention. We make the correspondence

$$\begin{aligned}
 & \left\langle \begin{pmatrix} h_1 & h_2 \\ q_1 \end{pmatrix} \left| \begin{bmatrix} p & 0 \\ q \end{bmatrix} \right| \begin{pmatrix} h'_1 & h'_2 \\ q'_1 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} j_1 + j_2 + j & j_1 + j_2 - j \\ j_1 + j_2 + m \end{pmatrix} \left| \begin{bmatrix} 2j_2 & 0 \\ j_2 + m_2 \end{bmatrix} \right| \begin{pmatrix} 2j_1 & 0 \\ j_1 + m_1 \end{pmatrix} \right\rangle \\
 (10) \quad & = \begin{bmatrix} j_1 & j_2 & j \\ m_1 & m_2 & m \end{bmatrix},
 \end{aligned}$$

where we have adopted the notation of [5] for the  $U(2)$  Wigner coefficient. The expression on the left of (9) can be identified [5(13.1c), p. 81] with

$$(11) \quad (-1)^{j_1 - j + m_2} \left[ \frac{(2j + 1)}{(2j_1 + 1)} \right]^{1/2} \begin{bmatrix} j & j_2 & j_1 \\ m & -m_2 & m_1 \end{bmatrix},$$

while the expression on the right of (9) can be identified [5(13.1b), p. 81] with

$$(12) \quad (-1)^{j_1 - m_1} \left[ \frac{(2j + 1)}{(2j_2 + 1)} \right]^{1/2} \begin{bmatrix} j_1 & j & j_2 \\ m_1 & -m & -m_2 \end{bmatrix},$$

both of which are equal to (10) by the elementary symmetries of the Wigner coefficient [5(13.2), (13.3), pp. 82–83].

The equality of (6) and (7) has therefore been established for  $n = 2$ ; let us assume that it holds for  $n = N$  and seek to prove it for  $n = N + 1$ . Since we are dealing with matrix elements of totally symmetric tensor operators, which are multiplicity-free, these matrix elements factor multiplicatively into the product of matrix elements of reduced totally symmetric tensor operators. The  $U(N + 1)$  Wigner coefficient in a multiplicity-free case is defined up to an invariant phase. Hence if the AJJ form of the  $U(N)$  Wigner coefficient (i.e., the  $U(N)$  Wigner coefficient constructed as a product of



integer and

$$\begin{aligned}
 & j \geq n, \\
 & A_{ir} - A_{is} = A_{sr} \quad \text{for } s < r, \\
 (16) \quad & a_{ir} - a_{sr} = A_{is} \quad \text{for } i < s, \\
 & b_{ir} - b_{sr} = A_{is} \quad \text{for } i < s, \\
 & b_{ii} = 1, \quad 1 \leq i \leq n.
 \end{aligned}$$

These conditions are stated somewhat confusingly in (3.2) of [4], where  $k$  and  $j$  are used to denote, respectively, the numbers of columns of numerator and denominator parameters in (3.1) and arbitrary indices in the second, third, and fourth relations of (3.2).

THEOREM 1.

$$\begin{aligned}
 (17) \quad & F^{(n)} \left( \begin{array}{cccc} -z_1 + z_2 - 1 & & & \\ -z_1 + z_3 - 2 & -z_2 + z_3 - 1 & & \\ \vdots & \vdots & \ddots & \\ -z_1 + z_n - n + 1 & -z_2 + z_n - n + 2 & \cdots & -z_{n-1} + z_n - 1 \end{array} \right. \\
 & \left. \begin{array}{cccc} q_1 - z_1 + 1 & q_2 - z_1 & \cdots & q_n - z_1 - n + 2 & -z_1 - n + 1 \\ q_1 - z_2 + 2 & q_2 - z_2 + 1 & \cdots & q_n - z_2 - n + 3 & -z_2 - n + 2 \\ \vdots & \vdots & & \vdots & \vdots \\ q_1 - z_n + n & q_2 - z_n + n - 1 & \cdots & q_n - z_n + 1 & -z_n \end{array} \right. \\
 & \left. \begin{array}{cccc} 1 & z_2 - z_1 & \cdots & z_n - z_1 - n + 2 & a - z_1 + 2 \\ z_1 - z_2 + 2 & 1 & \cdots & z_n - z_2 - n + 3 & a - z_2 + 3 \\ \vdots & \vdots & & \vdots & \vdots \\ z_1 - z_n + n & z_2 - z_n + n - 1 & \cdots & 1 & a - z_n + n + 1 \end{array} \right| \begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array} \Bigg) \\
 & = \frac{\Gamma(a + \sum_{i=1}^n (z_i - q_i) + 1) [\Gamma(a - z_1 + 2) \cdots \Gamma(a - z_n + n + 1)]}{\Gamma(a + n + 1) [\Gamma(a - q_1 + 1) \cdots \Gamma(a - q_n + n)]}
 \end{aligned}$$

when the series is terminating.

*Proof.* We equate the AJJ and CCB forms of the reduced matrix element of a totally symmetric tensor operator in  $U(n)$ , i.e., the right-hand sides of (6) and (7), as we are permitted to do by the lemma above. We then introduce the restrictions  $h'_1 = q'_i$ , for  $1 \leq i \leq n - 1$ , and  $h'_n = 0$ . The CCB form immediately reduces to a monomial, and then we are left with a sum of the form

$$\begin{aligned}
 (18) \quad & \sum_{r, j \neq i} \frac{(-1)^{q_i} S_{nn}^2([r]; [r])}{S_{nn}^2([h]; [r]) S_{nn-1}^2([r]; [q]) (r_1 + n - 1)! \cdots (r_{i-1} + n - i + 1)! (r_{i+1} + n - i - 1)! \cdots (r_n)!} \\
 & = \frac{\Gamma(\sum_{j=1}^n h_j - \sum_{j=1}^{n-1} q_j + 1)}{S_{nn-1}^2([h]; [q]) [\Gamma(h_1 + n) \cdots \Gamma(h_n + 1)]}.
 \end{aligned}$$

For the sum on the left of (18) we choose  $i = 1$ , relabel the parameters, and find that (18) corresponds to the  $F^{(n-1)}$  case of (17). We may perform analytic continuation in the parameters of (17) so long as all constituent series remain terminating.

We should note that the  $n = 2$  case of (18) reduces to the Gauss summation theorem, which holds even when the series is not terminating but still convergent. If we

choose  $i = 1$  we find in this case that the factor  $S_{11}^{22}([r]; [r])$  reduces to unity and we get

$$(19) \sum_{r_2} (-1)^{h_2-r_2} \frac{\Gamma(q_1+1-r_2)}{(r_2)! \Gamma(h_1+2-r_2) \Gamma(h_2+1-r_2)} = \frac{\Gamma(h_1+h_2-q_1+1) \Gamma(q_1-h_2+1)}{\Gamma(h_1+2) \Gamma(h_2+1) \Gamma(h_1-q_1+1)}.$$

Hence (17) gives us an infinite sequence of summation theorems analogous to the Gauss theorem in higher dimensions. We may conjecture that (17) holds when the series is nonterminating but convergent, but our proof holds only for the terminating case, i.e., the case in which at least one numerator parameter in each row is a nonpositive integer.

**THEOREM 2.**

$$(20) \begin{matrix} W_q^{(n)} \left( \begin{matrix} z_1 - z_2 + 1 & & & & \\ z_1 - z_3 + 2 & z_2 - z_3 + 1 & & & \\ \vdots & \vdots & \ddots & & \\ z_1 - z_n + n - 1 & z_2 - z_n + n - 2 & \cdots & \cdots & z_{n-1} - z_n + 1 \end{matrix} \right. \\ \left. \begin{matrix} z_1 - \omega_1 + 1 & z_1 - \omega_2 + 2 & \cdots & z_1 - \omega_{n-1} + n - 1 \\ z_2 - \omega_1 & z_2 - \omega_2 + 1 & \cdots & z_2 - \omega_{n-1} + n - 2 \\ \vdots & \vdots & & \vdots \\ z_n - \omega_1 - n + 2 & z_n - \omega_2 - n + 3 & \cdots & z_n - \omega_{n-1} \end{matrix} \right. \\ \left. \left. \begin{matrix} 1 & z_1 - z_2 + 2 & \cdots & z_1 - z_n + n \\ -z_1 + z_2 & 1 & \cdots & z_2 - z_n + n - 1 \\ \vdots & \vdots & & \vdots \\ -z_1 + z_n - n + 2 & -z_2 + z_n - n + 3 & \cdots & 1 \end{matrix} \right| \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \right) = 1. \end{matrix}$$

*Proof.* This theorem is simply the orthonormality condition for matrix elements of reduced totally symmetric tensor operators with maximal initial states. In this case the CCB form (6) reduces to a monomial, and the theorem (20) above follows immediately from the condition

$$(21) \sum_{\sum_{i=1}^n h_i = p + \sum_{i=1}^n h'_i} \left[ \left\langle \begin{pmatrix} h_1 & \cdots & h_n \\ q_1 & \cdots & q_{n-1} \end{pmatrix} \left| \begin{bmatrix} p & 0 & \cdots & 0 \\ q & 0 & \cdots & 0 \end{bmatrix} \right| \begin{pmatrix} h'_1 & \cdots & h'_n \\ h'_1 & \cdots & h'_{n-1} \end{pmatrix} \right\rangle \right]^2 = 1,$$

where the sum is taken over all representations  $[h]_n$  obeying the degree condition  $\sum_{i=1}^n h_i = p + \sum_{i=1}^n h'_i$  and the “betweenness” conditions

$$(22) \quad h_i \geq q_i \geq h_{i-1}, \quad h_i \geq h'_i \geq h_{i-1}$$

for  $1 \leq i \leq n - 1$ .

This theorem and its proof were presented in [4] and are included here for the sake of completeness. The theorem, however, was stated incorrectly in (3.4) of [4]. There should be  $(n - 1)$  rather than  $n$  columns of numerator parameters in the well-poised series. Equation (20) above gives us an infinite sequence of summation theorems which begin with the classical summation theorem for well-poised  ${}_4F_3(-1)$  as the initial case,  $n = 2$ .

A number of authors [1], [8] have noted the connection that exists between matrix elements of reduced totally symmetric tensor operators in  $U(n)$  and the multiplicity-

free Racah coefficients of  $U(n - 1)$ . This relationship is given by [8], [9]

$$\begin{aligned}
 & \left\{ \begin{matrix} [W, 0] & [\mu]_n & [m']_n \\ [p, 0] & [M]_n & [m]_n \end{matrix} \right\} \\
 &= (-1)^{\varepsilon_n(m'_{1n} - m'_{nn} + m_{1n} - m_{nn} + p + W)} \frac{1}{[\dim ([m']_n) \dim ([m]_n)]^{1/2}} \\
 & \cdot \left[ \frac{\mathcal{M}([M]_n) \mathcal{M}([\mu]_n)}{\mathcal{M}([m']_n) \mathcal{M}([m]_n)} \right]^{1/2} \\
 & \left\langle \left( \begin{matrix} M_{1n} & \cdots & M_{nn} & 0 \\ m_{1n} & \cdots & m_{nn} & \end{matrix} \right) \middle| \begin{matrix} [p & 0 & \cdots & 0] \\ [p & 0 & \cdots & 0] \end{matrix} \middle| \left( \begin{matrix} m'_{1n} & \cdots & m'_{nn} & 0 \\ \mu_{1n} & \cdots & \mu_{nn} & \end{matrix} \right) \right\rangle,
 \end{aligned}
 \tag{23}$$

where

$$\varepsilon_2 = \frac{1}{2}, \quad \varepsilon_n = 1, \quad n \geq 3,
 \tag{24}$$

$$W = \sum_{i=1}^n (M_{in} - m_{in}) = \sum_{i=1}^n (m'_{in} - \mu_{in}),
 \tag{25}$$

$$\mathcal{M}([m]_n) = \left[ \frac{(m_{1n} + n - 1)! \cdots (m_{nn})!}{\prod_{j=2}^n \prod_{i=1}^{j-1} (m_{in} - m_{jn} + j - i)} \right].
 \tag{26}$$

Since the  $U(2)$  Racah coefficient is proportional to a Saalschützian  ${}_4F_3(1)$  series (a classical hypergeometric series in which the sum of the denominator parameters is equal to the sum of the numerator parameters plus one), and we realize this form of the  $U(2)$  Racah coefficient by means of the AJJ form of the matrix element of a reduced totally symmetric tensor operator in  $U(3)$ , we should be able to realize Saalschütz' theorem and a sequence of multidimensional analogs of it by introducing appropriate degeneracies into the right side of (23) and, again, comparing the AJJ and CCB forms. In the case of the  $U(2)$  Racah coefficient, which is Saalschützian  ${}_4F_3(1)$ , we find a degeneracy which sets one numerator parameter equal to a denominator parameter, whereupon the Saalschützian  ${}_4F_3(1)$  series becomes Saalschützian  ${}_3F_2(1)$ , which is summable by Saalschütz' theorem. We now want to find an analogous procedure for the multiplicity-free Wigner coefficients of all  $U(n)$ .

**THEOREM 3.**

$$\begin{aligned}
 & F^{(n)} \left( \begin{matrix} z_1 - z_2 + 1 \\ z_1 - z_3 + 2 & z_2 - z_3 + 1 & \cdots \\ \vdots & \vdots & \ddots \\ z_1 - z_n + n - 1 & z_2 - z_n + n - 2 & \cdots & z_{n-1} - z_n + 1 \end{matrix} \right) \\
 & \left( \begin{matrix} z_1 - q_1 & z_1 - q_2 + 1 & \cdots & z_1 - q_n + n - 1 & z_1 - a & z_1 - b \\ z_2 - q_1 - 1 & z_2 - q_2 & \cdots & z_2 - q_n + n - 2 & z_2 - a - 1 & z_2 - b - 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ z_n - q_1 - n + 1 & z_n - q_2 - n + 2 & \cdots & z_n - q_n & z_n - a - n + 1 & z_n - b - n + 1 \end{matrix} \right) \\
 & \left( \begin{matrix} 1 & z_1 - z_2 + 2 & \cdots & z_1 - z_n + n & z_1 - c & z_1 - d \\ z_2 - z_1 & 1 & \cdots & z_2 - z_n + n - 1 & z_2 - c - 1 & z_2 - d - 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ z_n - z_1 - n + 3 & z_n - z_2 - n + 4 & \cdots & 1 & z_n - c - n + 1 & z_n - d - n + 1 \end{matrix} \right) \left| \begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \right. \\
 &= \frac{\Gamma(1 + d - b) \Gamma(1 + c - b) \left[ \frac{\Gamma(c - q_1 + 1) \cdots \Gamma(c - q_n + n)}{\Gamma(c - z_1 + 1) \cdots \Gamma(c - z_n + n)} \right] \left[ \frac{\Gamma(d - q_1 + 1) \cdots \Gamma(d - q_n + n)}{\Gamma(d - z_1 + 1) \cdots \Gamma(d - z_n + n)} \right]}{\Gamma(a - d) \Gamma(a - c)}.
 \end{aligned}
 \tag{27}$$



when the series on the left is terminating and the parameters are related by

$$(28) \quad a + b + \sum_{i=1}^n q_i = d + c + \sum_{i=1}^n z_i + 1.$$

*Proof.* Consider the matrix element of a reduced totally symmetric tensor operator in  $U(n + 1)$  in the degenerate case indicated in (23) and equate the CCB and AJJ forms. We shall take  $i = 1$  in the AJJ form given in (7), and we shall impose the further degeneracy

$$(29) \quad m_{1n} = \mu_{1n} + p, \quad m_{in} = \mu_{in} \quad \text{for } i \geq 2.$$

We find that the CCB form then becomes a sum over a single index. We obtain

$$(30) \quad \sum_{\rho_1} (-1)^{\rho_1} \frac{S_{nn}^2([\mu + \rho_{(1)}]; [\mu + \rho_{(1)}]) S_{nn}^2([M]; [\mu + \rho_{(1)}])}{S_{nn}^2([m']; [\mu + \rho_{(1)}]) S_{nn}^2([\mu + p_{(1)}]; [\mu + \rho_{(1)}]) S_{nn}^2([\mu + \rho_{(1)}]; [\mu])}$$

$$= \frac{S_{nn}^2([M]; [\mu + p_{(1)}]) S_{nn}^2([M]; [m'])}{S_{nn}^2([m']; [\mu]) S_{nn}^2([\mu + p_{(1)}]; [\mu])}$$

$$\cdot \sum_{r_2 \cdots r_n} \frac{(-1)^{\sum_{i=2}^n (M_{in} - r_i)} S_{nn}^2([r]; [r]) S_{nn}^2([r]; [\mu])}{S_{nn}^2([r]; [\mu + p_{(1)}]) S_{nn}^2([M]; [r]) S_{nn}^2([r]; [m'])},$$

where we have used the fact that the summation index  $r_{n+1}$  on the right is restricted to the single value  $r_{n+1} = 0$  and we denote

$$(31) \quad [\mu + \rho_{(1)}] = [\mu_{1n} + \rho_1, \mu_{2n}, \dots, \mu_{n-1n}, 0],$$

$$[\mu + p_{(1)}] = [\mu_{1n} + p, \mu_{2n}, \dots, \mu_{n-1n}, 0],$$

$$[m'] = [m'_{1n}, \dots, m'_{nn}],$$

$$[M] = [M_{1n}, \dots, M_{nn}].$$

Without loss of generality we can take  $\mu_{nn} = 0$  in (23) and in (30). The sum on the left of (30) can be performed by application of the summation theorem stated in (14) of [1]. When this is done (30) above becomes Saalschütz' theorem in the  $n = 2$  case and (27) in its most general form. We note that in both (23) and (30) the parameters are related by the degree condition

$$(32) \quad \sum_{i=1}^n M_{in} = \sum_{i=1}^n m'_{in} + p.$$

When we have relabeled the parameters for ease in comparison with the classical Saalschütz' theorem, condition (32) corresponds to (28) above. Again, both sides of (27) can be continued analytically in their parameters in a unique manner so long as the series on the left remains terminating, i.e., at least one numerator parameter in each row must be a nonpositive integer.

THEOREM 4.

$$(33) \quad W_q^{(n)} \begin{pmatrix} z_1 - z_2 + 1 \\ z_1 - z_3 + 2 & z_2 - z_3 + 1 & \ddots \\ \vdots & \vdots & \ddots \\ z_1 - z_n + n - 1 & z_2 - z_n + n - 2 & \cdots & z_{n-1} - z_n + 1 \end{pmatrix} \\
 \begin{vmatrix} z_1 - \omega_1 + 1 & z_1 - \omega_2 + 2 & \cdots & z_1 - \omega_{n-1} + n - 1 \\ z_2 - \omega_1 & z_2 - \omega_2 + 1 & \cdots & z_2 - \omega_{n-1} + n - 2 \\ \vdots & \vdots & \ddots & \vdots \\ z_n - \omega_1 - n + 2 & z_n - \omega_2 - n + 3 & \cdots & z_n - \omega_{n-1} \end{vmatrix} \\
 \begin{vmatrix} 1 & z_1 - z_2 + 2 & \cdots & -z_n + z_1 + n & -a + z_1 \\ -z_1 + z_2 & 1 & \cdots & -z_n + z_2 + n - 1 & -a + z_2 - 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -z_1 + z_n - n + 2 & -z_2 + z_n - n + 3 & \cdots & 1 & -a + z_n - n + 1 \end{vmatrix} \begin{matrix} -1 \\ -1 \\ \vdots \\ -1 \end{matrix} \\
 = \frac{[\Gamma(a - \omega_1 + 2) \cdots \Gamma(a - \omega_{n-1} + n)] [\Gamma(a - q - z_1 + 1) \cdots \Gamma(a - q - z_n + n)]}{[\Gamma(a - q - \omega_1 + 2) \cdots \Gamma(a - q - \omega_{n-1} + n)] [\Gamma(a - z_1 + 1) \cdots \Gamma(a - z_n + n)]}$$

where  $q$  is a nonnegative integer.

*Proof.* Theorem 2 above provides a sequence of summation theorems beginning with the classical theorem for well-poised  ${}_4F_3(-1)$  and is obtained from the orthonormality condition for degenerate matrix elements of reduced totally symmetric tensor operators. Similarly, we can obtain a sequence of summation theorems beginning with the classical theorem for well-poised  ${}_5F_4(1)$  from the orthonormality condition for degenerate cases of multiplicity-free Racah coefficients. We take

$$(34) \quad \sum_{\sum_{i=1}^n m'_{in} = W + \sum_{i=1}^{n-1} \mu_{in}} (\dim [m'_{in}])(\dim [m_n]) \left\{ \begin{matrix} [W, \hat{0}] & [\mu]_n & [m'_{in}]_n \\ [p, \hat{0}] & [m + W_{(1)}]_n & [m]_n \end{matrix} \right\}^2 = 1,$$

where the  $m'_{in}$  are indices of summation whose sum is fixed and, as before,

$$(35) \quad [m_n + W_{(1)}] = [m_{1n} + W, m_{2n}, \dots, m_{nn}].$$

We form the degenerate Racah coefficient using (23) and the AJJ form of the matrix element of a reduced totally symmetric tensor operator in  $U(n + 1)$ , and we choose  $i = 1$  in (7) above. We immediately obtain the sequence of theorems (33) above, since the Racah coefficient with degeneracies indicated in (34) will become a monomial when we construct it as indicated. Note that in the limit  $a \rightarrow +\infty$  we find that (33) becomes identical with (20).

For the sake of completeness we present here the other theorem, in addition to Theorem 2 above, which was established in [4]. The proof is given in [4] with reference to work contained in [2]. This theorem is an analog of Whipple's theorem [7] which establishes a relationship between well-poised  ${}_7F_6(1)$  and Saalschützian  ${}_4F_3(1)$ . This relationship is realized in the representation theory of  $U(2)$  as a degenerate case of the Biedenharn–Elliott identity, given in (1.25) of [4]. The following analog arises in the representation theory of  $U(3)$  and is an analog in the limited sense that it relates well-poised and Saalschützian forms. No sequence is yet known of which both Whipple's theorem and (36) below are individual members.

THEOREM 5.

$$\begin{aligned}
 W_q^{(3)} & \left( \begin{array}{ccc|ccc}
 x_3 + \Delta_1 - \Delta_2 & & & 1 & x_3 + \Delta_1 - \Delta_2 + 1 & -x_2 + \Delta_1 - \Delta_3 + 1 \\
 -x_2 + \Delta_1 - \Delta_3 & & & -x_3 - \Delta_1 + \Delta_2 + 1 & 1 & x_1 + \Delta_2 - \Delta_3 + 1 \\
 & x_1 + \Delta_2 - \Delta_3 & & x_2 - \Delta_1 + \Delta_3 + 1 & -x_1 - \Delta_2 + \Delta_3 + 1 & 1 \\
 \hline
 & & & \Delta_1 - q + 1 & x_3 + \Delta_1 - q + 1 & -x_2 + \Delta_1 - q + 1 \\
 & & & -x_3 + \Delta_2 - q + 1 & \Delta_2 - q + 1 & x_1 + \Delta_2 - q + 1 \\
 & & & x_2 + \Delta_3 - q + 1 & -x_1 + \Delta_3 - q + 1 & \Delta_3 - q + 1 \\
 & & & & & 1 \\
 \hline
 & & & & & 1 \\
 & & & & & 1
 \end{array} \right) \\
 (36) \quad & = q! \sum_{k_1, k_2, k_3} \frac{\Gamma(\Delta_1 + \Delta_2 + \Delta_3 - k_1 - k_2 - k_3 + 1)}{\Gamma(\Delta_1 + \Delta_2 + \Delta_3 - q + 1)(q - k_1 - k_2 - k_3)!} \\
 & \cdot \frac{\Gamma(\Delta_2 + x_1 - q + 1)\Gamma(\Delta_3 - x_1 - q + 1)\Gamma(\Delta_1 - q + 1)}{k_1! \Gamma(\Delta_2 + x_1 - k_1 + 1)\Gamma(\Delta_3 - x_1 - k_1 + 1)\Gamma(\Delta_1 - q - k_1 + 1)} \\
 & \cdot \frac{\Gamma(\Delta_3 + x_2 - q + 1)\Gamma(\Delta_1 - x_2 - q + 1)\Gamma(\Delta_2 - q + 1)}{k_2! \Gamma(\Delta_3 + x_2 - k_2 + 1)\Gamma(\Delta_1 - x_2 - k_2 + 1)\Gamma(\Delta_2 - q + k_2 + 1)} \\
 & \cdot \frac{\Gamma(\Delta_1 + x_3 - q + 1)\Gamma(\Delta_2 - x_3 - q + 1)\Gamma(\Delta_3 - q + 1)}{k_3! \Gamma(\Delta_1 + x_3 - k_3 + 1)\Gamma(\Delta_2 - x_3 - k_3 + 1)\Gamma(\Delta_3 - q + k_3 + 1)},
 \end{aligned}$$

where  $q$  is a nonnegative integer and  $x_1 + x_2 + x_3 = 0$ . There are no numerator parameters in the well-poised series on the left of (36).

REFERENCES

[1] S. J. ALIŠAUSKAS, A.-A. A. JUCYS AND A. P. JUCYS, *On the symmetric tensor operators of the unitary groups*, J. Mathematical Phys., 13 (1972), pp. 1329–1333.

[2] L. C. BIEDENHARN AND J. D. LOUCK, *On the structure of the canonical tensor operators in the unitary groups*, Ibid., 13 (1972), pp. 1985–2001.

[3] E. CHACÓN, M. CIFTAN AND L. C. BIEDENHARN, *On the evaluation of the multiplicity-free Wigner coefficients of  $U(n)$* , Ibid., 13 (1972), pp. 577–590.

[4] W. J. HOLMAN, III, L. C. BIEDENHARN AND J. D. LOUCK, *On hypergeometric series well-poised in  $SU(n)$* , this Journal, 7 (1976), pp. 529–541.

[5] A. P. JUCYS AND A. A. BANDZAITIS, *Teoriya Momenta Kolichestva Dvizheniya v Kvantovoi Mekhanike*, Izdatelstvo MINTIS, Vilinius, U.S.S.R., 1965. (In Russian.)

[6] F. J. W. WHIPPLE, *On well-poised series, generalized hypergeometric series having parameters in pairs, each pair with the same sum*, Proc. London Math. Soc., 24 (1926), pp. 247–263.

[7] ———, *Well-poised series and other generalized hypergeometric series*, Ibid., 25 (1926), pp. 525–544.

[8] M. F. K. WONG, *Multiplicity-free  $6-j$  symbols and Weyl coefficients of  $U(n)$ : Explicit evaluation*, J. Mathematical Phys., 19 (1978), pp. 1635–1643.

[9] ———, private communication.

## PERIODIC SOLUTIONS TO A NONLINEAR VOLTERRA INTEGRO-DIFFERENTIAL EQUATION\*

HARLAN W. STECH†

**Abstract.** A nonlinear Volterra integro-differential equation arising from the theory of population dynamics is shown to have a nonconstant periodic solution whenever the (biologically important) steady state is unstable. Existence of the periodic solution is proved by a fixed point argument.

**Introduction.** For  $\alpha, \beta$  and  $N_0 > 0$ , we consider the scalar equation

$$(1) \quad \dot{N}(t) = \alpha N(t) \left[ 1 - \frac{1}{N_0} \int_{-\infty}^{-r} N(t+s) |s+r|^n \frac{\beta^{n+1}}{n!} e^{\beta(s+r)} ds \right],$$

where  $r \geq 0$ , and  $n$  is a nonnegative integer. Equation (1) arises in the study of population fluctuations:  $N(t) > 0$  represents the population density of a species;  $\alpha$  is the “intrinsic rate of growth” of the species, and  $N_0$  is the “carrying capacity” of the environment. The choice of constants in the integrand is taken so that  $N(t) = N_0$  is a steady state for (1). The integral term represents a self-regulating or negative feed-back mechanism (with positive definite time delay when  $r > 0$ ). Both Cushing [4] and May [14] provide complete discussions of the model’s derivation.

Many authors have studied the stability of the steady state solution for variants of model (1). (See, for example [4], [18] and [21] and the references therein.) For the particular case we will consider it will be shown that whenever the steady state solution to (1) is unstable (in a linearized sense) then there exists a nonconstant, positive periodic solution. In particular, it will follow that for arbitrary  $n \geq 0, \beta > 0$ , (1) has a periodic solution of period larger than  $2r$  whenever  $\alpha r > \pi/2$ .

Periodicity results for equations similar to (1) have been obtained by Hopf bifurcation techniques (see, for example [3], [8], [12] and [19]) as well as topological fixed point methods ([1], [2], [5], [6], [15], [16], [20] among others). The first approach allows, in some cases, information concerning the stability of the periodic orbits, but often applies only to a restricted range of parameter values. The latter method applies to a wide range of parameter values but allows no conclusions concerning stability. The techniques used here fall in this latter category.

It is important to note that the analysis to follow relies heavily on the particular kernel in (1). Its special form allows one to relate (1) to a system of differential equations of finite delay type (when  $r > 0$ ) or (when  $r = 0$ ) an ordinary differential system. Equations with kernels similar to that in (1) have been the subject of considerable interest—especially in the case  $r = 0$ . See, for example, [4], [10], [13], [14], [21] and their extensive lists of references. We note also that the kernel of (1) arises in [16] in the special case when  $n = 0$  and  $r > 0$ . The results presented here are, in some sense, complementary to those in [1], [2] and [20] in that the kernels considered there all have compact support. While it would be of interest to know to what extent the periodicity theorem of this paper generalizes to equations with the same qualitative form as that found in (1) (e.g., kernels with “large” first moment and “small” variance), many of the arguments used in this paper do not lend themselves to this more general problem.

\* Received by the editors January 3, 1979, and in revised form October 5, 1979.

† Department of Mathematics, Iowa State University, Ames, Iowa 50010, and Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

Section 1 contains a statement of the principal result along with some preliminary analysis. The characteristic equation associated with the linearization of (1) about  $N_0$  is analyzed in Lemma 2.1. In § 2 we detail the construction of a cone  $K$  of “initial conditions” and a completely continuous map  $A : K \rightarrow K$ . A nontrivial fixed point of  $A$  will correspond to a nontrivial periodic solution to (1). Using an ejective fixed point theorem of Nussbaum [15] and an ejectivity criteria of Chow and Hale [2], we prove in § 3 the existence of such a nontrivial fixed point. In the case  $r = 0$ , it is evident that there is a close connection between the method used here and that of Hastings, Tyson and Webster [9] concerning a different negative feedback model.

**1. Preliminaries.** It is convenient to shift the equilibrium to the origin by setting  $x(t) = -1 + N(t)/N_0$ . Then  $x$  solves

$$(1.1) \quad \dot{x}(t) = -\alpha[x(t) + 1] \int_{-\infty}^{-r} x(t+u) \rho_n(u+r) du,$$

where  $\rho_n(v) = \beta^{n+1}|v|^n e^{\beta v}/n!$ . For any bounded continuous initial function  $x(t) = \varphi(t)$ ,  $t \leq 0$ , the solution of (1.1) exists for all  $t > 0$  and is unique.

We shall prove

**THEOREM 1.1.** *Let  $\beta > 0$ ,  $r \geq 0$  and  $n \geq 0$  (with not both  $r$  and  $n$  equal to zero). Define  $\nu_0 > 0$  to be the smallest solution of  $r\nu_0 + (n+1) \tan^{-1}(\nu_0/\beta) = \pi/2$  and  $\alpha_0 \equiv \nu_0(1 + (\nu_0/\beta)^2)^{(n+1)/2}$ . For  $\alpha > \alpha_0$  (1.1) has a nontrivial periodic solution  $x(\cdot) > -1$  with period larger than  $2r$ .*

For the case  $r = n = 0$  excluded by these hypotheses, it is well-known that the zero solution of (1.1) is attracting for all solutions  $x > -1$ . (See, for example, Staffans [17].) Thus, no positive nontrivial periodic solution of (1) exists.

To discuss the stability of the zero solution of (1.1) we linearize about  $x = 0$  to obtain

$$(1.2) \quad \dot{z}(t) = -\alpha \int_{-\infty}^{-r} z(t+u) \rho_n(u+r) du.$$

This will have the solution  $z(t) = e^{\lambda t}$  ( $\lambda$  complex with  $\text{Re } \lambda > -\beta$ ) if and only if  $\lambda$  satisfies the characteristic equation

$$(1.3) \quad \begin{aligned} \lambda &= -\alpha \int_{-\infty}^{-r} e^{\lambda u} \rho_n(u+r) du \\ &= -\alpha e^{-r\lambda} (1 + \lambda/\beta)^{-(n+1)}. \end{aligned}$$

Thus, the zero solution of (1.2) will be unstable whenever (1.3) is satisfied by a  $\lambda$  with positive real part. The existence of such a root is given by the following lemma.

**LEMMA 1.2.** *Fix  $n \geq 0$  and  $\beta > 0$ . Define  $\nu_0$  and  $\alpha_0$  as in Theorem 1.1. For  $0 < \alpha < \alpha_0$  all solutions  $\lambda = \mu + i\nu$  to (1.3) have negative real part. For  $\alpha > \alpha_0$  there is a simple zero of  $\lambda + \alpha e^{-\lambda r} (1 + \lambda/\beta)^{-(n+1)}$  with  $\mu > 0$ ,  $\nu_0 < \nu$  (and  $\nu < \pi/r$  if  $r > 0$ ).*

*Proof.* Considering the argument and modulus of each side of (1.3) we see that  $\lambda = \mu + i\nu$  will be a characteristic root if and only if

$$(1.4) \quad \tan^{-1}\left(\frac{\nu}{\mu}\right) = \pi - \nu r - (n+1) \arg(1 + \lambda/\beta), \quad \text{mod } 2\pi$$

and

$$(1.5) \quad \nu^2 + \mu^2 = \alpha^2 e^{-2r\mu} [(1 + \mu/\beta)^2 + (\nu/\beta)^2]^{-(n+1)}.$$

Since complex roots occur in conjugate pairs we may assume  $\nu \geq 0$ .

Case 1:  $r > 0$ . Considering, for the moment, purely imaginary roots  $\lambda = i\nu$ , (1.4) reduces to

$$(1.6) \quad r\nu + (n + 1) \tan^{-1}(\nu/\beta) = \frac{\pi}{2} + 2\pi m, \quad m \text{ an integer.}$$

The left side of (1.6) defines an increasing unbounded function of  $\nu$  that is zero when  $\nu = 0$ . Thus, for each  $m \geq 0$  there is a unique solution  $\nu_m > 0$  to (1.6), and  $i\nu_m$  is a root provided  $\alpha$  is taken so that (1.5) (with  $\mu = 0, \nu = \nu_m$ ) is satisfied. That is,  $\alpha = \nu_m [1 + (\nu_m/\beta)^2]^{(n+1)/2} \equiv \alpha_m$ . Clearly  $\nu_m$  (and therefore  $\alpha_m$ ) increases with  $m$ .

For  $\lambda = \mu + i\nu, \mu > 0$ , the existence of roots can be shown similarly. Equation (1.4) may be written

$$(1.7) \quad r\nu + \tan^{-1}(\nu/\mu) + (n + 1) \tan^{-1}(\nu/(\mu + \beta)) = \pi + 2\pi m, \quad m \geq 0.$$

For fixed  $\mu > 0$  and  $m \geq 0$ , we find a unique solution  $\nu_m(\mu)$  of (1.7)—then determine  $\alpha = \alpha_m(\mu)$  from (1.5) (substituting  $\lambda = \mu + i\nu_m(\mu)$ ). It is not difficult to show  $\nu_m(\mu)$  is an increasing function of  $\mu$ , and therefore (using (1.5)) that  $\alpha_m(\mu)$  increases monotonically without bound as  $\mu$  increases from zero. It follows that the minimum value of  $\alpha$  for which there are roots with nonnegative real part is  $\alpha_0(0) = \alpha_0$ . For  $\alpha > \alpha_0$  the monotonicity and unboundedness of  $\alpha_0(\mu)$  show that (1.7) (with  $m = 0$ ) is solved by  $\mu = \alpha_0^{-1}(\alpha) > 0$  and a unique  $\nu = \nu_0(\mu)$  with  $\pi/r > \nu_0(\mu) > \nu_0(0) = \nu_0$ .

Concerning the root's simplicity, we assume the opposite, and differentiate

$$\lambda(1 + \lambda/\beta)^{n+1} + \alpha e^{-r\lambda} = 0$$

with respect to  $\lambda$  at the root. Into the resulting expression one can substitute for  $\alpha e^{-r\lambda}$  the expression derived from (1.3). It is easy to see that  $\lambda$  must be a root of a polynomial that has all roots with negative real part—a contradiction.

Case 2:  $r = 0, n \geq 1$ . Arguing as in the previous case, we consider first purely imaginary roots. The left side of (1.6) increases monotonically from 0 to  $(n + 1)(\pi/2)$ . Thus, for each  $m \geq 0$  satisfying  $4m < n$ , there is a unique solution  $\nu_m = \tan((\pi/2)((1 + 4m)/(1 + n)))$  of (1.6). With minor adjustments (often simplifications) the line of reason in Case 1 applies here as well, so we will omit further details.  $\square$

Associated with (1.1) is an  $n + 2$  dimensional system of finite delay equations. For any solution  $x$  of (1.1) for  $t > 0$  we define

$$(1.8) \quad \begin{aligned} y_k(t) &= \int_{-\infty}^{-r} x(t+u)\rho_k(u+r) du \\ &= \int_{-\infty}^{t-r} x(s)\rho_k(s-t+r) ds \end{aligned}$$

for  $k = 0, 1, \dots, n$ . Differentiating, we find

$$(1.9) \quad \begin{aligned} \dot{x} &= -\alpha[x + 1]y_n, \\ \dot{y}_n &= -\beta y_n + \beta y_{n-1}, \\ &\vdots \\ \dot{y}_1 &= -\beta y_1 + \beta y_0, \\ \dot{y}_0 &= -\beta y_0 + \beta x(t-r). \end{aligned}$$

For (1.9) it is natural to use the space of initial conditions  $X \equiv C([-r, 0]) \times \mathbb{R}^{n+1}$ , where  $C[-r, 0]$  is the space of continuous real valued functions on the interval  $[-r, 0]$ .

Elements of  $X$  will be denoted by  $\psi = (\varphi(\cdot), y_n, y_{n-1}, \dots, y_1, y_0)$ , with the norm  $\|\psi\| = \max \{ \sup_{[-r, 0]} |\varphi(s)|, |y_n|, \dots, |y_0| \}$ . (When  $r = 0$ ,  $X \cong \mathbb{R}^{n+2}$ .)

Concerning periodic solutions, it is evident from (1.8) that each periodic solution of (1.1) defines a periodic solution of (1.9). The converse is true as well.

LEMMA 1.3. *If  $(x(\cdot), y_n(\cdot), \dots, y_0(\cdot))$  is a periodic solution of (1.9) then*

$$y_k(t) = \int_{-\infty}^{-r} x(t+u)\rho_k(u+r) du, \quad k = 0, 1, \dots, n.$$

Thus  $x(\cdot)$  is a periodic solution of (1.1).

*Proof.* The argument is essentially that found in [10] for an equation with similar kernel.

Let  $-\infty < \tau < t$  and integrate the last  $n + 1$  equations over  $[\tau, t]$  to obtain

$$y_0(t) = y_0(\tau) e^{\beta(t-\tau)} + \beta \int_{\tau}^t e^{\beta(t-s)} x(s-r) ds,$$

$$y_k(t) = y_k(\tau) e^{\beta(t-\tau)} + \beta \int_{\tau}^t e^{\beta(t-s)} y_{k-1}(s) ds$$

for  $k = 1, \dots, n$ . The periodic solution is bounded so that we may let  $\tau \rightarrow -\infty$ . Thus

$$y_0(t) = \beta \int_{-\infty}^t e^{\beta(t-s)} x(s-r) ds$$

and

$$y_k(t) = \beta \int_{-\infty}^t e^{\beta(t-s)} y_{k-1}(s) ds, \quad k = 1, \dots, n.$$

The conclusion now follows by an elementary induction argument.  $\square$

By this equivalence, we must show (1.9) to have a periodic solution with  $x > -1$ . Towards that end, we introduce the following convex cone of initial conditions for (1.9). Let  $K \equiv \{ \psi = (\varphi(\cdot), 0, y_{n-1}, \dots, y_0) \in X \mid \varphi \text{ is a continuous, nonnegative and increasing function defined on } [-r, 0], 0 \leq y_k \leq \varphi(0) \text{ for } k = 0, \dots, n \}$ . For  $0 \neq \psi \in K$ , we will show that there is a first  $t = \tau(\psi) > 2r$  such that the solution of (1.9) with initial condition  $\psi$  satisfies  $(x(\tau + \cdot), y_n(\tau), y_{n-1}(\tau), \dots, y_0(\tau)) \in K$ . (By  $x(\tau + \cdot)$  we mean the element of  $C[-r, 0]$  whose value at  $s$  is  $x(\tau + s)$ .) In addition, it will be necessary that we show  $\tau(\psi)$  bounded uniformly for bounded  $\|\psi\|$ . This will be shown with the aid of the following technical lemma.

LEMMA 1.4. *Let  $\psi \in X$  and  $x(\psi)$  denote the first coordinate of the solution to (1.9) with initial condition  $\psi$ . Then  $x(t) = x(\psi)(t)$  satisfies*

$$\begin{aligned} \dot{x}(t) &= -\alpha[x(t) + 1] \left[ \sum_{k=0}^n \frac{y_k(0)}{(n-k)!} (\beta t)^{n-k} e^{-\beta t} + \int_{-r}^{t-r} \varphi(u)\rho_n(u-t+r) du \right] \\ &\qquad\qquad\qquad \text{for } 0 \leq t \leq r \\ (1.10) \quad &= -\alpha[x(t) + 1] \left[ \sum_{k=0}^n \frac{y_k(0)}{(n-k)!} (\beta t)^{n-k} e^{-\beta t} + \int_{-r}^0 \varphi(u)\rho_n(u-t+r) du \right. \\ &\qquad\qquad\qquad \left. + \int_0^{t-r} x(u)\rho_n(u-t+r) du \right] \\ &\qquad\qquad\qquad \text{for } r \leq t. \end{aligned}$$

(The integrals over  $[-r, 0]$  are absent if  $r = 0$ .)

*Proof.* By direct computation one can show that the last  $n + 1$  coordinates of the solution of (1.9) with initial condition  $\psi$  are given by

$$y_j(t) = \sum_{k=0}^j \frac{y_k(0)}{(j-k)!} (\beta t)^{j-k} e^{-\beta t} + \int_{-r}^{t-r} x(u) \rho_j(u-t+r) du, \quad j=0, \dots, n$$

for  $t > 0$ .  $\square$

**2. Construction of a cone map.** Fix  $\beta > 0$ ,  $r$  and  $n \geq 0$ , and assume  $\alpha > \alpha_0$ . We proceed to show the existence and boundedness of the function  $\tau : K \setminus \{0\} \rightarrow [2r, \infty)$  defined previously. As above, we will write  $x(\cdot) = x(\psi)(\cdot)$  for the first coordinate of the solution of (1.9) with initial condition  $\psi$ .

**LEMMA 2.1.** *Assume  $c > 0$  and  $\psi \in K$  satisfies  $0 < \|\psi\| \leq C$ . Then there is a first  $t_1 = t_1(\psi) \geq 0$  at which  $x(t_1) = 0$ , and  $x(t_1 + \varepsilon) < 0$  for small  $\varepsilon > 0$ . As a function of  $\psi$ ,  $t_1(\psi)$  is continuous and bounded above by a constant dependent only on  $C$ .*

*Proof.* We will use the existence of a characteristic root  $\lambda = \mu + i\nu$  with positive real part to construct a comparison function with which we can show  $x$  to have the desired change in sign. As seen from the previous section,  $\xi(t) = e^{\mu t} \sin(\nu t)$  solves (1.2). Define for  $l > 0$ ,  $\xi^{(l)}(t) = \xi(t)$  for  $-l - r \leq t \leq 0$  and  $\xi^{(l)}(t) = 0$ , otherwise. Since  $\xi^{(l)} \rightarrow \xi$  uniformly on  $(-\infty, 0]$  as  $l \rightarrow \infty$ , the solution  $z^{(l)}(\cdot)$  to (1.2) with initial condition  $\xi^{(l)}$  will vary continuously with  $l$ . As  $l \rightarrow \infty$ ,  $z^{(l)} \rightarrow \xi$  uniformly on compact subsets of  $[0, \infty)$ . Thus one may find  $l$  sufficiently large so that (as is the case for  $\xi$ ) there is a unique  $0 < t^* < 3\pi/(2\nu)$  such that  $0 < z^{(l)}(t)$  for  $0 < t < t^*$ ,  $z^{(l)}(t) < 0$  for  $t^* < t \leq 3\pi/(2\nu)$ , and the maximum value of  $z^{(l)}(t)$  for  $-\infty < t \leq t^*$  lies in  $[0, 3\pi/(2\nu)]$ .

We claim that  $x$  must change sign on  $[0, r + l + t^*]$ . If not then  $\psi \in K$  and (1.10) imply  $x$  is nonincreasing on  $[0, r]$  and decreasing on  $[r, r + l + t^*]$ . Choose  $\delta > 0$  to be the maximum number such that  $\delta z^{(l)}(t - l - r) \leq x(t)$  for  $0 \leq t \leq r + l + 3\pi/(2\nu)$ . (Such a  $\delta$  exists since  $x$  is positive on this interval.) By our choice of  $\delta$  there is some  $0 \leq t^{**} < r + l + 3\pi/(2\nu)$  such that  $\delta z^{(l)}(t^{**} - l - r) = x(t^{**})$  and  $\delta z^{(l)}(t - l - r) < x(t)$  for  $t^* < t < r + l + 3\pi/(2\nu)$ . In fact, by our choice of  $l$  and the monotonicity of  $x$  we have  $r + l < t^{**} < r + l + t^*$ . Finally, for  $t^{**} < t < r + l + t^*$ , we have from (1.10)

$$\begin{aligned} \dot{x}(t) &\leq -\alpha [0 + 1] \int_0^{t-r} x(u) \rho_n(u-t+r) du \\ &\leq -\alpha \delta \int_0^{t-r} z^{(l)}(u-l-r) \rho_n(u-t+r) du \\ &= -\alpha \delta \int_{-t}^{-r} z^{(l)}(t-l-r+s) \rho_n(s+r) ds \\ &= -\alpha \delta \int_{-\infty}^{-r} z^{(l)}(t-l-r+s) \rho_n(s+r) ds \end{aligned}$$

since  $z^{(l)}(t-l-r+s) = 0$  if  $s < t$ . Now, integrate over  $[t^{**}, r + l + t^*]$  and use (1.2) to obtain

$$\begin{aligned} x(t^* + r + l) - x(t^{**}) &\leq \delta \int_{t^{**}}^{r+l+t^*} \dot{z}^{(l)}(t-l-r) dt \\ &= \delta z^{(l)}(t^*) - \delta z^{(l)}(t^{**} - l - r). \end{aligned}$$

From the definition of  $t^{**}$  and  $z^{(l)}(t^*) = 0$ , we obtain the contradiction  $x(t^* + r + l) \leq 0$ .



The existence of a first zero has been proved. Since  $\psi \neq 0$  and  $x(t) > 0$  for  $0 \leq t \leq t_1(\psi)$ , (1.10) shows  $\dot{x}(t_1(\psi)) < 0$ . Thus, the zero of  $x$  at  $t_1(\psi)$  is simple. The uniform bound on  $t_1(\psi)$  follows directly from the proof ( $l$  is independent of  $C$ ). Continuity of  $t_1$  at  $\psi \neq 0$  is a consequence of continuous dependence of  $x(\psi)$  on its initial data.  $\square$

As we have observed in the lemma, if  $0 \neq \psi \in K$  then (1.10) shows  $x$  to decrease for values of  $t$  slightly larger than  $t_1(\psi)$ . (In fact, if  $r > 0$  then  $x$  decreases on  $[t_1, t_1 + r]$ .) Clearly  $x(\psi)$  must continue to decrease until either  $y_n(t)$  or  $[x(t) + 1]$  change signs. The latter is not possible since (1.10) implies that as long as  $x$  decreases from  $x(0) = \|\psi\| \leq C$ ,

$$\begin{aligned} \dot{x}(t) &\geq -\alpha[x(t) + 1] \left( 1 + \int_{-r}^{t-r} \rho_n(u - t + r) du \right) \|\psi\| \\ &\geq -\alpha[x(t) + 1] 2C. \end{aligned}$$

(Note:  $1 = e^{\beta t} e^{-\beta t} \geq \sum_{k=0}^{n-1} (1/(n-k)!) (\beta t)^{n-k} e^{-\beta t}$ .) Thus  $(d/dt) \ln(x(t) + 1) \geq -2\alpha C$  and (integrating over  $[t_1(\psi), t]$ )

$$x(t) \geq -1 + \exp\{-(t - t_1(\psi))2\alpha C\}.$$

We may integrate the last  $n + 1$  equations of (1.9) to obtain the equivalent system

$$\begin{aligned} \dot{x}(t) &= -\alpha[x(t) + 1]y_n(t), \\ y_n(t) &= y_n(0) e^{-\beta t} + \beta \int_0^t e^{-\beta(t-s)} y_{n-1}(s) ds, \\ &\vdots \\ y_1(t) &= y_1(0) e^{-\beta t} + \beta \int_0^t e^{-\beta(t-s)} y_0(s) ds, \\ y_0(t) &= y_0(0) e^{-\beta t} + \beta \int_0^t e^{-\beta(t-s)} x(s - r) ds. \end{aligned} \tag{2.1}$$

From this and the positivity of the  $y_k(0)$ , it is clear that  $y_n(t)$  can change sign only after  $y_0(t)$ ,  $y_1(t)$ ,  $\dots$ , and  $y_{n-1}(t)$  have. In fact, as  $x(t) < 0$  decreases past  $t = t_1 + r$ , the last equation in (2.1) shows  $y_0(t)$  to eventually decrease and change sign at some  $t = t^0 > t_1 + r$ . For  $t > t^0$  we have

$$e^{\beta t} y_0(t) = \beta \int_0^t e^{\beta s} x(s - r) ds$$

so that  $y_0(t) e^{\beta t}$  remains negative and decreases as long as  $e^{\beta s} x(s - r) < 0$ . From the equation for  $y_1(t)$  in (2.1) we see that this forces  $y_1(t)$  to eventually decrease and change sign at some  $t = t^1 \geq t^0$ . For  $t > t^1$ ,

$$e^{\beta t} y_1(t) = \int_{t^1}^t e^{\beta s} y_0(s) ds$$

is negative and decreasing. Continuing this line of reasoning, we see that  $y_n$  will eventually change sign at some  $t \equiv t_2(\psi) > t_1(\psi) + r$ .

Observe also that since  $x$  decreases on  $(t_1, t_2]$ ,

$$\begin{aligned} e^{\beta t_2} y_0(t_2) &= \beta \int_{t^0}^{t_2} e^{\beta s} x(s - r) ds \\ &\geq \beta \int_{t^0}^{t_2} e^{\beta s} ds x(t_2) \end{aligned}$$

so that  $y_0(t_2) \geq x(t_2)$ . Similarly, since  $y_0$  decreases on  $[t^0, t_2]$ ,  $y_1(t_2) \geq y_0(t_2) \geq x(t_2)$ .

Continuing, we see that  $y_k(t_2) \geq x(t_2)$  for  $k = 0, \dots, n - 1$  so that  $(x(t_2 + \cdot), y_n(t_2), y_{n-1}(t_2), \dots, y_0(t_2))$  defines an element of  $-K$ .

The two technical lemmas to follow show  $t_2(\psi)$  to be uniformly bounded for  $0 < \|\psi\| \leq C$ .

LEMMA 2.2. *Let  $0 < \|\psi\| \leq C$ , and assume  $t_2(\psi) > t_1(\psi) + 1 + r$ . There is a  $\delta > 0$  dependent only on  $C$  such that  $x(t_1(\psi) + 1) \leq -\delta\|\psi\|$ .*

*Proof.* Since  $x$  decreases on  $[0, t_1]$ ,  $x(t) \leq x(0) = \|\psi\| \leq C$ . Thus (as above)

$$\begin{aligned} \dot{x}(t) &\geq -\alpha[C + 1] \left[ x(0) + \int_{-r}^{t-r} \rho_n(u - t + r) du \cdot x(0) \right] \\ &\geq -2\alpha[C + 1] \|\psi\| \end{aligned}$$

which we integrate over  $0 \leq t \leq u \leq t_1$  to obtain  $x(u) \geq \|\psi\| a(u)$ , where  $a(u) \equiv \max \{1 - 2\alpha(C + 1)u, 0\}$  for  $0 \leq u \leq t_1$ . (We define  $a(u) \equiv 0$  outside that interval.) Over  $[t_1, t_1 + 1]$ , (2.1) implies

$$\dot{x}(t) \leq -\alpha[x(t) + 1] \left[ 0 + 0 + \|\psi\| \int_0^{t-r} a(u) \rho_n(u - t + r) du \right].$$

We assume  $x(t + 1 + r) \geq -\frac{1}{2}$  (otherwise, take  $\delta = 1/(2C)$ ). Then

$$\dot{x}(t) \leq -\frac{\alpha}{2} \|\psi\| \int_0^{t-r} a(u) \rho(u - t + r) du$$

since  $x$  decreases on  $[t_1, t_1 + 1 + r]$ . Finally, we integrate over that interval to obtain the desired  $\delta$ .  $\square$

LEMMA 2.3. *For  $0 < \|\psi\| \leq C$ ,  $t_2(\psi)$  is uniformly bounded by a constant dependent on  $C$  alone.*

*Proof.* Without loss of generality we may restrict our attention to those  $\psi$  satisfying the hypotheses of the previous lemma. For  $t_1(\psi) + 1 + r < t < t_2(\psi)$  the monotonicity of  $x$  implies

$$\begin{aligned} y_n(t) &\leq \|\psi\| \left[ \sum_{k=0}^{n-1} \frac{1}{(n-k)!} (\beta t)^{n-k} e^{-\beta t} + \int_{-r}^0 \rho_n(u - t + r) du \right] \\ &\quad - \|\psi\| \delta \int_{t_1+1+r}^{t-r} \rho_n(u - t + r) du. \end{aligned}$$

The last term approaches  $-\|\psi\|\delta$  as  $t \rightarrow +\infty$  while the first approaches zero. Thus, the uniform bound on the zero of  $y_n, t_2(\psi)$ , follows from that of  $t_1(\psi)$ .  $\square$

From this lemma and the discussion following Lemma 2.1 we obtain

COROLLARY 2.4. *The minimum value of  $x$  on  $[0, t_2(\psi)]$ ,  $x(t_2(\psi))$ , is bounded above  $-1$  by a positive constant dependent only on  $C$ .*

LEMMA 2.5. *There is a first  $t_3 = t_3(\psi) > t_2(\psi)$  for which  $x(t_3) = 0$  and  $x(t_3 + \varepsilon) > 0$  for small  $\varepsilon > 0$ . As a function of  $\psi$ ,  $t_3(\psi)$  is continuous and bounded on  $0 < \|\psi\| \leq C$  by a constant dependent on  $C$  alone.*

*Proof.* Since (1.9) is autonomous, we may consider  $x(t)$  for  $t > t_2$  as the first coordinate of the solution to (1.9) for  $t > 0$  with initial data given by  $\bar{\varphi}(u) = x(t_2(\psi) + u)$ ,  $-r \leq u \leq 0$ , and  $\bar{y}_k = y_k(t_2(\psi))$ ,  $k = 0, 1, \dots, n$ . From (1.10) and  $(\bar{\varphi}, \bar{y}_n, \dots, \bar{y}_0) \in -K$ , we see that  $x$  must increase as long as  $x(t) < 0$ .

Assume for the moment that  $x(t_2) \geq \frac{1}{2}[\alpha_0/\alpha - 1] = (\alpha_0 + \alpha)/2\alpha - 1$ . Then from (1.10) we have  $\dot{x}(t) \geq -((\alpha_0 + \alpha)/2)y_n$ , and we may proceed exactly as in Lemma 2.1 to construct a comparison function to show the existence and boundedness of  $t_3(\psi)$ . Note

that since  $(\alpha_0 + \alpha)/2 > \alpha_0$ , Lemma 1.2 provides the needed growing exponential solution to the linear equation

$$\dot{z}(t) = -\frac{\alpha_0 + \alpha}{2} \int_{-\infty}^{-r} z(t+u) \rho_n(u+r) du.$$

The details are omitted.

Finally, if  $x(t_2) \cong \frac{1}{2}[(\alpha_0/\alpha) - 1]$  fails, one can use the bound given by Corollary 2.4 and an argument similar to that used in Lemma 2.3 to show that  $x(t)$  will increase until the inequality does hold at some  $\bar{t}_2 > t_2$ , with  $\bar{t}_2 - t_2$  uniformly bounded by a constant dependent on  $C$  alone. The details are again omitted.  $\square$

LEMMA 2.6. *There is a first  $t \cong \tau(\psi) \cong t_3(\psi) + r > 2r$  at which  $\dot{x}(\tau) = 0$ . As a function of  $\psi$ ,  $\tau$  is completely continuous on  $0 < \|\psi\| \leq C$ .*

*Proof.* The existence of a first positive maximum for  $x$  for  $t > t_3 + r$  can be shown using an argument similar to that following Lemma 2.1. Variants of Lemmas 2.2 and 2.3 may be obtained to show  $\tau(\psi)$  uniformly bounded for  $0 < \|\psi\| \leq C$ . Continuity follows from continuous dependence, as usual.  $\square$

Using the same line of argument as that presented after Lemma 2.1, we see that  $(x(\psi)(\tau(\psi) + \cdot), y_n(\tau), y_{n-1}(\tau), \dots, y_0(\tau)) \in K$ . We define  $A: K \rightarrow K$  by  $A\psi = (x(\psi)(\tau(\psi) + \cdot), y_n(\tau), \dots, y_0(\tau))$  and  $A0 = 0$ . ( $A$  is, in general, not continuous at 0.) Clearly, any nonzero fixed point  $\psi$  of  $A$  defines a nontrivial periodic solution of (1.9) with period  $\tau(\psi) > 2r$ .

**3. Existence of a fixed point.** Following the procedure outlined in Hale [7], we intend to apply the following fixed point theorem of Nussbaum.

THEOREM 3.1, [7]. *Assume  $K$  is a closed convex subset of a Banach space  $X$ ,  $A: K \setminus \{0\} \rightarrow K$  is completely continuous,  $0 \in K$  is an ejective point of  $A$  and there is an  $M > 0$  such that  $\|\psi\| = M$  and  $A\psi = \delta\psi$  implies  $\delta < 1$ . Then  $A$  has a fixed point in  $0 < \|\psi\| < M$  if either  $K$  is infinite dimensional or  $0$  is an extreme point of  $K$ .*

Recall that  $0 \in K$  is an ejective point of  $K$  provided there exists an open neighborhood  $\mathcal{O}$  of  $0$  in  $K$  such that for any  $\psi \in K \cap \mathcal{O} \setminus \{0\}$ ,  $A^m\psi \notin \mathcal{O}$  for some positive integer  $m = m(\psi)$ . Ejectivity of  $0$  is the most complicated hypothesis of Theorem 3.1 to verify. For this condition we will use a technique of Chow and Hale [2].

If (1.9) is linearized about the zero solution one has

$$(3.1) \quad \dot{z}(t) = A_1 z(t) + A_2 z(t-r),$$

where  $z(t) \in \mathbb{R}^{n+2}$  and

$$A_1 = \begin{bmatrix} 0 & -\alpha & 0 & 0 & \dots & 0 \\ 0 & -\beta & \beta & 0 & \dots & 0 \\ 0 & 0 & -\beta & \beta & \dots & 0 \\ \vdots & & & & & \vdots \\ 0 & \dots & 0 & -\beta & \beta & \\ 0 & \dots & 0 & 0 & -\beta & \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & & \vdots \\ 0 & 0 & & \\ \beta & 0 & \dots & 0 \end{bmatrix}$$

are both  $(n+2) \times (n+2)$  matrices. An elementary calculation shows that any root of (1.3) with  $\text{Re } \lambda > -\beta$  is a root of the characteristic equation associated with (3.1). Thus, under the condition  $\alpha > \alpha_0$ , we may use the root  $\lambda$  with positive real part to decompose  $X = P(\lambda) \oplus Q(\lambda)$ , where  $P(\lambda)$  and  $Q(\lambda)$  are invariant under the usual solution operator  $T(t): X \rightarrow X$ ,  $T(t)\psi = (x(t+\cdot), y_n(t), \dots, y_0(t))$ ,  $t \geq 0$ . The restriction of  $T(t)$  to  $P(\lambda)$  has only the eigenvalue  $\{e^{\lambda t}\}$ . The projection  $\Pi(\lambda): X \rightarrow P(\lambda)$  is described in Hale [7].

LEMMA 3.2. Assume  $\alpha > \alpha_0$  and  $\lambda = \mu + i\nu$  is the root given in Lemma 1.2. Then  $\inf \{ \|\Pi(\lambda)\psi\| \mid \psi \in K, \|\psi\| = 1 \} > 0$ .

Proof. The adjoint equation associated with (3.1) is

$$(3.2) \quad \dot{w}(t) = -w(t)A_1 - w(t+r)A_2,$$

where  $w(t)$  is a  $n + 2$  dimensional row vector. Since  $\lambda$  is a characteristic root,  $w = e^{-\lambda t}b(0 \neq b^T \in \mathbb{R}^{n+2})$  will solve (3.2). Substituting into (3.2) we find that the equations may be solved recursively for  $b = (1, (\lambda/\beta)(\lambda/\beta + 1)^n e^{r\lambda}, (\lambda/\beta)(\lambda/\beta + 1)^{n-1} e^{r\lambda}, \dots, (\lambda/\beta) e^{r\lambda})$ .

Let  $\psi = (\varphi(\cdot), 0, y_{n-1}(0), \dots, y_0(0)) \in K$ . Since  $\lambda$  is a simple root,  $\|\Pi(\lambda)\psi\|$  is a nonzero constant multiple of  $|(w, \psi)|$ , where

$$(3.3) \quad \begin{aligned} (w, \psi) &= w(0)\psi(0) + \int_0^r w(s)A_2\psi(s-r) ds \\ &= b\psi(0) + \int_0^r b e^{-\lambda s} (0, 0, \dots, 0, \beta\varphi(s-r))^T ds \\ &= \varphi(0) + \lambda e^{r\lambda} \int_0^r e^{-\lambda s} \varphi(s-r) ds + \frac{\lambda}{\beta} e^{r\lambda} \sum_{k=0}^{n-1} \left(\frac{\lambda}{\beta} + 1\right)^k y_k(0). \end{aligned}$$

Case 1:  $r > 0$ . Let  $u = s - r$  in the integral, and ( $\varphi$  being of bounded variation) integrate by parts to find

$$(3.4) \quad (w, \psi) e^{-r\lambda} = \varphi(-r) + \int_{-r}^0 e^{-\lambda(u+r)} d\varphi(u) + \frac{\lambda}{\beta} \sum_{k=0}^{n-1} \left(\frac{\lambda}{\beta} + 1\right)^k y_k(0).$$

Note that since  $0 < r\nu < \pi$  and  $\varphi$  is nondecreasing

$$(3.5) \quad -\pi < -r\nu \leq \arg \int_{-r}^0 e^{-\lambda(u+r)} d\varphi(u) \leq 0.$$

Assume for the moment that each term  $(\lambda/\beta)(\lambda/\beta + 1)^k$  lies in the angular sector  $0 \leq \arg(\zeta) < \pi - r\nu - \varepsilon$  for some  $\varepsilon > 0$ . If  $\psi^{(l)}$  is a sequence from  $K$  (all with unit norm), we may deduce from  $y_k^{(l)}(0) \geq 0, \varphi(-r) \geq 0, (3.4)$  and (3.5) that  $y_k^{(l)}(0) \rightarrow 0, k = 0, 1, \dots, n - 1$  as well as

$$(3.6) \quad \varphi^{(l)}(-r) + \int_{-r}^0 e^{-\lambda(u+r)} d\varphi^{(l)}(u) \rightarrow 0.$$

(Note, that for  $\varepsilon > 0 \{ \zeta \in \mathbb{C} \mid -r\nu < \arg(\zeta) \leq r\nu - \varepsilon + \pi \}$  defines a cone  $\tilde{K}$  in  $\mathbb{C}$  with the property: if  $\zeta_j \in \tilde{K}; j = 1, \dots, p$  then  $|\sum_{j=1}^p \zeta_j| \geq \delta \sum_{j=1}^p |\zeta_j|$  for some  $\delta = \delta(\varepsilon) > 0$ .)

Consider the imaginary part of the integral. Since  $0 < r\nu < \pi$  and each  $\varphi^{(l)}$  is monotone increasing, we have for each  $-r < s \leq 0$ ,

$$\begin{aligned} 0 &\leq \varphi^{(l)}(0) - \varphi^{(l)}(s) \\ &= \int_s^0 d\varphi^{(l)}(u) \\ &\leq \frac{1}{d} \int_{-r}^0 e^{-\mu(u+r)} \sin(\nu(u+r)) d\varphi^{(l)}(u), \end{aligned}$$

where  $d = d(s) > 0$  is the minimum value of  $e^{-\mu(u+r)} \sin(\nu(u+r))$  on  $(s, 0]$ . The integral approaches zero as  $l \rightarrow +\infty$ , thus the monotonicity of  $\varphi^{(l)}$  shows  $\varphi^{(l)}(s) \rightarrow \varphi^{(l)}(0) = 1$  on

compact subsets of  $(-r, 0]$ . Thus (from 3.3)  $(w, \psi^{(l)}) \rightarrow 1 + \lambda e^{r\lambda} \int_0^r e^{-\lambda s} ds \neq 0$ —a contradiction.

Finally, we must show for each  $k = 0, \dots, n - 1$ , that  $(\lambda/\beta)(\lambda/\beta + 1)^k$  lies in the sector  $0 < \arg(\zeta) < \pi + r\nu - \varepsilon$  for some  $\varepsilon > 0$ . Fix  $k$  and consider  $\theta(\sigma) = \arg((\zeta/\beta)(\zeta/\beta + 1)^k)$  for  $\zeta = \mu + i\sigma$  as  $\sigma$  increases from 0 to  $\nu$ . When  $\sigma = 0$ , clearly  $\theta = 0$ , while as  $\sigma$  increases  $0 < \arg(\zeta/\beta) \cong \theta(\sigma) < \arg((\zeta/\beta)(\zeta/\beta + 1)^n)$ . Thus

$$\begin{aligned} 0 < \arg\left(\frac{\mu + i\nu_0}{\beta}\right) &\cong \theta(\nu) < \arg\left(\frac{\lambda}{\beta}\left(\frac{\lambda}{\beta} + 1\right)^n\right) \\ &= \arg\left(\frac{\lambda}{\beta}\left(\frac{\lambda}{\beta} + 1\right)^{n+1}\right) - \arg\left(\frac{\lambda}{\beta} + 1\right) \\ &= \arg(-\alpha e^{-r\lambda}) - \arg\left(\frac{\lambda}{\beta} + 1\right) \\ &\cong \pi = r\nu - \arg\left(\frac{\mu + \nu_0}{\beta} + 1\right) \end{aligned}$$

using (1.3) and  $\nu \cong \nu_0$  (from Lemma 1.2).

Case 2:  $r = 0$ . Here, (3.3) reduces to

$$(3.7) \quad (w, \psi) = \varphi(0) + \sum_{k=0}^{n-1} \frac{\lambda}{\beta} \left(\frac{\lambda}{\beta} + 1\right)^k y_k(0).$$

Arguing as in the previous case we find that each  $(\lambda/\beta)(\lambda/\beta + 1)^k$  lies in the sector  $0 < \arg(\zeta) \cong \pi - \varepsilon$  for some  $\varepsilon > 0$ . Since  $\varphi(0) = 1$ , each term on the right side of (3.7) lies in the sector  $0 \cong \arg(\zeta) \cong \pi - \varepsilon$ , and the techniques of the previous case apply here as well.  $\square$

One final lemma is needed.

LEMMA 3.3. Define  $\eta > 0$  to be the unique solution of

$$\frac{1}{2} = \int_{-\infty}^{-\eta} \rho_n(u+r) du = \int_{-\eta}^{-r} \rho_n(u+r) du.$$

Then, for any  $0 \neq \psi \in K$ ,  $x(\tau(\psi)) < 2e^{\alpha\eta} - 1$ .

Proof. Without loss of generality we may assume  $x(\tau) > 1$ . However,  $x(t) \cong 1$  cannot hold on  $[\tau - \eta, \tau]$  since, if so, we arrive at the contradiction

$$\begin{aligned} \dot{x}(\tau) &= -\alpha[x(\tau) + 1] \left[ \sum_{k=0}^n \frac{y_k(0)}{(n-k)!} (\beta\tau)^{n-k} e^{-\beta\tau} + \int_{-r}^0 \varphi(u)\rho_n(u-\tau+r) du \right. \\ &\quad \left. + \int_0^{\tau-\eta} x(u)\rho_n(u-\tau+r) du + \int_{\tau-\eta}^{\tau-r} x(u)\rho_n(u-\tau+r) du \right] \\ &< -\alpha[x(\tau) + 1] \left[ 0 + 0 - \int_0^{\tau-\eta} \rho_n(u-\tau+r) du + \int_{\tau-\eta}^{\tau-r} \rho_n(u-\tau+r) du \right] \\ &< -\alpha[x(\tau) + 1] \left[ - \int_{-\infty}^{-\eta} \rho_n(s+r) ds + \int_{-\eta}^{-r} \rho_n(s+r) ds \right] \\ &= 0. \end{aligned}$$

Let  $t^* \in [\tau - \eta, \tau]$  satisfy  $x(t^*) = 1$ .

The previous inequalities can be modified to show  $\dot{x}(t) < \alpha[x(t) + 1]$  for  $t^* \leq t \leq \tau$ .

Thus

$$\ln(x(\tau) + 1) - \ln(x(t^*) + 1) < \alpha(\tau - t^*)$$

and

$$\ln(x(\tau) + 1) - \ln(2) < \alpha\eta. \quad \square$$

*Proof of Theorem 1.1.* We show that  $K$  and  $A$  satisfy the hypotheses of Theorem 3.1. Clearly  $K$  is closed and convex. When  $r > 0$ ,  $K$  is infinite dimensional and when  $r = 0$ ,  $\psi = 0$  is an extreme point of  $K$ .

Since  $\|A\psi\| = x(\tau(\psi))$ , we may take  $M = 2e^{\eta\alpha} - 1$ , the upper bound derived in the previous lemma. The continuity of  $A$  on  $K \setminus \{0\}$  is a consequence of continuous dependence. The previous lemma implies  $A(K \setminus \{0\})$  is uniformly bounded so that complete continuity follows in the case  $r = 0$  from the finite dimensionality of  $K$ . For  $r > 0$  we note that on  $[\tau(\psi) - r, \tau(\psi)]$ ,  $\dot{x}(t) \leq \alpha[x(t) + 1] < 2\alpha e^{\eta\alpha}$ , and apply the Ascoli-Arzelà lemma.

Finally, the ejectiveity of 0 follows from Lemmas 1.2, 2.6, 3.2 and the complete continuity of  $A$ . (See [7, Thm. 11.2.3].)  $\square$

Using the bound  $M$  derived in the previous theorem we can summarize as follows:

**COROLLARY 3.4.** Assume  $\beta > 0$ ,  $n \geq 0$ ,  $r \geq 0$  (with not both  $r$  and  $n$  zero). Define  $\alpha_0$  as in Theorem 1.1 and  $\eta$  as in Lemma 3.3 For  $\alpha > \alpha_0$  there is a nontrivial periodic solution of (1) with period  $> 2r$  and  $0 < N(t) < 2N_0 e^{\alpha\eta}$ .

#### REFERENCES

- [1] W. ALT, *Some periodicity criteria for functional differential equations*, Manuscripta Math. 23 (1978), pp. 295–318.
- [2] S.-N. CHOW AND J. K. HALE, *Periodic solution to autonomous equations*, J. Math. Anal. Appl., 3 (1978), pp. 495–506.
- [3] S.-N. CHOW AND J. MALLET-PARET, *Integral averaging and bifurcation*, J. Differential Equations, 26 (1977), pp. 112–159.
- [4] J. M. CUSHING, *Integro-differential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomathematics, 20, Springer-Verlag, New York,
- [5] G. M. DUNKEL, *Some mathematical models for population growth with lags*, Tech. Note, IFDAM, Univ. of Maryland, College Park, MD, 1967.
- [6] R. B. GRAFTON, *A periodicity theorem for autonomous functional differential equations*, J. Differential Equations, 6 (1969), pp. 87–109.
- [7] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1977.
- [8] ———, *Nonlinear oscillations in equations with delays*, preprint.
- [9] S. HASTINGS, J. TYSON AND D. WEBSTER, *Existence of periodic solutions for negative feedback cellular control systems*, J. Differential Equations, 25 (1977), pp. 39–64.
- [10] H. HETHCOTE, H. STECH AND P. VAN DEN DRIESSCHE, *Non-linear oscillations in epidemic models*, preprint.
- [11] G. S. JONES, *The existence of periodic solutions of  $f'(x) = -\alpha f(x-1)[1+f(x)]$* , J. Math. Anal. Appl., 5 (1962), pp. 435–450.
- [12] N. D. KAZARINOFF, Y.-H. WAN AND P. VAN DEN DRIESSCHE, *Hopf bifurcation and stability of periodic solutions of differential-difference and integro-differential equations*, J. Inst. Math. Appl., 21 (1978), pp. 461–478.
- [13] N. MACDONALD, *Time delay in prey-predator models*, Math. Biosci., 28 (1976), pp. 321–330.
- [14] R. M. MAY, *Stability and Complexity in Model Ecosystems*, 2nd edition, Princeton University Press, Princeton, NJ, 1974.
- [15] R. D. NUSSBAUM, *Periodic solutions of some nonlinear autonomous functional differential equations*, Ann. Mat. Pura. Appl., 101 (1974), pp. 263–306.
- [16] A. SOMOLINOS, *Periodic solutions of the sunflower equation:  $\ddot{x} + (a/r)\dot{x} + (b/r)\sin(x(t-r)) = 0$* , Quart. Appl. Math., 4 (1978), pp. 465–478.
- [17] O. J. STAFFANS, *Nonlinear Volterra integral equations with positive definite kernels*, Proc. Amer. Math. Soc., 51 (1975), pp. 103–108.

- [18] H. W. STECH, *The effect of time lags on the stability of the equilibrium state of a population growth equation*, J. Math. Biol., 5 (1978), pp. 115–120.
- [19] ———, *The Hopf bifurcation: A stability result and application*, J. Math. Anal. Appl., in press.
- [20] H. O. WALTHER, *Existence of a non-constant periodic solution of a non-linear autonomous functional differential equation representing the growth of a single species population*, J. Math. Biol., 1 (1975), pp. 227–240.
- [21] A. WÖRZ-BUSEKROS, *Global stability in ecological systems, with continuous time delay*, SIAM J. Appl. Math., 35 (1978), pp. 123–134.

# A BIFURCATION APPLICATION OF THE GENERALIZED INVERSE OF A LINEAR DIFFERENTIAL OPERATOR\*

W. S. LOUD†

**Abstract.** If  $L$  is a linear differential operator, and if  $E$  and  $F$  are projections such that the range of  $F$  is the range of  $L$  while the null space of  $E$  is the null space of  $L$ , there is a unique generalized inverse  $X$  of  $L$  such that  $LXL = L$ ,  $XLX = X$ ,  $LX = F$  and  $XL = E$  (restricted to the domain of  $L$ ). The freedom of choice of the projections  $E$  and  $F$  leads to simplifications in the analysis of branching problems for solutions of nonlinear boundary problems. This is illustrated with two examples.

**1. Introduction.** In recent years there has been much activity in the area of generalized inverses of linear operators. See the books of Ben-Israel and Greville [1] and Nashed [10], and the references given therein. For the case of a regular differential operator the generalized inverse leads to generalized Green's functions, an idea which goes back to Hilbert. See also the work of Reid [11], [12], [13].

The most common generalized inverse of an operator  $L$  is the Moore-Penrose pseudoinverse  $L^\dagger$ . It is characterized by the relations  $LL^\dagger L = L$ ,  $L^\dagger LL^\dagger = L^\dagger$ , and by the requirements that  $L^\dagger L$  and  $LL^\dagger$  be orthogonal projections. A larger class of generalized inverses, denoted in [1] as (1-2)-generalized inverses is characterized by the first two of the requirements for  $L^\dagger$ ; but if  $X$  is the generalized inverse, the projections  $XL$  and  $LX$  are arbitrary except that the range of  $LX$  must be the range of  $L$ , while the null space of  $XL$  must be the null space of  $L$ . It turns out that the freedom of choice of these projections can lead to simplification of calculations in some applications.

In § 2 we give a Hilbert space setting for the generalized inverses. Section 3 is a collection of known results in the matrix case; one lemma needed for a later application is given. Section 4 is a collection of known results for differential operators. As an example the generalized Green's matrix corresponding to the group inverse is given for a differential operator.

In § 5 we study a branching problem for solutions of a nonlinear boundary-value problem involving a differential equation. This leads to an equation  $Lx = Nx$ , where  $L$  is of necessity a noninvertible linear operator and  $N$  is a nonlinear operator with no linear terms. To obtain the determining equation for the branching problem, a generalized inverse  $X$  of  $L$  can be used. It turns out, that by a judicious choice of the projection  $LX$ , the calculations can be simplified. Section 6 consists of the discussion of two examples to illustrate the use of the generalized inverse with branching problems.

**2. Abstract setting.** Let  $\mathcal{H}$  be a separable Hilbert space, and let  $L$  be a densely defined, closed linear operator in  $\mathcal{H}$ , such that the null spaces of  $L$  and of its adjoint  $L^*$  are both finite-dimensional. This will imply that the ranges of  $L$  and of  $L^*$  are closed subspaces. The reason for this last assumption is that we have regular ordinary differential operators in mind.

An operator  $X$  with domain  $\mathcal{H}$  will be called a (1-2)-generalized inverse of  $L$  if the range of  $X$  is contained in the domain of  $L$  and if

$$(1) \quad LXL = L,$$
$$(2) \quad XLX = X.$$

\* Received by the editors April 28, 1978, and in final revised form September 18, 1979. This research was supported in part by the U.S. Army Research Office under Grant DA-AROD-31-124-73-G199.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.



The relations (1) and (2) imply that  $LX$  and  $XL$  are idempotents.  $LX$  is a projection operator, and the range of  $LX$  is the same as the range of  $L$ .  $XL$  is the restriction of a projection operator to the domain of  $L$ , and the null space of  $XL$  is the same as the null space of  $L$ .

If  $E$  and  $F$  are any projection operators such that the null space of  $E$  is the null space of  $L$ , and the range of  $F$  is the range of  $L$ , we shall construct the unique (1-2)-generalized inverse  $X$  such that  $XL = E$  (restricted to the domain of  $L$ ), and  $LX = F$ .

The study of the projections  $E$  and  $F$  will be simpler if we consider the supplementary projections  $P = I - E$  and  $Q = I - F$ . If the dimensions of the null spaces of  $L$  and  $L^*$  are  $k$  and  $k'$  respectively,  $P$  and  $Q$  are projection operators whose ranges have dimension  $k$  and  $k'$  respectively. Let  $(u_1, u_2, \dots, u_k)$  be a basis for the null space of  $L$ , which is the range of  $P$ . It is always possible to choose a basis  $(r_1, r_2, \dots, r_k)$  for the range of  $P^*$  such that

$$(r_i, u_j) = \delta_{ij}, \quad i, j = 1, \dots, k.$$

If this is done, the projection  $P$  is given by the formula

$$(2.1) \quad Px = \sum_{i=1}^k (x, r_i)u_i.$$

Similarly if  $(v_1, \dots, v_{k'})$  is a basis for the null space of  $L^*$ , which is the range of  $Q^*$ , a basis  $(s_1, \dots, s_{k'})$  can be found for the range of  $Q$  such that

$$(v_i, s_j) = \delta_{ij}, \quad i, j = 1, \dots, k',$$

and the projection  $Q$  is given by the formula

$$(2.2) \quad Qx = \sum_{i=1}^{k'} (x, v_i)s_i.$$

**THEOREM 2.1.** *Let  $E$  and  $F$  be projection operators such that the null space of  $E$  is the null space of  $L$  and the range of  $F$  is the range of  $L$ . Then there exists a unique (1-2)-generalized inverse of  $L$  such that  $LX = F$  and  $XL = E$  (restricted to the domain of  $L$ ).*

*Proof.* Let  $\underline{L}$  be any operator with domain  $\mathcal{H}$  such that  $\underline{L}\underline{L}\underline{L} = L$ . We claim that

$$(2.3) \quad X = \underline{E}\underline{L}F$$

is the desired generalized inverse. Note that the defining conditions on  $E$  and  $F$  imply that  $LE = FL = L$ . Hence  $LX = \underline{L}\underline{E}\underline{L}F = \underline{L}\underline{L}F$ , and  $XL = \underline{E}\underline{L}FL = \underline{E}\underline{L}\underline{L}$ . Now if  $y$  is any element of the range of  $L$ , the definition of  $\underline{L}$  gives that  $\underline{L}\underline{L}y = y$ . Since the range of  $F$  is the range of  $L$ , this means that for any vector  $z$ ,  $\underline{L}\underline{L}Fz = Fz$ , so that  $\underline{L}\underline{L}F = F$ . Also if  $x$  is in the domain of  $L$ ,  $\underline{L}\underline{L}x = x + u$ , where  $u$  is in the null space of  $L$ . Since  $u$  is in the null space of  $E$ , for any  $x$  in the domain of  $L$ ,  $\underline{E}\underline{L}\underline{L}x = Ex$ , so that  $\underline{E}\underline{L}\underline{L} = E$  (restricted to the domain of  $L$ ). This shows that  $LX = F$  and  $XL = E$  (restricted to the domain of  $L$ ).  $LXL = \underline{L}\underline{E}\underline{L}FL = \underline{L}\underline{L}\underline{L} = L$ , and  $XLX = \underline{E}\underline{L}F\underline{L}\underline{E}\underline{L}F = \underline{E}\underline{L}\underline{L}\underline{L}F = \underline{E}\underline{L}F = X$ , where we used  $FLE = L$  and  $\underline{L}\underline{L}F = F$ . This proves that  $X$  is a (1-2)-generalized inverse of  $L$  with the asserted properties.

To prove uniqueness, suppose that  $X$  and  $Y$  are such that  $XLX = X$ ,  $YLY = Y$ ,  $LX = LY = F$ ,  $XL = YL = E$  (restricted to the domain of  $L$ ). Then

$$X = XLX = XLY = YLY = Y.$$

This completes the proof.

There are many generalized inverses of  $L$  which satisfy  $LXL = L, LX = F, XL = E$  (restricted), but the above shows that there is only one which satisfies the additional requirement  $XLX = X$ . In what follows, the projections  $LX$  and  $XL$  are of principal interest, so it seems simplifying and not particularly restrictive to ask that  $XLX = X$ .

The two most important generalized inverses of  $L$  are the Moore–Penrose pseudo-inverse, for which  $LX$  and  $XL$  are Hermitian, and the group inverse, for which  $LX = XL$  on the domain of  $L$ . To obtain the Moore–Penrose pseudo-inverse,  $L^\dagger$ , let  $E$  be the orthogonal projection with range the range of  $L^*$ , and let  $F$  be the orthogonal projection with range the range of  $L$ . If  $(u_1, \dots, u_k)$  is an orthonormal basis for the null space of  $L, E = I - P$ , where

$$Px = \sum_{i=1}^k (x, u_i)u_i.$$

If  $(v_1, \dots, v_{k'})$  is an orthonormal basis for the null space of  $L^*, F = I - Q$ , where

$$Qx = \sum_{j=1}^{k'} (x, v_j)v_j.$$

If  $\underline{L}$  is any operator for which  $\underline{L}\underline{L}\underline{L} = \underline{L}$ ; we have for any vector  $x$

$$\begin{aligned} (2.4) \quad L^\dagger x &= \underline{E}\underline{L}Fx = (I - P)\underline{L}(I - Q)x = \underline{L}x - P\underline{L}x - \underline{L}Qx + P\underline{L}Qx \\ &= \underline{L}x - \sum_{i=1}^k (\underline{L}x, u_i)u_i - \sum_{j=1}^{k'} (x, v_j)\underline{L}v_j + \sum_{i=1}^k \sum_{j=1}^{k'} (x, v_j)(\underline{L}v_j, u_i)u_i. \end{aligned}$$

For the group inverse,  $L^\#$  to exist, we must have  $E = F$ ; which means that  $E$  and  $F$  are each equal to the projection with range, the range of  $L$ , and null space, the null space of  $L$ . This implies that  $k = k'$ , and that the range and null space of  $L$  have no nontrivial element in common. If  $(u_1, \dots, u_k)$  is a basis for the null space of  $L$ , a basis  $(v_1, \dots, v_k)$  can be found for the null space of  $L^*$  such that

$$(u_j, v_i) = \delta_{ij}.$$

We then have for any vector  $x$ ,

$$Px = Qx = \sum_{i=1}^k (x, v_i)u_i.$$

With these bases and with  $\underline{L}$  as before we have

$$(2.5) \quad L^\# x = \underline{L}x - \sum_{i=1}^k (\underline{L}x, v_i)u_i - \sum_{j=1}^k (x, v_j)\underline{L}u_j + \sum_{i=1}^k \sum_{j=1}^k (x, v_j)(\underline{L}u_j, v_i)u_i.$$

The question as to when the range of  $X$  is contained in the range of  $L$  is of interest. Since the range of  $X$  is the same as the range of  $XL$ , this implies that the range of  $E$  must be contained in the range of  $F$ , and since  $E$  and  $F$  are projections this gives  $FE = E$ . If we use  $P = I - E$  and  $Q = I - F$ , the requirement becomes  $QP = Q$ . Introduce the bases  $(r_1, \dots, r_k)$  and  $(s_1, \dots, s_{k'})$  for the ranges of  $P^*$  and  $Q$  used earlier. Formulas (2.1) and (2.2) give

$$QPx = \sum_{i=1}^{k'} \sum_{j=1}^k (x, r_j)(u_j, v_i)s_i, \quad Qx = \sum_{i=1}^{k'} (x, v_i)s_i.$$

The linear independence of the  $s_i$  gives that  $QPx = Qx$  implies

$$(x, v_i) = \sum_{j=1}^k (x, r_j)(u_j, v_i), \quad i = 1, \dots, k',$$

and since this holds for all  $x$ ,

$$(2.6) \quad v_i = \sum_{j=1}^k (v_i, u_j)r_j, \quad i = 1, \dots, k'.$$

This shows that the choice of the  $s_i$  is not significant, but that the  $r_j$  must be properly chosen.

Let  $M$  be the  $k' \times k$  matrix of which the  $i - j$  element is  $(v_i, u_j)$ . We claim that (2.6) can be solved for the  $r_j$  if and only if the rank of  $M$  is  $k'$ , so that *the desired (1-2)-generalized inverse with the range of  $X$  contained in the range of  $L$  exists if and only if the rank of  $M$  is  $k'$ .*

If the rank of  $M$  is less than  $k'$ , there is a nontrivial set of constants  $c_1, \dots, c_{k'}$  such that the corresponding linear combination of the rows of  $M$  is the zero row-vector. But then (2.6) would imply

$$c_1v_1 + \dots + c_{k'}v_{k'} = 0$$

which contradicts the linear independence of the  $v_i$ .

If the rank of  $M$  is  $k'$  and  $k' < k$ , we can solve (2.6) for some set of  $k'$   $r$ 's, with the and null space of  $L$  having no nontrivial element in common. In this case, the  $r_j$  are determined uniquely. The projections  $P$  and  $E = I - P$  are then the same as  $P$  and  $E$  as found for the group inverse.

If the rank of  $M$  is  $k'$  and  $k' < k$ , we can solve (2.6) for some set of  $k'$   $r$ 's, with the remaining  $r$ 's arbitrary. In this case the projections  $P$  and  $E$  can be determined in many ways so that the range of  $X$  is contained in the range of  $L$ . The projection  $F$  is restricted only by the requirement that the range of  $F$  be the range of  $L$ .

**3. The matrix case.** Let  $A$  be an  $m \times n$  matrix of rank  $r$ . In this case the generalized inverse is an  $n \times m$  matrix which satisfies

$$AXA = A, \quad XAX = X.$$

In this section we list the properties of (1-2)-generalized inverses of matrices needed in the application to linear ordinary differential operators. All the results presented in this section are either well-known or easily derived. The reader is referred to [1] or [10] for comprehensive discussions of generalized inverses of matrices.

If  $E$  is an  $n \times n$  projection matrix of rank  $r$ , such that the null space of  $E$  is the same as the null space of  $A$ , and if  $F$  is an  $m \times m$  projection matrix of rank  $r$ , such that the range of  $F$  is the same as the range of  $A$ ; there is a unique (1-2)-generalized inverse of  $A$  for which  $AX = F$  and  $XA = E$ . Indeed, if  $\underline{A}$  is any  $n \times m$  matrix such that  $\underline{A}\underline{A}A = A$ , then  $X = \underline{E}\underline{A}F$ ; the result being the same for every possible choice of  $\underline{A}$ .

Other formulas for  $X$  can be obtained with the use of full-rank factorizations. Let  $A$  have the full-rank factorization

$$(3.1) \quad A = HK^*,$$

where  $H$  is an  $m \times r$  matrix of rank  $r$  whose columns span the range of  $A$ , which is also the range of  $F$ , and where  $K$  is an  $n \times r$  matrix of rank  $r$  whose columns span the range of  $A^*$ , which is also the range of  $E^*$ . Let  $N_1$  be an  $n \times r$  matrix of rank  $r$  whose columns span the range of  $E$ . Then  $K^*N_1$  is a nonsingular  $r \times r$  matrix. If  $N = N_1(K^*N_1)^{-1}$ ,  $N$  is also an  $n \times r$  matrix of rank  $r$  whose columns span the range of  $E$  with the additional property that  $K^*N = I_r$ . We then have  $E = NK^*$ . Similarly let  $M_1$  be an  $m \times r$  matrix of rank  $r$  whose columns span the range of  $F^*$ . If  $M = M_1(H^*M_1)^{-1}$ ,  $M$  is an  $m \times r$  matrix of rank  $r$  whose columns span the range of  $F^*$  and  $M^*H = I_r$ . We then have  $F = HM^*$ . It is then readily verified that  $X = NM^*$  is the required (1-2)-generalized inverse of  $A$  with  $XA = E$  and  $AX = F$ .

When  $X$  is the Moore–Penrose pseudo-inverse,  $E$  and  $F$  are Hermitian. Choose  $N_1 = K$  and  $M_1 = H$ . The result is ([1, (1.25)])

$$A^\dagger = K(K^*K)^{-1}(H^*H)^{-1}H^*.$$

When  $X$  is the group inverse,  $E = F$ . This requires  $m = n$  so that  $A$  is a square matrix. The common value of  $E$  and  $F$  must be a projection with range the same as the range of  $A$ , and null space the same as the null space of  $A$ . Such a projection will exist if and only if the range and null space of  $A$  have no nontrivial vector in common, and this is equivalent to  $A$  and  $A^2$  having the same rank, and also to  $K^*H$  being a nonsingular  $r \times r$  matrix. We select  $M_1 = K$  and  $N_1 = H$  and obtain ([1, (4.12)])

$$A^\# = H(K^*H)^{-2}K^*, \quad AA^\# = A^\#A = H(K^*H)^{-1}K^*.$$

The requirement that the range of  $X$  be the same as the range of  $A$  makes  $m = n$ . Also the range of  $X$ , which is the range of  $E$ , must also be the range of  $A$ , while the null space of  $E$  must be the null space of  $A$ . We again must have  $K^*H$  nonsingular and

$$E = H(K^*H)^{-1}K^*.$$

We may still use  $F = HM^*$ , so that

$$X = H(K^*H)^{-1}M^*.$$

A possible choice for  $M^*$  is  $(K^*H)^{-1}K^*$ , which gives  $X = A^\#$ .

For later applications we need the following lemma.

**LEMMA 3.1.** *Let  $A$  be an  $m \times n$  matrix of rank  $r$ , and let  $E$  and  $F$  be projections such that the null space of  $E$  is the null space of  $A$ , and the range of  $F$  is the range of  $A$ . Then there exists an  $n \times m$  matrix  $R$ , having full rank, such that  $RA = E$  and  $AR = F$ .*

*Proof.* Let the four projections  $E, F, P = I_n - E$ , and  $Q = I_m - F$  have the full-rank factorizations

$$E = NK^*, \quad F = HM^*, \quad P = N_1K_1^*, \quad Q = H_1M_1^*.$$

Here  $A = HK^*$ ,  $M$  and  $N$  are as earlier,  $K_1^*N_1 = I_{n-r}$ ,  $M_1^*H_1 = I_{m-r}$ , and, for example,  $N_1$  is an  $n \times (n - r)$  matrix whose columns span the range of  $P$ , which is the null space of  $E$ . Then if  $G$  is any matrix of size  $(n - r) \times (m - r)$  having full rank, i.e. of rank equal to  $\min(n - r, m - r)$ , we may choose

$$R = NM^* + N_1GM_1^*.$$

Because  $K^*N_1$  and  $M_1^*H$  are zero, it follows at once that  $RA = E$  and  $AR = F$ .

To show that  $R$  has full rank, i.e. rank equal to  $\min(m, n)$ , let  $G_1$  be an  $(m - r) \times (n - r)$  matrix of full rank, chosen so that  $GG_1 = I_{n-r}$ , if  $n \leq m$ , and  $G_1G = I_{m-r}$ , if  $m \leq n$ . Then if

$$R_1 = HK^* + H_1G_1K_1^*,$$

it is readily verified that  $RR_1 = I_n$  if  $n \leq m$ , and  $R_1R = I_m$  if  $m \leq n$ . This shows that both  $R$  and  $R_1$  have full rank and completes the proof.

*Remarks.* (1) If  $m = n$ ,  $R$  is a nonsingular  $n \times n$  matrix. (2) If  $A$  is of full rank, the term with  $G$  is not present,  $R = NM^*$ , the (1-2)-generalized inverse found earlier, and  $R_1 = A$ .

**4. Differential operators.** In this section we apply the ideas of § 2 to the case that  $L$  is a differential operator. This leads to the notion of generalized Green's functions and generalized Green's matrices (see [1], [10], [11], [12] [13]). In [9] the author gave a

number of computational techniques for the case of the Moore–Penrose pseudo-inverse. In this section these ideas are extended to the wider class of generalized inverses discussed in § 2.

Let  $x$  be an  $n$ -vector, and let

$$(4.1) \quad lx = x' - A(t)x$$

be a differential expression. In (4.1)  $A(t)$  is a continuous  $n \times n$  matrix defined for  $a \leqq t \leqq b$ , and we consider  $x(t)$  as an element of  $L^2(a, b)$ . The Lagrange adjoint of (4.1) is the expression

$$(4.2) \quad l^+x = -x' - A(t)^*x.$$

The domain of the operator  $L$  consists of all absolutely continuous functions in  $L^2(a, b)$  for which  $x'$  is also in  $L^2(a, b)$ , and which satisfy the boundary conditions

$$(4.3) \quad Ax(a) + Bx(b) = 0.$$

In (4.3)  $A$  and  $B$  are  $m \times n$  matrices ( $0 \leqq m \leqq 2n$ ) such that the  $m \times 2n$  matrix  $A : B$  has rank  $m$ . The domain of the adjoint operator  $L^*$  will be the same set of functions, but which satisfy the adjoint boundary conditions

$$(4.4) \quad Mx(a) + Nx(b) = 0.$$

In (4.4)  $M$  and  $N$  are  $(2n - m) \times n$  matrices, and  $M : N$  has rank  $2n - m$ . The two boundary conditions are related by  $AM^* = BN^*$ .

For each  $x$  in the domain of  $L$ ,  $Lx = lx$ , and for each  $x$  in the domain of  $L^*$ ,  $L^*x = l^+x$ . The inner product in  $L^2(a, b)$  is given by

$$(x, y) = \int_a^b y(t)^*x(t) dt.$$

For each  $x$  in the domain of  $L$  and for each  $y$  in the domain of  $L^*$ ,

$$(Lx, y) = (x, L^*y).$$

Both of  $L$  and  $L^*$  have finite-dimensional null spaces of dimensions  $k$  and  $k'$  respectively. The dimensions  $k$  and  $k'$  are restricted by  $0 \leqq k \leqq n$ ,  $0 \leqq k' \leqq n$ ,  $k - k' = n - m$ . The range of  $L$  is the orthogonal complement of the null space of  $L^*$ , and the range of  $L^*$  is the orthogonal complement of the null space of  $L$ . Thus both ranges are closed subspaces with finite codimension.

A (1-2)-generalized inverse of  $L$  is an integral operator  $X$  with kernel  $G_X(t, s)$ . For our case we let  $P$  and  $Q$  be arbitrary projections such that the range of  $P$  is the null space of  $L$  and the range of  $Q^*$  is the null space of  $L^*$ . Finally let  $\underline{L}$  be an integral operator with kernel  $\underline{G}(t, s)$  such that  $\underline{L}\underline{L}\underline{L} = L$ . From (2.3) we have

$$X = (I - P)\underline{L}(I - Q),$$

and if  $G_P(t, s)$  and  $G_Q(t, s)$  are the kernels of the integral projection operators  $P$  and  $Q$ , we have for  $G_X(t, s)$  the formula

$$(4.5) \quad G_X(t, s) = \underline{G}(t, s) - \int_a^b G_P(t, u)\underline{G}(u, s) du - \int_a^b \underline{G}(t, v)G_Q(v, s) dv + \int_a^b \int_a^b G_P(t, u)\underline{G}(u, v)G_Q(v, s) dv du.$$

This formula appears in [8] and [9] for the Moore–Penrose pseudo-inverse.

Calculations with (4.5) are somewhat tedious but are quite straight-forward. As an example of the result of such a calculation we give the kernel  $G^\#(t, s)$  for the group inverse  $L^\#$  of  $L$ , provided that it exists. For  $X$  to be the group inverse, we require  $P = Q$ . This will imply that  $k = k'$  and  $m = n$ , so that  $L$  and  $L^*$  each have  $n$  boundary conditions. The range of  $P$  must be the null space of  $L$  and the range of  $P^*$  must be the null space of  $L^*$ . For such a  $P$  to exist,  $L$  must satisfy the existence condition for a group inverse, namely, that the range and null space of  $L$  have no nontrivial common element.

Let  $\Phi(t)$  be a fundamental matrix of  $x' = A(t)x$ . Let  $L_0$  denote the matrix  $A\Phi(a) + B\Phi(b)$ , and let  $L_0^+$  denote the matrix  $M\Phi(a)^{*-1} + N\Phi(b)^{*-1}$ . Each of  $L_0$  and  $L_0^+$  is  $n \times n$  and has rank  $n - k$ . Let  $P_0$  denote the projection matrix of rank  $k$  with the range of  $P_0$  being the null space of  $L_0$  and the range of  $P_0^*$  being the null space of  $L_0^+$ . The existence of  $P_0$  follows from the existence condition for  $L^*$ . Finally, since the null space of  $I - P_0$  is the null space of  $L_0$ , it follows from Lemma 1 that there exists a nonsingular matrix  $R$  such that  $RL_0 = I - P_0$ .

It is then found that

$$G_P(t, s) = \frac{1}{b - a} \Phi(t)P_0\Phi(s)^{-1}$$

and

$$G^\#(t, s) = \Phi(t)\Phi(s)^{-1}|_{t>s} - \left[ \frac{1}{2} + \frac{t-s}{b-a} \right] \Phi(t)P_0\Phi(s)^{-1} - \Phi(t)RB\Phi(b)(I - P_0)\Phi(s)^{-1}.$$

As is shown in [9] for the Moore–Penrose pseudo-inverse, the Green’s matrix  $G_X(t, s)$  can be found from a differential equation. In fact, considered as a function of  $t$ ,  $G_X(t, s)$  is determined by the following four conditions:

- (i)  $lG(t, s) = -G_Q(t, s)$ , ( $a \leq t < s \leq b$ ,  $a \leq s < t \leq b$ ).
- (ii) As  $t$  increases through  $s$ ,  $G_X(t, s)$  has a jump discontinuity equal to the identity matrix;
- (iii)  $G_X(t, s)$  satisfies the boundary conditions of  $L$ ,
- (iv)  $\int_a^b G_P(t, u)G_X(u, s) du = 0$ .

Calculations with (4.6) are of the same order of complexity as with (4.5). It is not difficult to verify that the kernel  $G^\#(t, s)$  given above satisfies (4.6) and is determined by (4.6).

A similar discussion can be given when  $L$  is a scalar differential operator of order  $n$  (see [9] for details).

**5. Application to branching of solutions of boundary-value problems.** In the analysis of branching (or bifurcation) of solutions of boundary-value problems for ordinary differential equations there always occurs a noninvertible linear differential operator. The purpose of this section is to discuss the use of (1–2) generalized inverses in this analysis. It turns out that allowing a wider choice of generalized inverses, rather than restricting consideration to the Moore–Penrose pseudoinverse, allows for simplification of the nonlinear determining equations.

We consider a boundary-value problem associated with the differential equation

$$(5.1) \quad x' = F(t, x, \mu),$$

where  $x$  and  $F$  are  $n$ -vectors and  $\mu$  is an  $m$ -vector. In (5.1)  $F$  is assumed sufficiently regular in the sense that all partial derivatives of  $F$  needed in the discussion are assumed to exist and be continuous in an open region containing all required values of the variables. Associated with (5.1) we have the boundary condition

$$(5.2) \quad Ax(a) + Bx(b) = k,$$

where  $A$  and  $B$  are  $n \times n$  matrices such that the  $n \times 2n$  matrix  $A : B$  has rank  $n$ , and where  $k$  is an  $n$ -vector. We do not consider cases in which the number of rows in  $A$  and  $B$  is different from  $n$ .

For  $\mu = 0$  let the problem (5.1), (5.2) have a solution  $x_0(t)$ . We shall study the solutions of (5.1), (5.2) for small  $\mu$  which reduce to  $x_0(t)$  as  $\mu \rightarrow 0$ . The method is based on the Lyapunov-Schmidt procedure as used by Cesari and Hale in a number of papers (cf. [2], [3], [6]). The analysis is motivated by Hale's paper [6].

If the change of variables  $x = x_0(t) + y$  is made, the problem (5.1), (5.2) becomes

$$(5.3) \quad y' = F(t, x_0(t) + y, \mu) - F(t, x_0(t), 0),$$

$$(5.4) \quad Ay(a) + By(b) = 0,$$

and we seek solutions of (5.3), (5.4) for small  $\mu$  which tend to zero as  $\mu \rightarrow 0$ . We rewrite (5.3) in the form

$$(5.5) \quad y' = A(t)y + B(t, \mu)\mu + C(t, \mu)\mu y + D(t, y, \mu)y^2,$$

where

$$A(t) = F_x(t, x_0(t), 0) \text{ is an } n \times n \text{ matrix,}$$

$$B(t, \mu)\mu = F(t, x_0(t), \mu) - F(t, x_0(t), 0) \text{ is an } n\text{-vector}$$

$$C(t, \mu)\mu = F_x(t, x_0(t), \mu) - F_x(t, x_0(t), 0) \text{ is an } n \times n \text{ matrix,}$$

$$D(t, y, \mu)y^2 = F(t, x_0(t) + y, \mu) - F(t, x_0(t), \mu) - F_x(t, x_0(t), \mu)y \text{ is an } n\text{-vector.}$$

We note that  $B(t, 0) = F_\mu(t, x_0(t), 0)$ ,  $C(t, 0) = F_{x\mu}(t, x_0(t), 0)$  and  $D(t, 0, \mu) = \frac{1}{2}F_{xx}(t, x_0(t), \mu)$ .

Let  $L$  be the linear differential operator generated by the differential expression  $y' - A(t)y$  with boundary conditions (5.4). We then have

$$(5.6) \quad Ly = B(t, \mu)\mu + C(t, \mu)\mu y + D(t, y, \mu)y^2$$

to be solved for  $y$ , when  $\mu$  is small, with  $y \rightarrow 0$  as  $\mu \rightarrow 0$ .

When  $L$  is invertible, standard perturbation theory for ordinary differential equations shows that for small  $\mu$  there is a unique solution  $y = y(t, \mu)$  of (5.6) with  $y(t, 0) = 0$ . Thus branching of solutions will occur only when  $L$  is not invertible. When branching is present, the number of small solutions  $y(t, \mu)$  will depend on the value of  $\mu$ . For some  $\mu$  there may be several; for other  $\mu$  there may be none.

Equation (5.6) has the abstract form

$$(5.7) \quad Ly = N(y, \mu),$$

where  $\|N\| = O(\|\mu\| + \|\mu\|\|y\| + \|y\|^2)$  for small  $y$  and  $\mu$ . If  $y$  is a solution of (5.7), and if  $X$  is a (1-2)-generalized inverse of  $L$  as discussed in earlier sections,  $y$  also satisfies

$$(5.8) \quad y = XN(y, \mu) + u,$$

where  $u$  is an element of the null space of  $L$ . Because  $X$  is given by the formula  $X = (I - P)\underline{L}(I - Q)$ , and because the range of  $P$  is the null space of  $L$ , the term

$\underline{P}\underline{L}(I - Q)N(y, \mu)$  is in the null space of  $L$ , and so can be included in the term  $u$ . Thus we can write (5.8) as

$$(5.9) \quad y = \underline{L}(I - Q)N(y, \mu) + u.$$

Now whether or not  $y$  is a solution of (5.7), (5.9) can be solved, when  $\mu$  and  $u$  are small, for  $y$  as a function of  $\mu$  and  $u$ . If  $y(\mu, u)$  is the solution of (5.9), it will be a solution of (5.7) as well, provided that  $N(y(\mu, u), \mu)$  is in the range of  $L$ , i.e. if

$$(5.10) \quad QN(y(\mu, u), \mu) = 0.$$

Equation (5.10) is called the determining equation. If  $k$  is the common dimension of the null spaces of  $L$  and  $L^*$ , (5.10) is equivalent to a system of  $k$  nonlinear equations for the unknown  $u$ , which in turn depends on  $k$  parameters. It is possible to determine for which small  $\mu$  there exist solutions of (5.10) and how many solutions there are. For each solution  $u = u(\mu)$  of (5.10),  $y(\mu) = y(\mu, u(\mu))$  is a solution of (5.7).

In the computation just outlined, the form of (5.9) will depend on the choice of  $X$ , or more precisely, on the choice of the projection  $Q$ . The form of (5.10) will also depend on the choice of  $Q$ , but the values of  $\mu$  for which solutions exist, and the ultimate solutions  $y(\mu)$  of (5.7), will naturally be independent of the choice of  $Q$ . Hence the choice of the projection  $Q$  may possibly simplify both the form and the solution process of the determining equation.

We now apply this reasoning to the differential equation problem (5.6). In view of the remark following (5.8) we can use for a (1-2) generalized inverse of  $L$  the operator  $X = \underline{L}(I - Q)$  which is an integral operator with kernel

$$(5.11) \quad G_X(t, s) = \underline{G}(t, s) - \int_a^b \underline{G}(t, v)G_Q(v, s) dv.$$

If  $U(t)$  is an  $n \times k$  matrix, the columns of which span the null space of  $L$ , (5.9) takes the form

$$(5.12) \quad y(t) = \int_a^b G_X(t, s)[B(s, \mu)\mu + C(s, \mu)\mu y(s) + D(s, y(s), \mu)y(s)^2] ds + U(t)p,$$

where  $p$  is a  $k$ -vector and  $G_X(t, s)$  is given by (5.11). If  $\mu$  and  $p$  are sufficiently small, (5.12) has a unique solution  $y(t, \mu, p)$ . The function  $y(t, \mu, p)$  will be a solution of (5.6) provided that

$$B(t, \mu)\mu + C(t, \mu)\mu y(t, \mu, p) + D(t, y(t, \mu, p), \mu)y(t, \mu, p)^2$$

is in the range of  $L$ . If  $V(t)$  is an  $n \times k$  matrix whose columns span the null space of  $L^*$ , this requirement is equivalent to

$$(5.13) \quad \int_a^b V(t)^*[B(t, \mu)\mu + C(t, \mu)\mu y(t, \mu, p) + D(t, y(t, \mu, p), \mu)y(t, \mu, p)^2] dt = 0.$$

Equation (5.13) is the determining equation and is a system of  $k$  equations for the  $k$  components of  $p$  in terms of the  $m$ -vector  $\mu$ . Note that if (5.13) is written in the form  $K(\mu, p) = 0$ , then  $K(0, 0) = 0$  and  $K_p(0, 0) = 0$ . Equation (5.13) determines for small nonzero  $\mu$  how many different small values of the vector  $p$  exist for which  $y(t, \mu, p)$  is a solution of (5.6).

The following procedure can be used to obtain asymptotic expressions for small  $\mu$  of the various branching solutions of (5.6) which reduce to zero as  $\mu$  tends to zero. First determine an asymptotic expression for the solution of (5.12) in terms of  $\mu$  and  $p$ . Then use (5.13) to determine asymptotic expressions for  $p$  in terms of  $\mu$ . Use this last result in



the asymptotic expression for  $y(t, \mu, p)$  to obtain the desired asymptotic expressions for the solutions of (5.6).

**6. Examples.** In this section we give examples to illustrate the application of the foregoing ideas to branching of solutions of boundary-value problems.

*Example 1.* The boundary-value problem in this example occurs in the calculus of variations problem of minimal area of a surface of revolution.

$$(6.1) \quad xx'' = (1 + \mu)^2 + x'^2, \quad x'(0) = 0, \quad x(\xi) = \cosh \xi.$$

In (6.1)  $\xi$  is the unique positive root of the equation  $\coth u = u$  ( $\xi = 1.20$  approximately). When  $\mu = 0$ , (6.1) has the solution  $x_0(t) = \cosh t$ .

The substitution  $x = \cosh t + y$  transforms (6.1) to

$$(6.2) \quad \cosh t y'' - 2 \sinh t y' + \cosh t y = 2\mu + \mu^2 - yy'' + y'^2, \\ y'(0) = 0, \quad y(\xi) = 0.$$

Let the linear differential operator  $L$  be defined by

$$Ly = \cosh t y'' - 2 \sinh t y' + \cosh t y$$

with domain determined by the boundary conditions  $y'(0) = 0, y(\xi) = 0$ . The problem (6.2) is then

$$(6.3) \quad Ly = 2\mu + \mu^2 - yy'' + y'^2.$$

The functions  $\phi(t) = \cosh t - t \sinh t$  and  $\psi(t) = \sinh t$  are a linearly independent set of solutions of the differential equation  $Ly = 0$ . Since  $\phi(t)$ , but not  $\psi(t)$ , satisfies the boundary conditions, the null space of  $L$  is one-dimensional and is spanned by  $\phi(t)$ .

The adjoint operator  $L^*$  is given by

$$L^*y = \cosh t y'' + 4 \sinh t y' + 4 \cosh t y$$

with boundary conditions  $y'(0) = 0, y(\xi) = 0$ , the same as for  $L$ . A linearly independent set of solutions of  $L^*y = 0$  is  $\text{sech}^3 t \phi(t)$  and  $\text{sech}^3 t \psi(t)$ . The null space of  $L^*$  is one-dimensional, and is spanned by  $\text{sech}^3 t \phi(t)$ . We need two kernels,  $\underline{G}(t, s)$  and  $G_Q(t, s)$ . These are given by

$$\underline{G}(t, s) = [\phi(s)\psi(t) - \psi(s)\phi(t)] \text{sech}^3 s |_{t>s}, \\ G_Q(t, s) = \beta(t)\phi(s) \text{sech}^3 s,$$

where  $\beta(t)$  is normalized by the requirement  $\int_0^\xi \beta(t)\phi(t) \text{sech}^3 t dt = 1$ .  $Q$  is then a projection operator with the range of  $Q^*$  being the null space of  $L^*$ . We leave  $\beta(t)$  arbitrary otherwise so that it can be chosen to simplify the calculations. Any solution of (6.3) also satisfies

$$(6.4) \quad y(t) = \int_0^\xi G_X(t, s)[2\mu + \mu^2 - y(s)y''(s) + y'(s)^2] ds + p\phi(t)$$

for some constant  $p$ . In (6.4)  $G_X(t, s)$  is given by

$$G_X(t, s) = \underline{G}(t, s) - \int_0^\xi \underline{G}(t, v)G_Q(v, s) dv \\ = [\phi(s)\psi(t) - \psi(s)\phi(t)] \text{sech}^3 s |_{t>s} \\ - \int_0^t [\phi(v)\psi(t) - \psi(v)\phi(t)] \text{sech}^3 v \beta(v)\phi(s) \text{sech}^3 s dv.$$

When  $\mu$  and  $p$  are small, (6.4) has a unique solution given by

$$(6.5) \quad y(t, \mu, p) = p\phi(t) + 2\mu \int_0^\xi G_X(t, s) ds + \text{higher order terms.}$$

The integral in the second term of (6.5) is found to be

$$\int_0^t [\phi(s)\psi(t) - \psi(s)\phi(t)] \operatorname{sech}^3 s \left[ 1 - \frac{\xi}{2} \beta(s) \right] ds.$$

Because the constant  $2/\xi$  is an allowable choice for  $\beta(t)$ , we may make this choice and make the second term of (6.5) equal to zero so that

$$y(t, \mu, p) = p\phi(t) + \text{higher order terms.}$$

The function  $y(t, \mu, p)$  is a solution of (6.3) provided that

$$(6.6) \quad \int_0^\xi \phi(s) \operatorname{sech}^3 s [2\mu + \mu^2 - y(s, \mu, p)y''(s, \mu, p) + y'(s, \mu, p)^2] ds = 0.$$

This is the determining equation. Using the asymptotic expression found above for  $y(s, \mu, p)$ , (6.6) becomes

$$(6.7) \quad \mu\xi + \frac{1}{4}p^2\xi^2 + \text{higher order terms} = 0.$$

We see from (6.7) that for small negative  $\mu$  there are two values of  $p$ , approximately  $\pm(2/\xi)\sqrt{-\mu}$ , while for small positive  $\mu$  there are no values of  $p$ . As a result, the problem (6.1) has two solutions near  $x_0(t)$  for small negative  $\mu$  and none for small positive  $\mu$ .

Because the differential equation in (6.1) can be solved by elementary means, we may verify the above directly. The solutions of the differential equation for which  $x'(0) = 0$  are given by

$$x = \frac{1 + \mu}{c} \cosh ct,$$

where  $c$  is a constant. The boundary condition  $x(\xi) = \cosh \xi$  gives

$$\frac{\cosh \xi}{\xi} = (1 + \mu) \frac{\cosh c\xi}{c\xi},$$

and since the function  $\cosh u/u$  has a strict minimum at  $u = \xi$ , we must have  $\mu \leq 0$  for any solution of (6.1).

*Example 2.* We consider branching of periodic solutions of the system

$$(6.8) \quad x'' + g(x) = \mu f(t).$$

In (6.8)  $f(t)$  is a  $T$ -periodic continuous function, while  $g(x)$  is a "restoring force" term such that  $xg(x) > 0$  when  $x \neq 0$ . It is also assumed that  $g(x)$  is sufficiently regular in the sense given earlier. The solutions of the unperturbed equation  $x'' + g(x) = 0$  are periodic, and we assume that  $g(x)$  is such that the periods vary with amplitude. Note that this excludes the case that  $g(x)$  is linear. Let there exist a nonconstant periodic solution  $x_0(t)$  when  $\mu = 0$  with least period  $L_0$ , where  $L_0$  is a rational multiple of  $T$ , say  $qL_0 = pT$ , where  $p$  and  $q$  are relatively prime integers. To fix the phase of  $x_0(t)$ , we assume that at  $t = 0$ ,  $x_0 = A > 0$  and  $x'_0 = 0$ . Since not only  $x_0(t)$  but also all translations  $x_0(t + \tau)$  are  $L_0$ -periodic solutions when  $\mu = 0$ , it is appropriate to seek periodic solutions of (6.8) near to  $x_0(t + \tau)$  of period  $L = pT = qL_0$  for small nonzero  $\mu$ . This question is considered in [7].

If (6.8) has a periodic solution near to  $x_0(t + \tau)$ , the modified equation

$$(6.9) \quad x'' + g(x) = \mu f(t - \tau)$$

has a corresponding solution near to  $x_0(t)$ . It is more convenient to study (6.9).

In (6.9) make the substitution  $x = x_0(t) + y$ . This gives

$$x_0''(t) + y'' + g(x_0(t) + y) = \mu f(t - \tau),$$

which we write as

$$(6.10) \quad y'' + g'(x_0(t))y = \mu f(t - \tau) + H(t, y)y^2,$$

where

$$H(t, y)y^2 = -g(x_0(t) + y) + g(x_0(t)) + g'(x_0(t))y.$$

Note that  $H(t, 0) = -\frac{1}{2}g''(x_0(t))$ .

Let the linear differential operator  $L_0$  be defined by

$$L_0y = ly \equiv y'' + g'(x_0(t))y$$

with boundary conditions  $y(0) = y(L)$ ,  $y'(0) = y'(L)$ . The operator  $L_0$  is self-adjoint. Let  $\phi(t)$  and  $\psi(t)$  be those solutions of  $ly = 0$  with initial conditions  $\phi(0) = \psi'(0) = 1$ ,  $\phi'(0) = \psi(0) = 0$ . Since  $y = x_0'(t)$  is a solution of  $ly = 0$  with initial conditions  $x_0'(0) = 0$ ,  $x_0''(0) = -g(A)$ ; it follows that  $x_0'(t) = -g(A)\psi(t)$ . Therefore  $\psi(t)$  is  $L$ -periodic and so is in the null space of  $L_0$ . We assume that  $g(x)$  is such that  $\phi(t)$  is not periodic, and in particular that  $\phi'(L) \neq 0$ . The assumption that periods of solutions of  $x'' + g(x) = 0$  vary with amplitude will guarantee this (cf. [7]). Therefore the null space of  $L_0$  and the null space of  $L_0^* = L_0$  are one-dimensional and are spanned by  $\psi(t)$ .

The kernels needed for the (1-2)-generalized inverse of  $L_0$  are

$$\underline{G}(t, s) = \phi(s)\psi(t) - \psi(s)\phi(t) |_{t>s} - \frac{\phi(t)\phi(s)}{\phi'(L)},$$

$$G_Q(t, s) = \beta(t)\psi(s),$$

where  $\beta(t)$  is normalized by the requirement  $\int_0^L \beta(t)\psi(t) dt = 1$ .

For the branching problem, we have that if  $y(t)$  is a solution of (6.10) of period  $L$ , then  $y(t)$  satisfies

$$(6.11) \quad y(t) = \int_0^L G_X(t, s)[\mu f(s - \tau) + H(s, y(s))y(s)^2] ds + p\psi(t),$$

where

$$G_X(t, s) = \underline{G}(t, s) - \int_0^L \underline{G}(t, v)G_Q(v, s) dv.$$

For small  $\mu$  and  $p$ , (6.11) has a unique solution  $y(t, \mu, p)$  given by

$$(6.12) \quad y(t, \mu, p) = \mu h(t) + p\psi(t) + \text{higher order terms},$$

where we have written

$$h(t) = \int_0^L G_X(t, s)f(s - \tau) ds.$$

Now  $y(t, \mu, p)$  will be a solution of (6.10) provided that

$$\mu f(t - \tau) + H(t, y(t, \mu, p))y(t, \mu, p)^2$$

is in the range of the differential operator  $L_0$ . This means that

$$(6.13) \quad \int_0^L \psi(t)[\mu f(t-\tau) + H(t, y(t, \mu, p))y(t, \mu, p)^2] dt = 0.$$

If the expression (6.12) is used for  $y(t, \mu, p)$ , the determining equation (6.13) has the form

$$(6.14) \quad \mu \int_0^L \psi(t)f(t-\tau) dt - \int_0^L \psi(t)\frac{1}{2}g''(x_0(t))(\mu h(t) + p\psi(t))^2 dt + \text{higher order terms} = 0.$$

The character of the locus of (6.14) in  $\mu$ - $p$  space depends on whether or not the integral in the first term is or is not zero. Suppose first that this integral is not zero. Then near  $\mu = p = 0$  the locus consists of a single branch tangent to the line  $\mu = 0$ . Now when  $\mu = 0$ , (6.9) has a one-parameter family of periodic solutions, namely the translations of  $x_0(t)$ . This implies that the locus of (6.14) near the origin is in fact a portion of the axis  $\mu = 0$ , and that there are no other  $L$ -periodic solutions of (6.9) near to  $x_0(t)$  for small  $\mu$ .

Now suppose that the first integral in (6.14) is zero. This implies that  $f(t-\tau)$  is in the range of  $L_0$ , so that  $h(t)$  is an  $L$ -periodic solution of  $ly = f(t-\tau)$ . Indeed it is that periodic solution with initial value of the derivative equal to zero. It is found that in this case  $h(t)$  does not depend on the choice of  $\beta(t)$ . The second term of (6.14), when worked out gives

$$\frac{-1}{g(A)} \left[ \mu^2 \int_0^L f(t-\tau)h'(t) dt + \mu p \int_0^L f(t-\tau)\psi'(t) dt \right].$$

If we assume that  $\int_0^L f(t-\tau)\psi'(t) dt \neq 0$ , these quadratic terms show that the locus of (6.14) near the origin consists of two branches, one tangent to the line  $\mu = 0$  (which is, of course,  $\mu = 0$  itself), and a second branch tangent at the origin to the line

$$\mu \int_0^L f(t-\tau)h'(t) dt + p \int_0^L f(t-\tau)\psi'(t) dt = 0.$$

This shows that when  $\tau$  is such that  $\int_0^L f(t-\tau)\psi(t) dt = 0$  and  $\int_0^L f(t-\tau)\psi'(t) dt \neq 0$ , there is a second one-parameter family of  $L$ -periodic solutions of (6.9) for small nonzero  $\mu$ . These are given using  $y(t, \mu, p)$ , where  $p$  is related to  $\mu$  by the determining equation (6.13).

In summary, if  $F(\tau) = \int_0^L f(t-\tau)\psi(t) dt$ , then if  $F(\tau) \neq 0$ , the only  $L$ -periodic solutions of (6.8) near to  $x_0(t+\tau)$  are translations of  $x_0(t)$  with  $\mu = 0$ . If  $F(\tau) = 0$  and  $F'(\tau) \neq 0$ , then there is additionally a one-parameter family of  $L$ -periodic solutions of (6.8) near  $x_0(t+\tau)$  for small nonzero  $\mu$ .

**7. Concluding remarks.** The ideas in this paper have been strongly influenced by the paper of Hale [6], particularly his section V. It should be noted that the matrix  $K$  found by Hale on p. 243 of [6] is exactly the group inverse of the matrix  $D$ . Hale's application of this group inverse to the question of periodic solutions of a periodic vector system is similar to ours. We have been able to avoid the very complicated quantity  $\xi$  introduced by Hale. In this paper we deal with more general boundary-value problems and use a class of generalized inverses which includes the group inverse and the Moore-Penrose inverse as special cases.

The ultimate calculations made in the method used here are naturally the same as those which occur when construction is done by the implicit function theorem. The

determining equation (6.14) in the second example of § 6 was obtained in [7] using the implicit function theorem. A possible advantage of the present method is that the determining equation is obtained more directly, and some complicated calculations can be avoided.

The technique used here can also be used to study other cases of branching, including the Hopf bifurcation in autonomous systems where a nonconstant periodic solution branches from an equilibrium.

#### REFERENCES

- [1] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses, Theory and Applications*, Wiley, New York, 1974.
- [2] L. CESARI, *Functional Analysis and Periodic Solutions of Nonlinear Differential Equations*, Contributions to Differential Equations, 1 (1963), pp 149–187.
- [3] S. N. CHOW, J. K. HALE AND J. MALLET-PARET, *Applications of generic bifurcation I*. Arch. Rat. Mech. Anal., (1975), pp. 159–188.
- [4] M. J. ENGLEFIELD, *The commuting inverses of a square matrix*, Proc. Cambridge Philos. Soc., 62 (1966), pp. 667–671.
- [5] I. ERDÉLYI, *On the matrix equation  $Ax = \lambda Bx$* , J. Math. Anal. Appl., 17 (1967), pp. 119–132.
- [6] J. K. HALE, *On differential equations containing a small parameter*, Contributions to Differential Equations, 1 (1963), pp. 215–250.
- [7] W. S. LOUD, *Periodic solutions of  $x'' + cx' + g(x) = ef(t)$* , Memoirs of the American Mathematical Society No. 31, Providence, RI, 1959.
- [8] ———, *Generalized inverses and generalized Green's functions*, SIAM J. Appl. Math., 14 (1966), pp. 342–369.
- [9] ———, *Some examples of generalized Green's functions and generalized Green's matrices*, SIAM Rev., 12 (1970), pp. 194–210.
- [10] M. Z. NASHED, *Generalized Inverses and Applications*, Academic Press, New York, 1976.
- [11] W. T. REID, *Generalized Green's matrices for compatible systems of differential equations*, Amer. J. Math., 53 (1931), pp. 443–459.
- [12] ———, *Generalized Green's matrices for two-point boundary problems*, SIAM J. Appl. Math., 15 (1967), pp. 856–873.
- [13] ———, *Generalized inverses of differential and integral operators*, Theory and Application of Generalized Inverses of Matrices: Symposium Proceedings, Texas Technological College Mathematics Series, No. 4, Lubbock, TX, 1968, pp. 1–25.
- [14] ———, *Ordinary Differential Equations*, Wiley, New York, 1971.

## NONEXISTENCE OF GLOBAL SOLUTIONS FOR AN INTEGRODIFFERENTIAL SYSTEM IN REACTOR DYNAMICS\*

C. V. PAO†

**Abstract.** This paper is concerned with the instability behavior of an integrodifferential system arising in nuclear reactor dynamics. The spatial domain under consideration can be either bounded, subject to certain boundary conditions, or the whole space  $R^n$ . It is shown that if the physical parameter  $\beta(x)$  is non-negative and  $\beta(x) \neq 0$ , which corresponds to positive feedback reactivity in the reactor system, then for certain classes of initial functions the corresponding solution of the initial boundary-value problem (or the Cauchy problem) grows unbounded in finite time. This blowing-up property holds for a large class of nonlinear functions, including the physically most interesting one, and for very small initial perturbations from its equilibrium solution. An explicit instability region as well as an upper bound for the finite escape time are obtained.

**1. Introduction.** In this paper we consider the following nonlinear integro-differential system

$$p'(t) = p(t) \int_{\Omega} \beta(x)u(t, x) dx, \tag{1.1}$$

$(t > 0, x \in \Omega)$

$$u_t - Lu \equiv u_t - \left( \sum_{i,j=1}^n a_{ij}(x)u_{x_i x_j} + \sum_{i=1}^n a_i(x)u_{x_i} \right) = f(t, x, p(t) - p^*)$$

and the boundary and initial conditions

$$B[u] \equiv \alpha_1(x)\partial u/\partial \nu + \alpha_2(x)u = 0 \quad (t > 0, x \in \partial\Omega), \tag{1.2}$$

$$u(0, x) = u_0(x) \quad (x \in \Omega), \tag{1.3}$$

$$p(0) = p_0, \tag{1.4}$$

where  $L$  is a uniformly elliptic operator on the bounded domain  $\Omega$  in  $R^n$ ;  $\partial/\partial \nu$  is the outward normal derivative on the boundary  $\partial\Omega$ ,  $\alpha_i(x) \geq 0$  with  $\alpha_1(x) + \alpha_2(x) \neq 0$  on  $\partial\Omega$ ,  $p^* \geq 0$  is a constant and  $f$  is, in general, a nonlinear function of  $p$ . In addition to the initial boundary-value problem (1.1)–(1.4) for a bounded domain  $\Omega$ , we also consider the Cauchy problem (1.1), (1.3), (1.4), when  $\Omega$  is the whole space  $R^n$ . The system (1.1)–(1.4) occurs in nuclear reactor dynamics in which  $u$  and  $p$  represent the incremental temperature and instantaneous power from their corresponding steady-state  $0, p^*$ , respectively, while  $f$  is given by (cf. [1], [2], [7], [9]–[13], [18])

$$f(t, x, p(t) - p^*) \equiv \mu(x)(p(t) - p^*) \quad (\mu(x) \geq 0). \tag{1.5}$$

The term  $\int_{\Omega} \beta(x)u(t, x) dx$  measures the increment temperature feedback reactivity which plays an important role in the existence and nonexistence of a global solution.

The coupled system for a one-dimensional model has been investigated in a series of papers by Levin and Nohel [9]–[11] and by Miller [12], Bronikowski and Hall [1], [2], Suhadolc [18] and Infante and Walker [7]. The principle interests of these works are the global existence and the asymptotic behavior of the solution for the rod model, where  $\Omega$  is considered either as a finite interval or as the whole real line. Particular attention has been given to the function in (1.5). On the other hand, the existence of global solutions

\* Received by the editors January 11, 1979.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27607.

for a more general system in a multi-dimensional domain has been discussed by Pao [13]. In many of the above papers, it is assumed that the reactor system has “negative” feedback reactivity (i.e.,  $\beta(x) \leq 0$ ). An interesting problem is to predict the behavior of the solution when the reactor has positive feedback reactivity ( $\beta(x) \geq 0$ ). From a physical point of view, if the feedback is positive then the reactor gains energy which tends to destabilize the reactor system. The purpose of this paper is to show that if  $\beta(x) \geq 0$  but  $\beta(x) \not\equiv 0$ , then for a certain class of functions  $f$  the system is not only unstable but the solution also grows unbounded in finite time. We also give an upper bound for the “finite escape time” and an explicit instability region for the equilibrium solution  $(0, p^*)$ . In particular, these results hold for the function  $f$  in (1.5) and for very small initial perturbations from its equilibrium state. This is in sharp contrast to the asymptotic stability behavior obtained in [1], [2], [7], [9]–[12]. In terms of semi-group theory, our conclusion demonstrates that the associated operator cannot generate a nonlinear semi-group, at least not in the space of continuous functions.

The problem of nonexistence of global solutions for the traditional parabolic type system has been investigated by a number of authors and various methods have been developed (cf. [4]–[6], [15], [16]). In the present paper, we follow the approach of [16] using the notion of a lower solution. A novelty of this approach is that the analysis is elementary and the required conditions on  $f$  are simple. In § 2 we state the main theorems. Proofs of these theorems are given in § 3.

**2. The main results.** Throughout the paper we assume that the coefficients of  $L$  and the first partial derivatives of  $a_{ij}$  are Hölder continuous (of exponent  $\alpha \in (0, 1)$ ) in  $\Omega$ ; the matrix  $(a_{ij})$  is symmetric positive definite in  $\bar{\Omega}$ ,  $f(t, x, z)$  is Hölder continuous in every bounded subset of  $R^+ \times \bar{\Omega} \times R^+$ ;  $\beta(x)$  is bounded continuous in  $\Omega$ , and the boundary  $\partial\Omega$  is of class  $C^{2+\alpha}$ , where  $R^+ = [0, \infty)$  and  $\bar{\Omega}$  is the closure of  $\Omega$ . We also assume that  $\alpha_1, \alpha_2 \in H^{1+\alpha}(\partial\Omega)$ ,  $u_0 \in H^{2+\alpha}(\bar{\Omega})$  and  $u_0$  satisfies the boundary condition (1.2) at  $t = 0$ , where  $H^{1+\alpha}(\bar{\Omega})$  are the function spaces in the sense of [3], [8]. When  $\Omega = R^n$  we assume that  $u_0, f$  are bounded as  $|x| \rightarrow \infty$  and  $\beta$  is integrable in  $R^n$ . The above smoothness assumption will only be used to insure the existence of a solution for the corresponding linear problem (1.1)–(1.3), where  $f$  is replaced by a known function. In addition to the above requirements we assume, for simplicity, that  $f_z(t, x, z)$  exists and is bounded on bounded subsets of  $R^+ \times \bar{\Omega} \times R^+$ .

By solving the first equation in (1.1) for  $p$  and then substituting it into the second equation, we obtain

$$(2.1) \quad u_t - Lu = f\left(t, x, p_0 \exp\left(\int_0^t \int_{\Omega} \beta(x)u(s, x) \, dx \, ds\right) - p^*\right) \quad (t > 0, x \in \Omega),$$

where we have used (1.4). We shall study the problem (1.1)–(1.4) through the system (2.1), (1.2), (1.3) by using the notion of a lower solution. We call a smooth function  $v(t, x)$  in  $D \equiv (0, T] \times \Omega$  a lower solution if it satisfies the inequalities:

$$(2.2) \quad \begin{aligned} v_t - Lv &\leq f\left(t, x, p_0 \exp\left(\int_0^t \int_{\Omega} \beta(x)v(s, x) \, dx \, ds\right) - p^*\right) \quad ((t, x) \in D), \\ B[v] &\leq 0 \quad (t \in (0, T], x \in \partial\Omega), \\ v(0, x) &\leq u_0(x) \quad (x \in \Omega), \end{aligned}$$

where  $T$  is finite but arbitrary. In the case of  $\Omega = R^n$ , we replace the second condition in (2.2) by “ $v$  is bounded as  $|x| \rightarrow \infty$ ”. Here by a smooth function is meant a

continuous function on  $\bar{D}$  whose first derivative in  $t$  and second derivative in  $x_i$  are continuous in  $D$ , and  $\partial v/\partial \nu$  exists on  $\partial\Omega$ .

When  $\Omega$  is a bounded domain we need to consider the linear eigenvalue problem

$$(2.3) \quad \begin{aligned} L\phi + \lambda\phi &= 0 & (x \in \Omega), \\ B[\phi] &= 0 & (x \in \partial\Omega). \end{aligned}$$

It is well-known that the least eigenvalue  $\lambda_0$  of (2.3) is real nonnegative and its corresponding eigenfunction  $\phi$  is positive in  $\Omega$  (cf. [17]). In case  $\alpha_2(x) \neq 0$ , then  $\lambda_0$  is positive and if  $\alpha_1(x) > 0$ , then  $\phi(x)$  is positive on  $\bar{\Omega}$ . We normalize  $\phi$  so that  $\max\{\phi(x); x \in \bar{\Omega}\} = 1$ . With this notation we now state our main results in the following two theorems.

**THEOREM 1.** *Let  $\beta(x) \geq 0$  ( $\beta(x) \neq 0$ ),  $p_0 > p^*$  and  $u_0(x) \geq \delta\phi(x)$  for some  $\delta > 0$ . Assume that for  $t > 0$ ,  $x \in \Omega$ ,*

$$(2.4) \quad f(t, x, z_1 - p^*) \geq f(t, x, z_2 - p^*) \quad \text{when } z_1 \geq z_2 \geq 0$$

and for some constant  $b > 0$ ,

$$(2.5) \quad f(t, x, z - p^*) \geq b(z - p^*) \quad \text{when } z \geq p^*.$$

Then there exists a finite  $T_0$  such that a unique solution  $(u, p)$  to (1.1)–(1.4) exists on  $[0, T_0) \times \bar{\Omega}$  and  $[0, T_0)$ , respectively, such that

$$(2.6) \quad \lim_{t \rightarrow T_0} \left( \max_{x \in \bar{\Omega}} u(t, x) \right) = \infty \quad \text{and} \quad \lim_{t \rightarrow T_0} p(t) = \infty.$$

In particular, the above blowing-up property holds for the function given by (1.5).

**THEOREM 2.** *Let  $\Omega = R^n$ ,  $\beta(x) \geq 0$ ,  $p_0 > p^*$ ,  $u_0(x) \geq 0$  and  $\beta(x)u_0(x) \neq 0$ . Assume that  $f$  satisfies the conditions (2.4), (2.5). Then there exists finite  $T_0$  such that a unique solution  $(u, p)$  to (1.1), (1.3), (1.4) exists on  $[0, T_0) \times R^n$  and  $[0, T_0)$ , respectively, such that*

$$(2.7) \quad \lim_{t \rightarrow T_0} \left( \sup_{x \in R^n} u(t, x) \right) = \infty \quad \text{and} \quad \lim_{t \rightarrow T_0} p(t) = \infty.$$

In particular, (2.7) holds for the function  $f$  in (1.5).

**Remarks.** (a) The result in Theorem 1 implies that if  $f(t, x, 0) = 0$  (so that  $(0, p^*)$  is an equilibrium state) then an instability region of  $(0, p^*)$  is given by the set  $\{(u_0, p_0); u_0(x) \geq \delta\phi(x), p_0 > p^*\}$ , where  $\delta, (p_0 - p^*)$  can be arbitrarily small. Since  $\beta$  is required to be nonnegative and not identically zero, this instability behavior holds even when positive feedback occurs only in a small neighborhood of the reactor. (b) An upper bound for the finite escape time  $T_0$  in Theorem 1 is  $\gamma_0^{-1}$ , where  $\gamma_0$  is the largest positive number satisfying

$$(2.8) \quad \gamma_0(\gamma_0 + \lambda_0) \leq (b\bar{\beta}/2)(p_0 - p^*)$$

and  $\bar{\beta} = \int_{\Omega} \beta(x)\phi(x) dx$ . Similarly the finite escape time  $T_0$  in Theorem 2 is bounded by  $[(b\beta^*/2)(p_0 - p^*)]^{-1/2}$ , where  $\beta^*$  is given by (3.13).

**3. Proof of the Theorems.** In order to prove the theorems, we describe an iterative process which leads to the existence of a (local) solution of (1.1)–(1.4) as well as a lower bound of the solution. Suppose there exists a nonnegative lower solution  $v(t, x)$ . Then by using  $v$  as the initial iteration, we can construct a sequence  $\{u^{(k)}\}$  successively from



the linear system

$$\begin{aligned}
 (3.1) \quad & u_i^{(k)} - Lu^{(k)} = f\left(t, x, p_0 \exp\left(\int_0^t \int_\Omega \beta(x) u^{(k-1)}(s, x) dx ds\right) - p^*\right) \\
 & B[u^{(k)}] = 0 \\
 & u^{(k)}(0, x) = u_0(x)
 \end{aligned}
 \left. \begin{aligned}
 & (t \in (0, T], x \in \Omega) \\
 & (t \in (0, T], x \in \partial\Omega) \\
 & (x \in \Omega)
 \end{aligned} \right\} k = 1, 2, \dots$$

When  $\Omega = R^n$  this sequence is determined from the first and third equations in (3.1). The existence of such a sequence follows from the hypothesis in § 2. It can easily be shown from the property of a lower solution and the monotone property of  $f$  that the sequence  $\{u^{(k)}\}$  is monotone nondecreasing (cf. [14], [16]). Thus if the sequence is bounded from above then it converges to a function  $u$  such that  $u \geq v$  on  $[0, T] \times \Omega$ . In fact,  $u$  is the unique solution of (2.1), (1.2), (1.3). Our first step is to construct a lower solution from which the blowing-up property of the solution can be deduced.

*Proof of Theorem 1.* For the initial boundary-value problem (2.1), (1.2), (1.3) we seek a lower solution in the form  $v = q(t)\phi(x)$ , where  $\phi$  is the positive eigenfunction of (2.3) and  $q$  is a positive differentiable function with  $q(0) \leq \delta$ . Since  $B[\phi] = 0$  and  $q(0)\phi(x) \leq u_0(x)$ ,  $v$  fulfills the requirements of a lower solution if  $q$  satisfies the inequality

$$(3.2) \quad q' \phi - qL\phi \leq f\left(t, x, p_0 \exp\left(\bar{\beta} \int_0^t q(s) ds\right) - p^*\right) \quad (t \in [0, T], x \in \Omega),$$

where  $\bar{\beta} = \int_\Omega \beta(x)\phi(x) dx > 0$ . In view of (2.3), (2.5) and the fact that  $\phi(x) \leq 1$ , it suffices to find  $q$  such that

$$(3.3) \quad q' + \lambda_0 q \leq b\left(p_0 \exp\left(\bar{\beta} \int_0^t q(s) ds\right) - p^*\right) \quad (t \in (0, T]).$$

A convenient choice of  $q$  is given in the form

$$(3.4) \quad q(t) = a(1 - \gamma t)^{-1} \quad (t \in [0, \gamma^{-1})),$$

where  $a, \gamma$  are positive constants to be chosen. With this special form of  $q$ , (3.3) holds if for some  $\gamma > 0$ ,

$$(3.5) \quad a\gamma\eta^2 + \lambda_0 a\eta \leq b[p_0\eta^{(a\bar{\beta}/\gamma)} - p^*] \quad \text{for } \eta \in [1, \infty).$$

Let  $\gamma = a\bar{\beta}/2$ . Then (3.5) holds if

$$(3.6) \quad Q(\eta) \equiv (bp_0 - a\gamma)\eta^2 - \lambda_0 a\eta - bp^* \geq 0 \quad \text{for } \eta \in [1, \infty).$$

This is obviously the case when  $Q(1) \equiv bp_0 - a(\gamma + \lambda_0) - bp^* \geq 0$  and  $Q'(\eta) = 2(bp_0 - a\gamma)\eta - \lambda_0 a \geq 0$  for  $\eta \geq 1$ . Both requirements are satisfied if

$$(3.7) \quad b(p_0 - p^*) \geq a(\gamma + \lambda_0) = (2\gamma/\bar{\beta})(\gamma + \lambda_0).$$

Let  $\gamma_0$  be the largest positive constant satisfying  $\gamma_0 \leq \delta\bar{\beta}/2$  and (3.7) (or equivalently (2.8)). Then (3.5) holds with  $\gamma = \gamma_0$ ,  $a = 2\gamma_0/\bar{\beta}$  and  $q(0) = a \leq \delta$ . With this choice of  $a, \gamma_0$ ,  $v(t, x) = q(t)\phi(x)$  is a lower solution of (1.1)–(1.4) on  $[0, T] \times \Omega$  for every  $T < \gamma_0^{-1}$ .

We next define, for any given constant  $M > 0$ , a function  $f_M(t, x, z - p^*)$  such that  $f_M$  coincides with  $f$  on  $[0, T] \times \Omega$  when  $|z| \leq M$  and is uniformly bounded, monotone nondecreasing for all  $|z| < \infty$ , where  $T < \gamma^{-1}$ . For example, we may define  $f_M$  by

$f_M(t, x, z - p^*) = f(t, x, z - p^*)$  when  $|z| \leq M$  and  $f_M(t, x, z - p^*) = f(t, x, \pm M - p^*)$  when  $z > M$  and  $z < -M$ , respectively. Consider the modified equation

$$(3.8) \quad u_t - Lu = f_M\left(t, x, p_0 \exp\left(\int_0^t \int_{\Omega} \beta(x)u(s, x) dx ds\right) - p^*\right) \quad (t \in [0, T], x \in \Omega)$$

together with the boundary and initial conditions (1.2), (1.3), where  $T < \gamma^{-1}$  is fixed and  $M \geq a(1 - \gamma T)^{-1}$ . Since  $v(t, x) \leq M$  on  $[0, T] \times \bar{\Omega}$ , we see that  $v$  is also a lower solution for (3.8), (1.2), (1.3). Using  $v$  as the initial iteration, we construct a sequence  $\{u^{(k)}\}$  from (3.1) but with  $f$  replaced by  $f_M$ . This sequence is again monotone nondecreasing since  $f_M$  preserves the monotone property of  $f$ . In view of the uniform boundedness of  $f_M$  we conclude from the well-known estimate for linear parabolic system that the sequence  $\{u^{(k)}\}$  is also bounded (cf. [3], p. 146). This implies that  $\{u^{(k)}\}$  converges monotonically to a function  $u$  satisfying  $u(t, x) \geq v(t, x)$  on  $[0, T] \times \bar{\Omega}$ . A regularity argument shows that  $u$  is the unique solution of the modified problem (3.8), (1.2), (1.3) (cf. [14]). By the definition of  $f_M$ ,  $u$  is also a solution of the original system (2.1), (1.2), (1.3) for as long as  $|u| \leq M$ .

To show the blowing-up property (2.6) for  $u$ , we assume, by contradiction, that the solutions of (2.1), (1.2), (1.3) were bounded on  $[0, \gamma^{-1}] \times \bar{\Omega}$  (say, by  $K$ ). Choose  $T_1 < \gamma^{-1}$  but sufficiently close to  $\gamma^{-1}$  such that  $v(T_1, x_1) \geq K + 1$  for some  $x_1 \in \Omega$ . Define the modified function  $f_M$  with  $M \geq K + 1$ ,  $T = T_1$ . Then from the above discussion, the modified problem (3.8), (1.2), (1.3) has a unique solution  $u$  such that  $u(t, x) \geq v(t, x)$  on  $[0, T_1] \times \bar{\Omega}$ . Since  $v(T_1, x_1) \geq K + 1$ , there exists  $T_0 \leq T_1$  such that  $u(t, x) \leq K + 1$  on  $\bar{D}_0 \equiv [0, T_0] \times \bar{\Omega}$  and  $u(T_0, x_0) = K + 1$  for some  $x_0 \in \bar{\Omega}$ . This shows that  $u$  is the solution of the original problem on  $D_0$  and  $u(T_0, x_0) = K + 1$ . This contradiction leads to the conclusion in (2.6) for  $u$ . To show the result for  $p(t)$ , we observed that for every  $t < T_0$ ,  $u(t, x)$  is finite in  $\bar{\Omega}$ . In view of (1.1),  $p(t)$  is also finite for  $t < T_0$ . Now if  $p(t)$  were bounded at  $t = T_0$ , then  $f(t, x, p(t) - p^*)$  is bounded on  $\bar{D}_0$ . This implies that  $u$  is also bounded on  $\bar{D}_0$  which is absurd. This completes the proof of the theorem.

*Proof of Theorem 2.* For the Cauchy problem (1.1), (1.3), (1.4) we construct a lower solution in the form  $v(t, x) = \delta q(t)w(t, x)$ , where  $q(t)$  is again a positive differential function,  $\delta > 0$  is a constant and  $w$  is the solution of the linear problem

$$(3.9) \quad \begin{aligned} w_t - Lw &= 0 & (t > 0, x \in R^n) \\ w(0, x) &= u_0(x) & (x \in R^n). \end{aligned}$$

In fact,  $w$  is given by

$$(3.10) \quad w(t, x) = \int_{R^n} \Gamma(t, x; 0, y)u_0(y) dy,$$

where  $\Gamma$  is the fundamental solution of  $u_t - Lu = 0$ . Since  $u_0 \geq 0$ , ( $u_0 \not\equiv 0$ ) and is bounded in  $R^n$ , the function  $w$  is positive, bounded in  $(0, \infty) \times R^n$ . We choose  $\delta$  such that  $\delta w \leq 1$  on  $[0, \infty) \times R^n$ . By requiring  $q(0) \leq \delta^{-1}$ , the function  $v = \delta qw$  becomes a lower solution of (1.1), (1.3), (1.4) if

$$(3.11) \quad \delta(q'w + qw_t - qLw) \leq b\left[p_0 \exp\left(\delta \int_0^t \int_{R^n} \beta(x)q(s)w(s, x) dx ds\right) - p^*\right],$$

where we have used the hypothesis (2.5). Since  $w$  satisfies (3.9) and  $\delta w \leq 1$ , the above inequality holds if

$$(3.12) \quad q' \leq b\left[p_0 \exp\left(\beta^* \int_0^t q(s) ds\right) - p^*\right] \quad (t \in (0, T]),$$

where

$$(3.13) \quad \beta^* \equiv \delta \inf \left\{ \int_{R^n} \beta(x) w(s, x) dx, 0 \leq s \leq T \right\}.$$

Notice from the positivity of  $w$  and the hypothesis  $\beta(x)u_0(x) \neq 0$  that  $\beta^* > 0$ . By choosing  $q$  as in the proof of Theorem 1 except with  $\lambda_0 = 0$  and  $\beta$  replaced by  $\beta^*$ , we see that  $v$  is a lower solution. The remaining proof follows from the same argument as in the proof of Theorem 1.

#### REFERENCES

- [1] T. A. BRONIKOWSKI, J. E. HALL AND J. A. NOHEL, *Quantitative estimates for a nonlinear system of integrodifferential equations arising in reactor dynamics*, this Journal, 3 (1972), pp. 567–588.
- [2] T. A. BRONIKOWSKI, *An integrodifferential system which occurs in reactor dynamics*, Arch. Rational Mech. Anal., 37 (1970), pp. 363–380.
- [3] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [4] ———, *Remarks on nonlinear parabolic equations*, Symposium on Applied Mathematics, Proc. Amer. Math. Soc., 17 (1965), pp. 3–23.
- [5] H. FUJITA, *On the blowing up of solutions of the Cauchy problem for  $u_t = \Delta u + u^{1+\alpha}$* , J. Fac. Sci. Univ. Tokyo Sect. IA, 13 (1966), pp. 109–124.
- [6] K. HAYAKAWA, *On non-existence of global solutions of some semi-linear parabolic differential equations*, Proc. Japan Acad., 49 (1973), pp. 503–505.
- [7] E. F. INFANTE AND J. A. WALKER, *On the stability properties of an equation arising in reactor dynamics*, J. Math. Anal. Appl., 55 (1976), pp. 112–124.
- [8] O. A. LADYZENSKAYA, V. A. SOLONNIKOV AND N. N. URALCERA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematics Society, Providence, RI, 1968.
- [9] J. J. LEVIN AND J. A. NOHEL, *On a system of integrodifferential equations occurring in reactor dynamics*, J. Math. Mech., 9 (1960), pp. 347–368.
- [10] ———, *A system of nonlinear integrodifferential equations*, Michigan Math. J., 13 (1966), 257–270.
- [11] ———, *A nonlinear system of integrodifferential equations*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 398–405.
- [12] R. K. MILLER, *On the linearization of Volterra integral equations*, J. Math. Anal. Appl., 23 (1968), pp. 198–208.
- [13] C. V. PAO, *Solution of a nonlinear integrodifferential system arising in nuclear reactor dynamics*, Ibid., 48 (1974), pp. 470–492.
- [14] ———, *Positive solution of a nonlinear boundary-value problem of parabolic type*, J. Differential Equations, 22 (1976), pp. 145–163.
- [15] ———, *Non-existence of global solutions and bifurcation analysis for a boundary-value problem of parabolic type*, Proc. Amer. Math. Soc., 65 (1977), pp. 245–251.
- [16] ———, *Asymptotic behavior and non-existence of global solutions for a class of nonlinear boundary-value problems of parabolic type*, J. Math. Anal. Appl., 65 (1978), pp. 616–637.
- [17] M. H. PROTTER AND H. F. WEINBERGER, *On the spectrum of general second order operators*, Bull. Amer. Math. Soc., 72 (1966), pp. 251–255.
- [18] A. SUHADOLC, *On a system of integro-differential equations*, SIAM J. Appl. Math., 21 (1971), pp. 195–206.

## A MIXED BOUNDARY VALUE PROBLEM VIEWED AS A TYPE V PROBLEM\*

H. L. JOHNSON†

**Abstract.** This paper formulates a mixed boundary value problem for Laplace's equation in an axial symmetric domain in  $E^3$  as a type V boundary value problem. This formulation enables the boundary value problem to be transformed into a singular integral equation. The paper also considers the necessity of imposing an orthogonality condition on the boundary data to insure the existence of a solution. The principal theorem of the paper is that when the domain is a sphere no orthogonality condition is necessary.

**1. Introduction.** Mixed boundary value problems for elliptic partial differential equations occur frequently in mathematical physics. Examples of such problems are contained in the book by Sneddon [8] and in the papers of W. D. Collins [1], [2] and [3]. Compared with our knowledge of the Dirichlet and the Neumann boundary value problems, our knowledge of mixed boundary value problems is meager. Martin Schechter [7] has proven a Fredholm alternative theorem for quite general elliptic mixed boundary value problems. This theorem implies that a solution will exist if a finite number of orthogonality conditions is placed on the boundary data. For Laplace's equation,  $\nabla^2 u = 0$ , it is well known that the Dirichlet problem requires no orthogonality condition on the boundary data and that the Neumann problem requires one orthogonality condition. A basic question in potential theory that, to the author's knowledge, has not been answered is the necessity of orthogonality conditions on the boundary data for a mixed boundary value problem for Laplace's equation. In a previous paper [5], the author showed that a Dirichlet-Neumann type mixed boundary value problem for Laplace's equation on a sphere could be transformed into a Fredholm integral equation of the second kind with a weakly singular kernel. The mathematics used in [5] required an orthogonality condition to be placed on the boundary data.

This paper extends in two ways the analysis of [5]. First, the mixed boundary value problem is posed on a quite general axial symmetric, bounded domain in  $E^3$ , and it is shown that this problem can be transformed into type V boundary value problem. The solvability of type V boundary value problems and their representation as singular integral equations is the topic of Chapter 9 in the classical text [6] by I. N. Muskhelishvili. Secondly, we are able to use one form of the solvability conditions stated in [6] to prove the following theorem.

**THEOREM.** *No orthogonality condition need be placed on the boundary data to insure the existence of a solution of the mixed boundary value problem  $\nabla^2 w = 0$  in a solid sphere  $0 \leq \rho < R$  with the axial symmetric boundary conditions  $w(R, \phi) = H_1(\phi)$ ,  $0 \leq \rho < \alpha$ ,  $(\partial w / \partial \rho)(R, \phi) = H_2(\phi)$ ,  $\alpha < \phi < \pi$ .*

**2. The mixed boundary value problem.** Let  $D: 0 \leq \rho < R(\phi)$ ,  $0 \leq \phi \leq \pi$ ,  $0 \leq \theta \leq 2\pi$ , where  $\rho$ ,  $\phi$ , and  $\theta$  denote spherical variables. Let  $P.C.^{(m)}(S)$  denote the class of functions with piecewise continuous  $m$ th partial derivatives on a set  $S$ . Let  $w = w(\rho, \phi) \in C^2(D) \cap P.C.^{(1)}(\bar{D})$ . We consider the boundary value problem

$$(1) \quad \nabla^2 w = \frac{1}{\rho^2} \left[ \left( \frac{\partial \rho^2}{\partial \rho} \frac{\partial w}{\partial \rho} \right) + \frac{1}{\sin(\phi)} \frac{\partial}{\partial \phi} \left( \sin(\phi) \frac{\partial w}{\partial \phi} \right) \right] = 0, \quad 0 \leq \rho < R(\phi), \quad 0 \leq \phi \leq \pi,$$

$$(2) \quad w(R(\phi), \phi) = H_1(\phi), \quad \phi \in L_D^+ = \{\phi: 0 \leq \phi < \alpha\},$$

\* Received by the editors January 29, 1979, and in revised form July 12, 1979.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

$$(3) \quad \frac{\partial w}{\partial n}(R(\phi), \phi) = H_2(\phi), \quad \phi \in L_N^+ = \{\phi: \alpha < \phi \leq \pi\},$$

where  $H_1 \in C^2(\overline{L_D^+})$ ,  $H_2 \in C^1(\overline{L_N^+})$ .

Let  $R = R(\phi) \in P.C.^{(3)}(L_D^+ \cup L_N^+)$ , and  $R > 0$  on  $0 \leq \phi \leq \pi$ . We continue  $R$  into  $-\pi \leq \phi < 0$  as an even function and without loss of generality, assume that

$$R(\phi) = \hat{R}(\cos(\phi)), \quad -\pi \leq \phi < \pi.$$

In addition, we continue the boundary conditions (2) and (3) into  $-\pi \leq \phi < 0$  by

$$w(R(\phi), \phi) = H_1(-\phi), \quad -\alpha < \phi < 0,$$

$$\frac{\partial w}{\partial n}(R(\phi), \phi) = H_2(-\phi), \quad -\pi \leq \phi < -\alpha.$$

Let  $S^+ = \{(\rho, \phi): 0 \leq \rho < \hat{R}(\cos(\phi)), -\pi \leq \phi \leq \pi\}$ ,  $L = \{(R(\phi), \phi), -\pi \leq \phi \leq \pi\}$ ,  $L_D = \{(R(\phi), \phi) - \alpha < \phi < \alpha\}$ , and  $L_N = \{(R(\phi), \phi), -\pi \leq \phi < -\alpha, \alpha < \phi \leq \pi\}$ . Let  $f = f(\delta)$  be an analytic function of  $\delta$  in  $S^+$  subject to the symmetry condition

$$(4) \quad f(\bar{\delta}) = \overline{f(\delta)}.$$

Let  $\delta = \rho \cos(\phi) + i\rho \sin(\phi) \cos(\theta)$ . It is well-known [4] that

$$(5) \quad w(\rho, \phi) = \frac{Re}{\pi} \int_0^\pi f(\delta) d\theta = \frac{1}{\pi} \int_0^\pi f(\delta) d\theta$$

is a solution of (1). For fixed  $(\rho, \phi)$ , it follows that

$$d\theta = \frac{d\delta}{-i\rho \sin(\phi) \sin(\theta)} = i \frac{|\sin(\phi)|}{\sin(\phi)} \frac{d\delta}{\sqrt{(\delta-z)(\delta-\bar{z})}},$$

where  $z = \rho e^{i\phi}$ . As a function of  $\delta$ ,  $\sqrt{(\delta-z)(\delta-\bar{z})}$  has branch cuts along the vertical segments:  $\delta = \rho \cos(\phi) + iy$ ,  $|y| > \rho \sin(\phi)$ , and  $\sqrt{(-z)(-\bar{z})} = |z| \geq 0$ . Equation (5) becomes

$$(6) \quad w(\rho, \phi) = \frac{i|\sin(\phi)|}{\sin(\phi)\pi} \int_z^{\bar{z}} \frac{f(\delta) d\delta}{\sqrt{(\delta-z)(\delta-\bar{z})}}.$$

Letting  $\rho \rightarrow R(\phi)$ , deforming the path of integration into the portion of  $L_D$  between  $z = R(\phi)e^{i\phi}$  and  $\bar{z} = R(\phi)e^{-i\phi}$ , and using property (4), one obtains

$$(7) \quad \frac{\sin(\phi)}{|\sin(\phi)|} w(R(\phi), \phi) = Re \left( \frac{2}{\pi i} \int_0^\phi \frac{f(\delta)\delta' d\eta}{\sqrt{(\delta-z)(\delta-\bar{z})}} \right), \quad 0 < |\phi| < \alpha,$$

where  $\delta = R(\eta)e^{i\eta} = d\delta/d\eta$ .

The normal derivative of  $w$  on the surface  $\partial D: \rho - R(\phi) = 0$  is

$$\frac{\partial w}{\partial n} = \frac{R(\phi)(\partial w/\partial \rho) - R'(\phi)/R(\phi)(\partial w/\partial \phi)}{\sqrt{R^2(\phi) + (R'(\phi))^2}}.$$

The derivatives  $\partial w/\partial \rho$  and  $\partial w/\partial \phi$  for  $\rho < R(\phi)$  can be obtained by differentiating (5).

This leads to

(8)

$$R(\phi) \frac{\sin(\phi)}{|\sin(\phi)|} \Big|_{z'} \frac{\partial w}{\partial n} = \frac{1}{\pi i} \int_z^z \frac{f'(\delta)(R(\phi)\hat{R}'(\cos(\phi)) - \delta(R(\phi) + \cos(\phi)\hat{R}'(\cos(\phi)))) d\delta}{\sqrt{(\delta - z)(\delta - \bar{z})}}, \quad \alpha < |\phi| < \pi$$

where  $z = R(\phi)e^{i\phi}$  and  $z' = dz/d\phi$ .

A type V boundary value problem is in the problem of finding an analytic function  $\Phi(z)$ ,  $z \in S^+$  which satisfies the boundary condition

$$(9) \quad R_e(L(\Phi))(t_0) = R_e\left(\sum_{j=0}^m a_j(t_0)\Phi^{(j)}(t_0) + \int_{\partial S^+} K_j(t_0, t)\Phi^{(j)}(t) ds\right) = F(t_0), \quad t_0 \in \partial S^+,$$

where  $s$  denotes arclength along  $\partial S^+$  and  $F(t_0)$  is a prescribed function. It is necessary that  $a_m(t_0) \neq 0$ ,  $t_0 \in \partial S^+$ .

Our next goal is to transform the boundary conditions as given by (7) and (8) into the form of (9). In doing this, we shall follow the notation, when possible, of Chapter 9 of [6] and evaluate all of the important parameters and coefficient functions pertinent to our problem. The reader who is unfamiliar with the properties of type V problems should consult [6] before proceeding further with this paper.

For our boundary value problem  $m = 1$ . In this case, it is known that  $\Phi(z)$  should have the form

$$(10) \quad \Phi(z) = \int_L \ln(1 - z/t)\mu(t) ds + \int_L \mu(t) ds + iC,$$

where  $t$  is a complex variable on  $L$ ,  $\mu$  is a real-valued function,  $C$  is a real constant, and  $s$  denotes arclength along  $L$ . The form of (10) and the natural symmetry condition  $f(\bar{z}) = f(z)$  imposed on  $f$  lead one to relate  $\Phi$  and  $f$  by

$$(11) \quad \Phi(z) = if(z),$$

and to ask that  $\mu(\bar{t}) = -\mu(t)$ . In addition, it is convenient to use the angle  $\eta$  instead of the arclength  $s$  to parameterize  $L$ . With these changes, (10) becomes

$$(12) \quad \Phi(z) = \int_{-\pi}^{\pi} \ln(1 - z/\delta)\mu(\eta) d\eta + iC, \quad \delta = R(\eta)e^{i\eta}.$$

We need to modify (7) and (8) so they take the form (9) with  $a_1(t_0) \neq 0$ . To do this, we first note the following lemma.

LEMMA 1. For  $\delta = R(\eta)e^{i\eta}$ ,  $z = R(\phi)e^{i\phi}$ ,

$$(\delta - z)(\delta - \bar{z}) = \frac{(\cos(\eta) - \cos(\phi))}{2} G(\cos(\phi), \eta),$$

where

$$G(x, \eta) = 4R^2(\eta)e^{i\eta} + 4R(\eta)(xe^{i\eta} - 1) \int_0^1 R'(y) d\sigma + 2(\cos(\eta) - x) \left( \int_0^1 R'(y) d\sigma \right)^2 \neq 0,$$

$$y = (1 - \sigma)x + \sigma \cos(\eta).$$

Letting  $K(\phi, \eta) = \sqrt{|\sin^2(\phi/2) - \sin^2(\eta/2)|}$  and using the identity

$$\frac{d}{d\phi} \left( \int_0^\phi \frac{\sin(t)}{2K(\phi, t)} \frac{1}{\pi} \int_0^t \frac{U(s) ds}{K(t, s)} \right) = U(\phi),$$

(7) can be transformed into

$$(13) \quad R_e \left( a_1(\phi)\Phi'(z) + a_0(\phi)\Phi(z) + \int_0^\phi \Phi(\delta) \frac{\partial K_D}{\partial \phi}(\phi, \eta) d\eta \right) = F_D(\phi),$$

$-\alpha < \phi < \alpha,$

where

$$(14) \quad a_1(\phi) = (-iz')^{3/2}, \quad z' = \frac{dz}{d\phi}, \quad -\alpha < \phi < \alpha,$$

$$(15) \quad a_0(\phi) = K_D(\phi, \phi) - \frac{z''}{2\sqrt{-iz'}},$$

$$(16) \quad K_D(\phi, \eta) = \frac{-\sin(\phi)}{\pi} \delta' R(\phi) \int_0^1 \frac{x^2 G'(y, \eta) dx}{\sqrt{1-x^2} (G(y, \eta))^{3/2}},$$

$$y = \cos(\eta)(1-x^2) + \cos(\phi)x^2, \quad G'(y, \eta) = \frac{\partial G}{\partial y}(y, \eta),$$

$$(17) \quad F_D(\phi) = \frac{d}{d\phi} \sqrt{R(\phi)} \frac{d}{d\phi} \int_0^\phi \frac{|\sin(t)|}{2} \frac{H_1(t) dt}{K(\phi, t)}, \quad |\phi| < \alpha.$$

Setting  $\phi = \eta$  in (8), multiplying both sides of (8) by  $\sin(\eta)/2K(\eta, \phi)$  for  $-\pi < \phi < -\alpha$ , using the above lemma and the identity

$$\frac{d}{d\phi} \left( \frac{1}{\pi} \int_{-\pi}^\phi \frac{\sin(\eta)}{2K(\eta, \phi)} \left( \int_{-\pi}^\eta \frac{U(s) ds}{K(s, \eta)} \right) d\eta \right) = U(\phi),$$

and integrating by parts yields

$$(18) \quad R_e \left( a_1(\phi)\Phi'(z) + a_0(\phi)\Phi(z) + \int_{-\pi}^\phi \frac{\Phi(\delta)}{\sqrt{R(\phi)}} \frac{\partial^2 K_N}{\partial \phi \partial \eta}(\phi, \eta) d\eta \right) = F_N(\phi),$$

$-\pi < \phi < -\alpha,$

where

$$(19) \quad a_1(\phi) = i(iz')^{3/2}, \quad a_0(\phi) = \frac{1}{\sqrt{R(\phi)}} \frac{\partial K_N}{\partial \phi}(\phi, \eta)|_{\eta=\phi},$$

$$(20) \quad K_N(\phi, \eta) = \frac{\delta^4}{\pi} \int_0^1 \frac{(\hat{R}(y)\hat{R}'(y) - \delta(\hat{R}(y) + y\hat{R}'(y)))}{\sqrt{1-x^2}\sqrt{-G(y, \eta)}} dx,$$

$$y = \cos(\phi) - 2K^2(\eta, \phi)x^2,$$

$$(21) \quad F_N(\phi) = \frac{1}{\sqrt{R(\phi)}} \frac{d}{d\phi} \int_{-\pi}^{\phi} \frac{\sin(\eta)R(\eta)|\delta'|}{2K(\eta, \phi)} H_2(\eta) d\eta \quad -\pi < \phi - \alpha.$$

Finally,

$$(22) \quad R_e \left( a_1(\phi)\Phi'(z) + a_0(\phi)\Phi(z) - \int_{\phi}^{\pi} \frac{\Phi(\delta)}{\sqrt{R(\phi)}} \frac{\partial^2 K_N}{\partial\phi\partial\eta}(\phi, \eta) d\eta \right) = F_N(\phi),$$

$$\alpha < \phi < \pi,$$

where  $a_1(\phi)$  and  $a_0(\phi)$  are given by (19) and

$$(23) \quad F_N(\phi) = \frac{1}{\sqrt{R(\phi)}} \frac{d}{d\phi} \int_{\phi}^{\pi} \frac{\sin(\eta)}{2} \frac{R(\eta)|\delta'|}{K(\eta, \phi)} H_2(\eta) d\eta, \quad \alpha < \phi < \pi.$$

It can be shown from (17), (21) and (23) that if  $H_1 \in C^2(\bar{L}_D)$ , then  $F_D \in C^1(\bar{L}_D)$ , and if  $H_2 \in C^1(\bar{L}_N)$ , then  $F_N \in C^1(\bar{L}_N)$ .

The function  $a_1$  as given by (14) and (19) is nonzero on  $-\pi \leq \phi \leq \pi$ , and it is continuous on  $-\pi < \phi < \pi$ . Moreover, the change in the argument of  $a_1$  around the contour  $L$  is  $\Delta_L(\arg(a_1)) = 3\pi$ . Thus, the parameter

$$(24) \quad n = \frac{1}{2\pi} \Delta_L(\arg(\bar{a}_1)) = -\frac{3}{2},$$

and the index

$$(25) \quad \kappa = 2(m + n) = -1.$$

Muskhelishvili requires that  $n$  be integer-valued, but this assumption is unnecessary and too restrictive for our problem.

Setting

$$(26) \quad h_0(\phi, \eta) = \begin{cases} \frac{1}{\sqrt{R(\phi)}} \frac{\partial^2 K_N}{\partial\phi\partial\eta}(\phi, \eta), & -\pi < \eta < \phi, -\pi < \phi < -\alpha, \\ 0, & \phi < \eta < \pi, -\pi < \phi < -\alpha, \\ \frac{\partial K_D}{\partial\phi}(\phi, \eta), & 0 < \eta < \phi < \alpha \\ 0, & -\alpha < \phi < \eta < 0, |\phi| < \alpha, \\ 0, & |\eta| > |\phi|, |\phi| < \alpha, \\ -\frac{1}{\sqrt{R(\phi)}} \frac{\partial^2 K_N}{\partial\phi\partial\eta}(\phi, \eta), & \phi < \eta < \pi, \alpha < \phi < \pi, \\ 0, & -\pi < \eta < \phi, \alpha < \phi < \pi, \end{cases}$$

$$(27) \quad F(\phi) = \begin{cases} F_N(\phi), & -\pi < \phi < \alpha, \alpha < \phi < \pi, \\ F_D(\phi), & -\alpha < \phi < \alpha. \end{cases}$$

Equations (13), (18), and (23) can, under the above notation, be written as

$$(28) \quad R_e \left( a_1(\phi)\Phi'(z) + a_0(\phi)\Phi(z) + \int_{-\pi}^{\pi} h_0(\phi, \eta)\Phi(\delta) d\eta \right) = F(\phi), \quad |\phi| < \pi;$$

again  $z = R(\phi)e^{i\phi}$ ,  $\delta = R(\eta)e^{i\eta}$ .



The singular integral equation form of (28) is given in [6]. Correlating the notation of [6] and our own, we set  $t_0 = z = R(\phi)e^{i\phi}$ ,  $t'_0 = dt_0/ds = z'/|z'|$ ,  $z' = dz/d\phi$ ;

$$(29) \quad A(t_0) = A(z) = R_e\left(-\pi i \frac{dt_0}{ds} a_1(t_0)\right) = R_e\left(-\pi \frac{iz'}{|z'|} a_1(\phi)\right),$$

$$(30) \quad N_0(z, \delta) = -\log\left(1 - \frac{z}{\delta}\right) + 1,$$

$$(31) \quad N_1(z, \delta) = \frac{-1}{\delta - z},$$

$$(32) \quad N(z, \delta) = R_e\left(a_0(\phi)N_0(z, \delta) + \int_{-\pi}^{\pi} h_0(z, t)N_0(t, \delta) d\psi + a_1(\phi)N_1(z, \phi)\right),$$

$t = R(\psi)e^{i\psi}$ ,

$$(33) \quad \sigma(\phi) = R_e\left(ia_0(\phi) + i \int_{-\pi}^{\pi} h_0(\phi, \eta) d\eta\right),$$

$$(34) \quad A(z)\mu(\phi) + \int_{-\pi}^{\pi} N(z, \delta)|\delta'| \mu(\eta) d\eta = F(\phi) - C\sigma(\phi),$$

$$\delta = R(\eta)e^{i\eta}, \quad |\phi| < \pi.$$

The standard Green's identity

$$\int\int_{\partial D} w(\partial w/\partial n) dS = \int\int\int_D |\nabla w|^2 dV,$$

applied to a harmonic function  $w$  in P.C.<sup>(1)</sup> ( $\bar{D}$ ) implies that the number of linearly independent solutions of  $R_e(L(\Phi)) = 0$  as given by (28) is  $k = 0$ . Therefore,  $k' = k - \kappa = 1$ . Hence, there is one linearly independent real eigenfunction  $v = v(\phi)$  of an adjoint homogeneous, singular integral equation  $N'(v) = 0$  of (34). Our boundary value problem is solvable for arbitrary functions  $H_1 \in C^2(\bar{L}_D)$  and  $H_2 \in C^1(\bar{L}_N)$  if and only if

$$(35) \quad (\sigma, v) = \int_{-\pi}^{\pi} \sigma(\phi)v(\phi) d\phi \neq 0.$$

We proceed to investigate condition (35) when  $D$  is a solid sphere.

**3.  $(\sigma, v)$  when  $R(\phi) = \text{constant}$ .** When  $R = R(\phi) = \text{constant}$ ,

$$G(\cos(\phi), \eta) = 4R^2 e^{i\eta}, \quad h_0(\phi, \eta) \equiv 0,$$

$$a_1(\phi) = R^{3/2} e^{(i3/2)\phi}, \quad a_0(\phi) = \frac{\sqrt{2}}{2} e^{i\phi/2} u(\alpha - |\phi|),$$

where  $u(x) = 1, x > 0, 0, x < 0$ ,

$$(36) \quad \sigma(\phi) = R_e(ia_0(\phi)) = -\frac{\sqrt{R}}{2} \sin(\phi/2)u(\alpha - |\phi|),$$

$$w_0(\phi) = R_e(a_0(\phi)) = \frac{\sqrt{R}}{2} \cos(\phi/2)u(\alpha - |\phi|),$$

and

$$w_j(\phi) = L(z^j) = a_1(\phi) \frac{dz^j}{dz} + a_0(\phi)z^j = \sqrt{R}\left(j + \frac{1}{2}u(\alpha - |\phi|)\right) e^{(j+(1/2))i\phi}.$$

It is known, again see [6], that the eigenfunction  $v$  is characterized by the orthogonality conditions

$$(37) \quad (w_j, v) = 0, \quad j = 0, 1, 2, \dots$$

In the special case of  $\alpha = \pi$ , the Dirichlet problem,

$$(38) \quad v(\phi) = \sin(\phi/2), \quad |\phi| < \pi$$

and hence  $(\sigma, v) = -\frac{1}{2}\sqrt{R} \int_{-\pi}^{\pi} (\sin(\phi/2))^2 d\phi = -(\sqrt{R}/2)\pi$ . For a general value of  $\alpha$ ,  $0 < \alpha < \pi$ , we seek  $v$  in the form

$$(39) \quad v(\phi) = \sin(\phi/2) + \sum_{k=1}^{\infty} v_k \sin\left(\left(k + \frac{1}{2}\right)\phi\right).$$

The orthogonality conditions (37) imply that

$$(40) \quad v_k = \frac{1}{D_k} \left( C_k + \sum_{\substack{n=1 \\ n \neq k}}^{\infty} a_{k,n} v_n \right), \quad k = 1, 2, \dots,$$

where

$$(41) \quad D_k = \pi + \frac{\alpha}{2k} \left( 1 - \frac{\sin((2k+1)\alpha)}{(2k+1)\alpha} \right) \cong \pi,$$

$$(42) \quad C_k = \frac{-1}{2k} \left( \frac{\sin(k\alpha)}{k} - \frac{\sin((k+1)\alpha)}{k+1} \right),$$

$$(43) \quad a_{k,n} = \frac{-1}{2k} \left( \frac{\sin((n-k)\alpha)}{n-k} - \frac{\sin((n+k+1)\alpha)}{n+k+1} \right), \quad n \neq k.$$

Let  $u_k = k^2 v_k$ ,  $g_k = (1/D_k) k^2 C_k$ ,  $A_{k,n} = k^2 a_{k,n} / n^2 D_k$ ,  $n \neq k$ ,  $A_{k,k} = 0$ ,  $u = (u_1, u_2, \dots)'$ ,  $g = (g_1, g_2, \dots)'$ ,  $A = A_{k,n}$ ,  $\|u\| = \sup_k |u_k|$ ,  $\|A\| = \sup_k \left( \sum_{n=1}^{\infty} |A_{k,n}| \right)$ .

The infinite system of equations (40) can, under the above notation, be put in the form

$$(44) \quad u = g + Au.$$

One can prove the following lemma.

LEMMA 2.  $\|A\| \leq .5681$ .

*Proof.* It immediately follows from the definitions of  $A_{k,n}$  and  $D_k$  that

$$\begin{aligned} \sum_{n=1}^{\infty} |A_{k,n}| &\leq \frac{k}{2\pi} \sum_{\substack{n=1 \\ n \neq k}}^{\infty} \frac{1}{n^2} \left( \frac{1}{|n-k|} + \frac{1}{n+k+1} \right) \\ &= \frac{k}{2} \left( \sum_{n=1}^{\infty} \frac{1}{n^2(n+k+1)} + \sum_{n=1}^{k-1} \frac{1}{n^2(k-n)} + \sum_{n=k+1}^{\infty} \frac{1}{n^2(n-k)} \right). \end{aligned}$$

First observe that

$$\frac{1}{n^2(n+k+1)} = \frac{1}{(k+1)^2} \left( \frac{1}{(n+k+1)} - \frac{1}{n} \right) + \frac{1}{(k+1)n^2};$$

hence

$$\sum_{n=1}^{\infty} \frac{1}{n^2(n+k+1)} = \frac{1}{(k+1)^2} \lim_{M \rightarrow \infty} \left( \sum_{n=1}^M \frac{1}{n+k+1} - \sum_{n=1}^M \frac{1}{n} \right) + \frac{1}{(k+1)} \left( \sum_{n=1}^{\infty} \frac{1}{n^2} \right).$$

The term

$$\sum_{n=1}^M \frac{1}{(n+k+1)} - \sum_{n=1}^M \frac{1}{n} = \sum_{n=k+2}^{M+k+1} \frac{1}{n} - \sum_{n=1}^M \frac{1}{n} = \sum_{n=1}^{M+k+1} \frac{1}{n} - \sum_{n=1}^M \frac{1}{n} - \sum_{n=1}^{k+1} \frac{1}{n}.$$

Using the fact that

$$(45) \quad \lim_{M \rightarrow \infty} \left( \sum_{n=1}^M \frac{1}{n} - \log(M) \right) = \gamma = .55721 \dots$$

and that  $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$ , it follows that

$$(46) \quad \sum_{n=1}^{\infty} \frac{1}{n^2(n+k+1)} = \frac{1}{(k+1)} \pi^2/6 - \frac{1}{(k+1)^2} \left( \sum_{n=1}^{k+1} \frac{1}{n} \right).$$

Next, the term  $1/n^2(k-n) = (1/k^2)(1/n + 1/(k-n)) + 1/kn^2$ .

It follows that

$$(47) \quad \sum_{n=1}^{k-1} \frac{1}{n^2(k-n)} = \frac{2}{k^2} \sum_{n=1}^{k-1} \frac{1}{n} + \frac{1}{k} \sum_{n=1}^{k-1} \frac{1}{n^2}.$$

The term  $1/n^2(n-k) = (1/k^2)(1/(n-k) - 1/n) - (1/k)(1/n^2)$ , and hence

$$(48) \quad \begin{aligned} \sum_{n=k+1}^{\infty} \frac{1}{n^2(n-k)} &= -\frac{1}{k} \sum_{n=k+1}^{\infty} \frac{1}{n^2} + \frac{1}{k^2} \lim_{M \rightarrow \infty} \left( \sum_{n=1}^{M-k} \frac{1}{n} - \sum_{n=k+1}^M \frac{1}{n} \right) \\ &= -\frac{1}{k} \frac{\pi^2}{6} + \frac{1}{k} \sum_{n=1}^k \frac{1}{n^2} + \frac{1}{k^2} \sum_{n=1}^k \frac{1}{n}, \end{aligned}$$

where we have again used (45). Adding the expressions (46), (47), and (48) together, multiplying by  $k/2\pi$ , and rearranging some terms gives

$$(49) \quad \sum_{n=1}^{\infty} |A_{k,n}| \leq B(k),$$

where

$$(50)$$

$$B(k) = \frac{1}{\pi} \left( P(k) + \frac{1}{2(k+1)} P(k+1) + \frac{1}{2} (P(k) - P(k+1)) + \left( \sum_{n=1}^k \frac{1}{n^2} \right) - \frac{3}{2k^2} \right) - \frac{\pi}{12(k+1)},$$

$$(51) \quad P(k) = \frac{1}{k} \sum_{n=1}^k \frac{1}{n}.$$

Equation (45) implies that  $\lim_{k \rightarrow \infty} P(k) = 0$ , and it follows that

$$\lim_{k \rightarrow \infty} B(k) = \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi}{6} = .52359 \dots$$

We have numerically calculated the values of  $B(k)$ ,  $1 \leq k \leq 60$ . A partial listing of these values are given in Table 1.

The values of  $B(k)$ ,  $11 \leq k \leq 60$  are monotone decreasing with

$$.5391015 \leq B(k) \leq .5634350.$$

TABLE 1

<i>k</i>	<i>B(k)</i>
1	.1078327
2	.4786178
3	.5418810
4	.5602098
5	.5662986
6	.5680313
7	.5679856
8	.5671676
9	.5660166
10	.5647388

Moreover, one can show that

$$(52) \quad B(k) \leq \hat{B}(k) = \frac{1}{\pi} \left( \frac{3}{2k} (1 + \log(k)) \right) + \frac{\pi}{6}.$$

The sequence  $\hat{B}(k)$  is monotone decreasing as  $k$  increases and  $\hat{B}(56) = .56643 \leq B(6)$ . Hence, it follows that

$$\sum_{n=1}^{\infty} |A_{k,n}| \leq B(6), \quad 1 \leq k.$$

This completes the proof of Lemma 2.

From the definition

$$\begin{aligned} g_k &= \frac{1}{Dk} k^2 C_k = \frac{k}{2D_k} \left( \frac{\sin((k+1)\alpha)}{k+1} - \frac{\sin(k\alpha)}{k} \right) \\ &= \frac{1}{2D_k} \left( \frac{k}{k+1} (\sin((k+1)\alpha) - \sin(k\alpha)) - \frac{1}{k+1} \sin(k\alpha) \right) \\ &= -\frac{1}{2D_k} \left( \frac{k}{k+1} \left( \sin(k\alpha) 2 \sin\left(\frac{\alpha}{2}\right) + \cos(k\alpha) \sin(\alpha) \right) + \frac{k\alpha}{k+1} \frac{\sin(k\alpha)}{k\alpha} \right) \end{aligned}$$

and the inequalities  $|\sin(x)/x| \leq 1$ ,  $D_k \geq \pi$ , it follows that

$$(53) \quad |g_k| \leq \left( 2 + \frac{\alpha}{2} \right) \frac{\alpha}{2\pi}, \quad 0 \leq \alpha \leq \pi.$$

Lemma 2 and the inequalities (53) imply that there is a unique solution  $u$  of (44) and

$$(54) \quad \|u\| \leq \frac{\|g\|}{1 - \|A\|}.$$

The inequalities (53) and (54) imply that

$$(55) \quad |v_k| \leq \frac{.3685(2 + (1/2)\alpha)\alpha}{k^2}, \quad 0 \leq \alpha \leq \pi.$$

An immediate consequence of (36) and (39) is

$$(56) \quad I(\alpha) = -\frac{(\sigma, v)}{\sqrt{R}} = \frac{1}{2} \left( \alpha - \sin(\alpha) + \sum_{k=1}^{\infty} v_k b_k(\alpha) \right),$$

where

$$(57) \quad b_k(\alpha) = \frac{\sin(k\alpha)}{k} - \frac{\sin((k+1)\alpha)}{k+1}.$$

For  $0 < \alpha < 1$ ,

$$(58) \quad \alpha - \sin(\alpha) = \sum_{j=2}^{\infty} \frac{\alpha^{2j-1}}{(2j-1)!} (-1)^j \cong \frac{\alpha^3}{6} - \frac{\alpha^5}{5!}.$$

The inequalities (55) imply that

$$(59) \quad -|v_1 b_1(\alpha)| \cong -.3685 \left( 2 + \frac{1}{2} \alpha \right) \alpha \left( \sin(\alpha) - \frac{\sin(2\alpha)}{2} \right) \cong -\frac{.3685(2 + (1/2)\alpha)\alpha^4}{2},$$

$$(60) \quad \begin{aligned} -|v_2 b_2(\alpha)| &\cong -\frac{.3685(2 + (1/2)\alpha)\alpha \left( \frac{\sin(2\alpha)}{2} - \frac{\sin(3\alpha)}{3} \right)}{4} \\ &\cong -.3685 \left( 2 + \frac{1}{2} \alpha \right) \alpha^4 \frac{(5 - 8 \sin^2(\alpha/2))}{24}, \quad 0 \leq \alpha \leq 1. \end{aligned}$$

For general  $k$ ,

$$(61) \quad \begin{aligned} \frac{1}{\sqrt{k}} |b_k(\alpha)| &= \left| \frac{\alpha^3}{\sqrt{k}} \left( \frac{k}{2(k+1)} \frac{\sin(k\alpha)}{k\alpha} \left( \frac{\sin(\alpha/2)}{\alpha/2} \right)^2 + \frac{\cos(k\alpha)}{k+1} \left( \frac{1 - \sin(\alpha)/\alpha}{\alpha^2} \right) \right) \right. \\ &\quad \left. + \alpha^{5/2} \frac{k}{k+1} \left( \frac{1}{(\alpha k)^{3/2}} \left( \frac{\sin(k\alpha)}{k\alpha} - \cos(k\alpha) \right) \right) \right|. \end{aligned}$$

By considering the maximum values of the functions  $h_1(x) = (1 - \sin(x)/x/x^2)$  and  $h_2(x) = (1/x^{3/2})(\sin(x)/x - \cos(x))$ ,  $0 < x < \infty$ , one can show that

$$(62) \quad \frac{1}{\sqrt{k}} |b_k(\alpha)| \leq \frac{1}{3} (1 + \alpha) \alpha^{5/2}, \quad k \geq 3, \quad 0 \leq \alpha \leq 1.$$

Moreover

$$(63) \quad \sum_{k=3}^{\infty} k^{-3/2} \leq 3^{-3/2} + \int_3^{\infty} x^{-3/2} dx = (7)3^{-3/2}.$$

The inequalities (59), (60), (62) and (63) yield

$$(64) \quad I(\alpha) = -\frac{(\sigma, v)}{\sqrt{R}} \geq \frac{1\alpha^3}{12} C(\alpha),$$

where

$$(65) \quad C(\alpha) = 1 - .05\alpha^2 - .3685\sqrt{\alpha} \left( 2 + \frac{1}{2} \alpha \right) \left( \frac{17}{4} \sqrt{\alpha} + \frac{14}{3\sqrt{3}} (1 + \alpha) \right).$$

$C(\alpha)$  is monotone decreasing and  $C(.095) = .008829$ . Hence, we now know that  $(\sigma, v) < 0$  for  $0 < \alpha \leq .095$ . To show that  $(\sigma, v) < 0$  for  $.095 < \alpha \leq \pi$ , we first obtain an upper bound for  $|dI/d\alpha|$ .

To obtain an expression for  $dI/d\alpha$ , we proceed as follows. Equation (40) is formally differentiated with respect to  $\alpha$ . We set

$$\begin{aligned}
 u' &= (u'_1, u'_2, \dots)', & u'_k &= kv'_k = k \frac{dv_k}{d\alpha}, \\
 E &= (E_1, E_2, \dots)', \\
 E_k &= -\frac{1}{D_k} \left( \sin^2 \left( \left( k + \frac{1}{2} \right) \alpha \right) + \sin \left( \frac{\alpha}{2} \right) \sin \left( \left( k + \frac{1}{2} \right) \alpha \right) \right. \\
 &\quad \left. + \sum_{\substack{n=1 \\ n \neq k}}^{\infty} \frac{1}{2} (\cos((n-k)\alpha) - \cos((n+k+1)\alpha)) v_n(\alpha) \right), \\
 A'_{k,n} &= -\frac{1}{2D_k n} \left( \frac{\sin((n-k)\alpha)}{n-k} - \frac{\sin((n+k+1)\alpha)}{n+k+1} \right) n \neq k, & A'_{k,k} &= 0,
 \end{aligned}$$

$A' = [A'_{k,n}]$  to write the differentiated form of (40) as

$$(66) \quad u' = E + A' u'.$$

In a manner similar to that used to obtain an upper bound for  $\|A\|$ , one can show that

$$(67) \quad \sum_{n=1}^{\infty} |A'_{k,n}| \leq \frac{1}{2\pi} \left( 3P(k) - P(k+1) - \frac{2}{k^2} \right) \leq .4.$$

Using the bound (55) and the above definition of  $E_k$ , one obtains

$$(68) \quad |E_k| \leq \|E\| \leq \alpha(.78 + .156\alpha).$$

Hence,

$$(69) \quad |u'_k| \leq \|u'\| \leq \frac{\|E\|}{1 - \|A'\|} \leq \alpha(1.3 + .26\alpha)$$

and

$$(70) \quad |v'_k| = \frac{|u'_k|}{k} \leq \alpha \frac{(1.3 + .26\alpha)}{k}.$$

By formally differentiating (56), one obtains

$$(71) \quad \frac{dI}{d\alpha} = \sin^2 \left( \frac{\alpha}{2} \right) + \frac{1}{2} \sum_{k=1}^{\infty} \left( v'_k b_k + v_k \frac{db_k}{d\alpha} \right).$$

It is easy to show that

$$(72) \quad |b_k(\alpha)| = \left| \frac{\sin(k\alpha)}{k} - \frac{\sin((k+1)\alpha)}{k+1} \right| \leq \frac{\alpha}{k+1} (2 + \alpha/2),$$

$$(73) \quad \left| \frac{db_k}{d\alpha} \right| = |\cos(k\alpha) - \cos((k+1)\alpha)| \leq |2 \sin(\alpha/2)|.$$

Using the bounds (55), (70), (72) and (73), equation (71) yields

$$(74) \quad \left| \frac{dI}{d\alpha} \right| \leq J(\alpha) = .6062 \sin(\alpha/2) \alpha (2 + \alpha/2) + \sin^2(\alpha/2) + \alpha^2 (1 + \alpha/4) (1.3 + .26\alpha).$$

Finally, we compute  $I(\alpha)$  at a finite number of points  $\alpha_n$ ,  $1 \leq n \leq M$  with  $\alpha_1 = .095$ ,  $\alpha_{n+1} = \alpha_n + \min(.05, I(\alpha_n)/1.1J(\alpha_n + .05))$ ,  $\alpha_M \cong \pi$ . The computation of  $I(\alpha_n)$  is carried out by truncating system (40) to thirty equations in thirty unknowns and approximating  $I(\alpha)$  by  $\frac{1}{2}(\alpha - \sin(\alpha) + \sum_{k=1}^{30} v_k b_k)$ .

The above described algorithm shows that

$$(75) \quad .59(10^{-4}) \leq I(\alpha_1) \leq I(\alpha_n) < I(\alpha_{n+1}) \leq 1.57.$$

The mean value theorem applied to  $I(\alpha)$  over the subinterval  $[\alpha_n, \alpha_{n+1}]$  yields

$$(76) \quad I(\alpha) = I(\alpha_n) + \frac{dI}{d\alpha}(\hat{\alpha})(\alpha - \alpha_n) \geq I(\alpha_n) - J((\alpha_n + .05)(\alpha - \alpha_n)) \geq I(\alpha_n) \frac{.1}{1.1} > 0,$$

$$\alpha_n \leq \alpha \leq \alpha_{n+1}.$$

Hence  $-(\sigma, v) = \sqrt{R} I(\alpha) \geq 0$ ,  $.095 \leq \alpha \leq \pi$ . This completes the proof of the theorem stated in § 1.

Whether these arguments can be extended to cover the case of a general axial symmetric body and a general setting of the angle  $\alpha$  remains an open question.

#### REFERENCES

- [1] W. D. COLLINS, *On some dual series equations and their application to electrostatic problems for spherical caps*, Proc. Cambridge Philos. Soc., 57 (1961), pp. 367-384.
- [2] ———, *Some scalar diffraction problems for spherical caps*, Arch. Rational Mech. Anal., 10 (1962), pp. 249-266.
- [3] ———, *On some triple series equations and their applications*, Ibid., 11 (1962), pp. 122-137.
- [4] PETER HENRICI, *Complete systems of solutions for a class of elliptic partial differential equations*, Boundary Problems and Differential Equations, Rudolph E. Langer, ed, University of Wisconsin Press, Madison, WI, 1960.
- [5] H. L. JOHNSON, *An integral equation formulation of a mixed boundary value problem on a sphere*, this Journal, 6 (1975), pp. 417-426.
- [6] N. I. MUSKHELISHVILI, *Singular Integral Equations*, P. Noordhoff, Groningen, the Netherlands, 1953.
- [7] MARTIN SCHECHTER, *Mixed boundary problems for general elliptic equations*, Comm. Pure Appl. Math., 13 (1960), pp. 183-201.
- [8] IAN N. SNEDDON, *Mixed Boundary Value Problems in Potential Theory*, North-Holland, Amsterdam, 1966.

## SADDLE POINTS AND RITZ-GALERKIN APPROXIMATIONS\*

VICTOR L. SHAPIRO†

**Abstract.** This paper establishes a theorem for Ritz-Galerkin approximations to a unique saddle point of a function defined on a reflexive Banach space. The functional value at the saddle point is given in terms of an infimum and supremum over two closed, possibly infinite dimensional subspaces whose direct sum determines the original Banach space. Also, an application is given to a nonlinear boundary value problem involving the biharmonic operator.

**1. Introduction.** It is the purpose of this paper to extend the results obtained by Landesman, Lazer and Meyers in [5]. In particular, we intend to establish theorems concerning saddle points similar to those in [5], having the same (or similar) conclusions but under considerably weaker hypotheses. In the final section of the paper, we apply the results established to obtain Ritz-Galerkin approximations to a nonlinear boundary value problem involving the biharmonic operator (a type of problem which arises in elasticity theory, e.g., [4, p. 288]).

Before proceeding, the author would like to acknowledge past conversations on subject matter related to the material in [5] with E. Landesman and his Ph.D. student, I. Walton.

In the sequel,  $W$  (and  $V$ ) will be a real reflexive Banach space and  $\mathcal{B}(W, R^1)$  will be its dual, i.e., the set of real bounded linear functionals defined on  $W$  (see [7]). We shall say the real-valued function  $f$  defined on  $W$  has a  $\mathcal{G}-\mathcal{F}$  derivative at the point  $w$  if there exists a functional  $\nabla f(w)$  in  $\mathcal{B}(W, R^1)$  such that

$$(1.0) \quad \lim_{t \rightarrow 0} [f(w + tw_1) - f(w)]t^{-1} = \nabla f(w)(w_1)$$

for every  $w_1$  in  $W$ . In the sequel, we shall designate the  $\mathcal{G}-\mathcal{F}$  derivative of  $f$  at  $w$  by  $\nabla f(w)$ .

The  $\mathcal{G}-\mathcal{F}$  derivative lies midway between the Fréchet derivative and the Gateaux derivative (and hence the symbol  $\mathcal{G}-\mathcal{F}$ ). In particular, if  $f$  has a Fréchet derivative at  $w$ ,  $f$  has a  $\mathcal{G}-\mathcal{F}$  derivative at  $w$ . Also, if the latter holds, then  $f$  has a Gateaux derivative at  $w$  (see [3, p. 117]).

The first theorem we prove is the following

**THEOREM 1.** *Let  $W$  be a reflexive Banach space with  $W = X \oplus Y$  where  $X$  and  $Y$  are closed subspaces of  $W$ . Suppose that  $f$  is a real-valued function defined on  $W$  which is continuous in the norm topology of  $W$  and has a  $\mathcal{G}-\mathcal{F}$  derivative, designated by  $\nabla f(w)$ , at each point  $w$  of  $W$ . Suppose furthermore that*

- (i) *for each  $x$  in  $X$ ,  $f(x + y)$  is strictly convex on  $Y$ ;*
- (ii) *for each  $y$  in  $Y$ ,  $f(x + y)$  is strictly concave on  $X$ ;*
- (iii) *for each  $x$  in  $X$ ,  $\lim_{\|y\| \rightarrow \infty} f(x + y) = +\infty$ ;*
- (iv) *there is a  $\tilde{y}$  in  $Y$  such that  $\lim_{\|x\| \rightarrow \infty} f(x + \tilde{y}) = -\infty$ .*

*Then the following holds:*

- (a) *there exists one and only one  $w_0$  such that  $\nabla f(w_0) = 0$ ;*
- (b)  $f(w_0) = \sup_{x \text{ in } X} [\inf_{y \text{ in } Y} f(x + y)]$   
 $= \inf_{y \text{ in } Y} [\sup_{x \text{ in } X} f(x + y)];$

\*Received by the editors February 28, 1979, and in revised form October 2, 1979.

†Department of Mathematics, University of California, Riverside, Riverside, California 92521. This work was supported in part by the National Science Foundation under Grant MCS 76-02163-01.



(c) with  $w_0 = x_0 + y_0$ ,  $f(x + y_0) < f(x_0 + y_0) < f(x_0 + y)$  for  $x$  in  $X - \{x_0\}$  and  $y$  in  $Y - \{y_0\}$ .

It is clear that Theorem 1 above is an extension of Theorem 1 in [5] in a number of different ways. In particular, we note that in our Theorem 1 we do not require  $X$  to be finite dimensional.

To be explicit about the concept of strict convexity, (i) above means that if  $\alpha > 0$ ,  $\beta > 0$ ,  $\alpha + \beta = 1$ , and  $y_1$  and  $y_2$  are in  $Y$ , then  $f(x + \alpha y_1 + \beta y_2) < \alpha f(x + y_1) + \beta f(x + y_2)$  (i.e., the inequality is strict). For the definition of  $X \oplus Y$ , see [7, p. 100].

To prove the theorem, we set

$$(1.1) \quad K(x, y) = f(x + y)$$

and observe that

$$(1.2) \quad K \text{ is strongly continuous on } X \times Y,$$

i.e., if  $\|x_n - x\| \rightarrow 0$  and  $\|y_n - y\| \rightarrow 0$  as  $n \rightarrow \infty$ , then  $K(x_n, y_n) \rightarrow K(x, y)$ .

Also, we see from (i) and (ii) that

$$(1.3) \quad K(x, y) \text{ is strictly convex on } Y \text{ for each } x \text{ in } X$$

and

$$(1.4) \quad K(x, y) \text{ is strictly concave on } X \text{ for each } y \text{ in } Y.$$

Next, we intend to establish the following result.

$$(1.5) \quad \text{For each } y \text{ in } Y, K(x, y) \text{ is upper semi-continuous on } X \text{ with respect to the weak topology of } X.$$

Let  $\gamma$  be a fixed but arbitrary real number. It follows from [6, p. 140] that (1.5) will be established if we can show

$$(1.6) \quad \{x : K(x, y) \geq \gamma\} \text{ is a closed set in the weak topology of } X.$$

In order to establish (1.6) and other results in the sequel, we shall need the following fact (see [6, p. 178] or [7, p. 641]).

$$(1.7) \quad \text{If } A \text{ is a convex set in the real Banach space } Z, \text{ then } A \text{ is closed in the norm topology of } Z \text{ if and only if } A \text{ is closed in the weak topology of } Z.$$

It follows from (1.2) that the set in (1.6) is closed in the norm topology of  $X$ . It follows from (1.4) that this same set is a convex set in  $X$ . So (1.6), and hence (1.5), follows from (1.7).

In a similar manner using (1.3) instead of (1.4), we see also that the following fact holds.

$$(1.8) \quad \text{For each } x \text{ in } X, K(x, y) \text{ is lower semi-continuous on } Y \text{ with respect to the weak topology of } Y.$$

Next we define

$$(1.9) \quad G(x) = \inf \{K(x, y) : y \text{ in } Y\}.$$

We propose to show

$$(1.10) \quad \text{for each } x' \text{ in } X, \text{ there is a unique } y', \text{ designated by } y' = \phi(x'), \text{ in } Y \text{ such that } G(x') = K(x', y'). \text{ (Hence, } K(x', y') < K(x', y) \text{ for } y \text{ in } Y \text{ and } y \neq y'.)$$

To establish (1.10), we fix  $x'$  in  $X$  and see from (iii) and (1.1) that  $\lim_{\|y\| \rightarrow \infty} K(x', y) = +\infty$ . Therefore there exists an  $r > 0$  such that if  $\|y\| \geq r$  then

$K(x', 0) < K(x', y)$ . Consequently, it follows from (1.9) that there is a sequence  $\{y_n\}_1^\infty$  with  $\|y_n\| \leq r$  such that  $\lim_{n \rightarrow \infty} K(x', y_n) = G(x')$ . But then it follows from the Eberlein-Smulyan theorem (see [8, p. 141] or [7, p. 86]) that there is a subsequence  $\{y_{n_k}\}_1^\infty$  and a  $y'$  such that  $y_{n_k} \rightarrow y'$  (i.e., in the weak topology of  $Y$ ). We consequently obtain from (1.8) that  $\lim_{k \rightarrow \infty} K(x', y_{n_k}) \cong K(x', y')$ , and therefore that  $K(x', y') \leq G(x')$ . We conclude from (1.9) that  $G(x') = K(x', y')$ . It follows immediately from (1.3) and (1.9) that  $y'$  is unique, and (1.10) is therefore established.

We next show that

(1.11)  $G(x)$  is upper semi-continuous on  $X$  with respect to the weak topology of  $X$ .

To establish this fact, let  $\gamma$  be a real number, and set  $A_\gamma = \{x: G(x) \cong \gamma\}$ . It then follows from (1.9) that

$$A_\gamma = \bigcap_{y \text{ in } Y} \{x: K(x, y) \cong \gamma\}.$$

From (1.2), we see that each set in the intersection above is closed in the norm topology of  $X$ . Consequently, it follows that  $A_\gamma$  is closed in the norm topology of  $X$ . Also, we see that if  $x_1$  and  $x_2$  are in  $A_\gamma$  and  $\alpha + \beta = 1$  and  $\alpha$  and  $\beta > 0$ , then

$$\begin{aligned} G(\alpha x_1 + \beta x_2) &= K[\alpha x_1 + \beta x_2, \phi(\alpha x_1 + \beta x_2)] \\ &\cong \alpha K[x_1, \phi(\alpha x_1 + \beta x_2)] + \beta K[x_2, \phi(\alpha x_1 + \beta x_2)] \\ (1.12) \quad &\cong \alpha G(x_1) + \beta G(x_2) \\ &\cong \gamma \end{aligned}$$

from (1.10), (1.4), and (1.9). Consequently,  $A_\gamma$  is a convex set in  $X$ . We conclude therefore from (1.7) that  $A_\gamma$  is closed in the weak topology of  $X$ , and (1.11) is established.

Next for  $\rho > 0$ , we define

$$(1.13) \quad H_\rho(y) = \sup \{K(x, y): \|x\| \leq \rho\}.$$

It follows immediately from the Eberlein-Smulyan theorem, from (1.5), and from (1.4) that

$$(1.14) \quad \text{for each } y' \text{ in } Y \text{ and } \rho > 0, \text{ there is a unique } x', \text{ designated by } x' = \psi_\rho(y'), \text{ in the set } \{x: \|x\| \leq \rho\}, \text{ such that } H_\rho(y') = K(x', y'). \text{ (Hence } K(x, y') < K(x', y') \text{ for } \|x\| \leq \rho \text{ and } x \neq x'.)$$

Also, using a proof very similar to that used to establish (1.11), we see that the following fact obtains:

$$(1.15) \quad \text{for every } \rho > 0, H_\rho(y) \text{ is lower semi-continuous on } Y \text{ with respect to the weak topology of } Y.$$

Next, we note from (1.9) that  $G(x) \leq K(x, \tilde{y})$  for  $x$  in  $X$ . From (iv) in the hypothesis of the theorem in conjunction with (1.1), we see that  $\lim_{\|x\| \rightarrow \infty} K(x, \tilde{y}) = -\infty$ . Therefore,  $\lim_{\|x\| \rightarrow \infty} G(x) = -\infty$ . Consequently, we see that there is an  $r^* > 0$  such that  $G(x) < G(0)$  for  $\|x\| \cong r^*$ . We conclude that

$$(1.16) \quad \sup_{x \text{ in } X} G(x) = \sup_{\|x\| \leq r^*} G(x).$$

Next, we see from (1.13) that  $K(0, y) \leq H_\rho(y)$  for  $\rho > 0$  and  $y$  in  $Y$ . We consequently obtain from (iii) in the hypothesis of the theorem that  $\lim_{\|y\| \rightarrow \infty} H_\rho(y) =$

$+\infty$  for  $\rho > 0$ . Therefore, there is an  $r_\rho > 0$  such that

$$(1.17) \quad \inf_{y \text{ in } Y} H_\rho(y) = \inf_{\|y\| \leq r_\rho} H_\rho(y).$$

Returning to the  $r^*$  in (1.16), we set

$$(1.18) \quad Q_1 = \{x : \|x\| \leq r^*\},$$

and observe (since  $X$  is a reflexive Banach space) from [6, p. 174] (or [7, p. 105]) that

$$(1.19) \quad Q_1 \text{ is a compact Hausdorff space with respect to the weak topology of } X.$$

We consequently conclude from this last fact, (1.11) and [6, p. 140] that there is an  $x_0$  in  $Q_1$  such that  $G(x) \leq G(x_0)$  for  $x$  in  $Q_1$ . But then it follows from (1.16) and (1.18) that  $\sup_{x \text{ in } X} G(x) = G(x_0)$ . Using (1.9) and (1.10), we record this as follows:

$$(1.20) \quad \sup_{x \text{ in } X} \left[ \inf_{y \text{ in } Y} K(x, y) \right] = \sup_{x \text{ in } Q_1} \left[ \inf_{y \text{ in } Y} K(x, y) \right] = K(x_0, y_0),$$

where  $x_0$  is  $Q_1$  and  $y_0 = \phi(x_0)$ .

Next, we define the set  $E$  as follows:

$$(1.21) \quad E = \{x : K(x_0, y_0) \leq K(x_0 + x, y_0) \text{ and } x \neq 0\}.$$

Also, we define

$$(1.22) \quad r^{**} = \inf \{\|x\| : x \text{ in } E\},$$

and observe that  $r^{**}$  is a finite real number. In particular, we note that if  $E$  is the empty set,  $r^{**} = 0$ .

Continuing with our definitions, we set

$$(1.23) \quad \xi = r^* + r^{**} + 1,$$

and define

$$(1.24) \quad Q = \{x : \|x\| \leq \xi\}.$$

From (1.18), we see that  $Q_1 \subset Q$  and conclude in particular from (1.20) that

$$(1.25) \quad \sup_{x \text{ in } Q} \left[ \inf_{y \text{ in } Y} K(x, y) \right] = K(x_0, y_0).$$

From a theorem of Fan, [2, Thm. 2], we next obtain from (1.3), (1.4), (1.5), and (1.19) with  $Q$  instead of  $Q_1$  that

$$(1.26) \quad \sup_{x \text{ in } Q} \left[ \inf_{y \text{ in } Y} K(x, y) \right] = \inf_{y \text{ in } Y} \left[ \sup_{x \text{ in } Q} K(x, y) \right].$$

Using the same  $\xi$  as in (1.23), we see from (1.17) that there is an  $r' > 0$  such that

$$(1.27) \quad \inf_{y \text{ in } Y} H_\xi(y) = \inf_{\|y\| \leq r'} H_\xi(y).$$

Now using [6, p. 174] (or [7, p. 105]) once again, we see that  $\{y : \|y\| \leq r'\}$  is a compact Hausdorff space in the weak topology of  $Y$ . Since  $H_\xi(y)$  is lower semi-continuous on  $Y$  [see (1.15)], we conclude from [6, p. 140] that there is a  $y_1$  with  $\|y_1\| \leq r'$  such that

$\inf_{\|y\| \leq r} H_\xi(y) = H_\xi(y_1)$ . Using (1.14) and (1.17), we record this as follows:

$$(1.28) \quad \inf_{y \text{ in } Y} H_\xi(y) = K(x_1, y_1), \quad \text{where } x_1 = \psi_\xi(y_1) \text{ and } \|x_1\| \leq \xi.$$

From (1.13) and (1.24), we see that the right-hand side of (1.26) is  $\inf_{y \text{ in } Y} H_\xi(y)$ . Consequently, we see from (1.28) that the right-hand side of (1.26) is equal to  $K(x_1, y_1)$ . But from (1.25), we see that the left-hand side of (1.26) is equal to  $K(x_0, y_0)$ ; we conclude

$$(1.29) \quad K(x_0, y_0) = K(x_1, y_1).$$

Now  $y_0 = \phi(x_0)$  and  $x_1 = \psi_\xi(y_1)$ . We therefore obtain from (1.10) that  $K(x_0, y_0) \leq K(x_0, y_1)$  with equality if and only if  $y_0 = y_1$ . From (1.20), we see that  $x_0$  is in  $Q_1 \subset Q$ , and therefore  $\|x_0\| \leq \xi$ . Consequently, we obtain from (1.14) that  $K(x_0, y_1) \leq K(x_1, y_1)$ . We conclude that

$$K(x_0, y_0) \leq K(x_0, y_1) \leq K(x_1, y_1).$$

Using (1.29), we see from this last fact that  $K(x_0, y_0) = K(x_0, y_1)$ . But then, as stated in the preceding paragraph, this implies that  $y_0 = y_1$ .

From the fact that  $y_0 = y_1$ , we obtain from (1.29) that  $K(x_0, y_1) = K(x_1, y_1)$ . But  $x_1 = \psi_\xi(y_1)$  and  $\|x_0\| \leq \xi$ . So using (1.14) once again, we conclude also that  $x_0 = x_1$ . We record this fact as follows:

$$(1.30) \quad K(x, y_0) < K(x_0, y_0) \quad \text{for } \|x\| \leq \xi \text{ and } x \neq x_0.$$

Also, using (1.20) and (1.10), we record the following fact:

$$(1.31) \quad \begin{aligned} K(x_0, y_0) &< K(x_0, y) \quad \text{for } y \text{ in } Y - \{y_0\}. \\ \text{Also } \|x_0\| &\leq r^*. \end{aligned}$$

Next, we set  $w_0 = x_0 + y_0$ . By hypothesis, the  $\mathcal{G}$ - $\mathcal{F}$  derivative of  $f$  exists at  $w_0$ . We consequently conclude from (1.31) and (1.1) that

$$(1.32) \quad \lim_{t \rightarrow 0} [f(w_0 + ty) - f(w_0)]t^{-1} = 0 \quad \text{for } y \text{ in } Y.$$

From (1.23), we see that  $\xi \geq r^* + 1$ . From (1.31), we also see that  $\|x_0\| \leq r^*$ . Therefore given any  $x$  in  $X$ , we conclude that  $\|x_0 + tx\| \leq \xi$  for  $|t|$  sufficiently small (depending on  $x$ ). Consequently, we obtain from (1.30) that

$$\lim_{t \rightarrow 0} [f(w_0 + tx) - f(w_0)]t^{-1} = 0 \quad \text{for } x \text{ in } X.$$

From this last fact and (1.0), we obtain  $\nabla f(w_0)(x) = 0$  for  $x$  in  $X$ . From (1.32) and (1.0), we obtain  $\nabla f(w_0)(y) = 0$  for  $y$  in  $Y$ . Since  $W = X \oplus Y$  and  $\nabla f(w_0)$  is in  $\mathcal{B}(W, \mathbf{R}^1)$ , we conclude from these last two facts that  $\nabla f(w_0)(w) = 0$  for  $w$  in  $W$ . We record this as follows:

$$(1.33) \quad \nabla f(w_0) = 0, \quad \text{where } w_0 = x_0 + y_0.$$

Suppose

$$(1.34) \quad \nabla f(w_2) = 0, \quad \text{where } w_2 = x_2 + y_2.$$

We see from (1.33) that the proof of (a) in the conclusion of the theorem will be

complete, once we show

$$(1.35) \quad w_0 = w_2.$$

To establish this last fact, we next show that the set  $E$  defined in (1.21) satisfies the following:

$$(1.36) \quad E \text{ is the empty set.}$$

Suppose not. Then from (1.22), we see there exists an  $x_3$  in  $E$  with  $0 < \|x_3\| \leq r^{**} + \frac{1}{2}$ . Since  $\|x_0\| \leq r^*$ , we see from (1.23) that  $\|x_0 + x_3\| \leq \xi$ . We obtain, consequently, from (1.30) that  $K(x_0 + x_3, y_0) < K(x_0, y_0)$ . Therefore  $x_3$  is not in  $E$ , and (1.36) is established.

Using (1.36) in conjunction with (1.1), (1.21), and (1.31), we obtain that  $w_0 = x_0 + y_0$  is a strict global saddle point for  $f$ , i.e.,

$$(1.37) \quad f(x_0 + y_0) < f(x_0 + y) \quad \text{for } y \text{ in } Y \text{ and } y \neq y_0$$

and

$$f(x + y_0) < f(x_0 + y_0) \quad \text{for } x \text{ in } X \text{ and } x \neq x_0.$$

We propose to show that  $w_2 = x_2 + y_2$ , defined in (1.34), is a strict global saddle point for  $f$  of the same nature, i.e.,

$$(1.38) \quad f(x_2 + y_2) < f(x_2 + y) \quad \text{for } y \text{ in } Y \text{ and } y \neq y_2$$

and

$$f(x + y_2) < f(x_2 + y_2) \quad \text{for } x \text{ in } X \text{ and } x \neq x_2.$$

To establish the first inequality in (1.38), we suppose there exists a  $y_3$  such that

$$(1.39) \quad f(x_2 + y_3) \leq f(x_2 + y_2), \quad \text{where } y_3 \neq y_2.$$

Next, we set

$$(1.40) \quad q(t) = f[x_2 + y_2 + t(y_3 - y_2)].$$

From condition (i) in the hypothesis of the theorem, it follows that  $q$  is a convex function for  $-\infty < t < \infty$ . From the fact that  $f$  has a  $\mathcal{G} - \mathcal{F}$  derivative everywhere in  $W$ , it follows from (1.0) and (1.40) that  $dq/dt$  exist for all  $t$  in  $(-\infty, \infty)$  and furthermore that

$$(1.41) \quad dq(t)/dt = \nabla f[w_2 + t(y_3 - y_2)](y_3 - y_2).$$

Since  $q$  is convex on  $(-\infty, \infty)$ , we obtain that  $dq/dt$  is a nondecreasing function on  $(-\infty, \infty)$  (see [9, p. 22]). From (1.34) and (1.41), we see that  $dq(t)/dt$  evaluated at  $t = 0$  is itself 0. Consequently,  $dq(t)/dt \geq 0$  for  $0 \leq t < \infty$ , and therefore  $q(t)$  is a nondecreasing function for  $0 \leq t < \infty$ . In particular,  $q(0) \leq q(\frac{1}{2}) \leq q(1)$ , and we obtain from (1.40) and (1.39) that

$$f(x_2 + y_2) \leq f[x_2 + (y_2 + y_3)/2] \leq f(x_2 + y_3) \leq f(x_2 + y_2).$$

Consequently,

$$f[x_2 + (y_2 + y_3)/2] = 2^{-1}[f(x_2 + y_2) + f(x_2 + y_3)].$$

But this last contradicts condition (i) in the hypothesis of the theorem. We conclude that (1.39) does not hold, and consequently the first inequality in (1.38) is established. A similar proof using condition (ii) in the hypothesis of the theorem establishes the second inequality in (1.38).

It is a simple matter to conclude from (1.37) and (1.38) that  $x_0 = x_2$  and  $y_0 = y_2$ . (See [5, p. 595].) Equation (1.35) is therefore established, and the proof of (a) in the conclusion of the theorem is complete.

Inequality (1.37) gives part (c) in the conclusion of the theorem. It remains to prove part (b). To do this, we set

$$(1.42) \quad H(y) = \sup_{x \in X} f(x + y).$$

It follows from (1.37) that  $H(y_0) = f(w_0)$ . Suppose there exists  $y_3$  such that  $H(y_3) < f(w_0)$ . Then from (1.42) and (1.37), we have  $f(x_0 + y_3) \leq H(y_3) < f(x_0 + y_0) < f(x_0 + y_3)$ . This is a clear contradiction, and we conclude  $H(y) \geq f(w_0)$  for all  $y$  in  $Y$  with equality when  $y = y_0$ . This fact, coupled with (1.1) and (1.20) gives part (b) in the conclusion of the theorem, and the proof of the theorem is complete.

**2. A fundamental lemma.** For a real-valued function  $g$  defined on a reflexive Banach space  $V$ , we shall set

$$(2.1) \quad \bar{D}^2 g(v, v_1) = \limsup_{t \rightarrow 0} [g(v + tv_1) + g(v - tv_1) - 2g(v)]t^{-2}.$$

$\underline{D}^2 g(v, v_1)$  will be defined analogously using  $\lim \inf$ .

In the sequel, we shall need the following lemma.

LEMMA 1. *Let  $V$  be a reflexive Banach space, and let  $g$  be a real-valued  $\mathcal{G} - \mathcal{F}$  differentiable function on  $V$  which is continuous in the norm topology of  $V$ . Suppose there is a positive constant  $k$  such that for every  $v$  and  $v_1$  in  $V$ ,  $\bar{D}^2 g(v, v_1) \geq k\|v_1\|^2$ . Then the following holds:*

- (a)  $g$  is strictly convex on  $V$ ;
- (b)  $\nabla g(v + v_1)(v_1) - \nabla g(v)(v_1) \geq k\|v_1\|^2$  for  $v$  and  $v_1$  in  $V$ ;
- (c)  $\lim_{\|v\| \rightarrow \infty} g(v) = +\infty$ .

To prove part (a) of the lemma, let  $v_2$  and  $v_3$  be arbitrary but fixed points in  $V$  with  $v_2 \neq v_3$ . Set

$$(2.2) \quad q(t) = g[v_2 + t(v_3 - v_2)] - k\|v_3 - v_2\|^2 t^2 / 2.$$

Then from the hypothesis of the lemma and (1.0), we see that  $q(t)$  is a differentiable and continuous function for  $t$  in the infinite interval  $(-\infty, \infty)$ , and

$$(2.3) \quad Dq(t) = \nabla g[v_2 + t(v_3 - v_2)](v_3 - v_2) - k\|v_3 - v_2\|^2 t$$

where  $Dq = dq/dt$ .

Following the notation in [9, p. 23], we set

$$\bar{D}^2 q(t) = \limsup_{s \rightarrow 0} [q(t + s) + q(t - s) - 2q(t)]s^{-2}$$

with  $\underline{D}^2 q(t)$  defined similarly using  $\lim \inf$ . An easy computation from (2.1) and (2.2) shows that

$$\bar{D}^2 q(t) = \bar{D}^2 g[v_2 + t(v_3 - v_2), v_3 - v_2] - k\|v_3 - v_2\|^2.$$

Consequently, it follows from the hypothesis of the lemma that  $\bar{D}^2 q(t) \geq 0$  for  $-\infty < t < \infty$ . But then it follows from [9, p. 23] that  $q(t)$  is a convex function for  $-\infty < t < \infty$ . Therefore if  $\alpha > 0$ ,  $\beta > 0$ , and  $\alpha + \beta = 1$ , we have  $q(\alpha \cdot 0 + \beta \cdot 1) \leq \alpha q(0) + \beta q(1)$ . From (2.2), we see this is the same as  $g(\alpha v_2 + \beta v_3) - k\|v_3 - v_2\|^2 \beta^2 2^{-1} \leq \alpha g(v_2) + \beta g(v_3) - k\|v_3 - v_2\|^2 2^{-1}$ . Since  $0 < \beta < 1$ , we conclude that  $g(\alpha v_2 + \beta v_3) < \alpha g(v_2) + \beta g(v_3)$ , and (a) of the lemma is established.

To show that (b) holds, we assume without loss of generality that  $v_1 \neq 0$ , and take  $v_2 = v$  and  $v_3 = v_1 + v_2$  in (2.2). Since  $q(t)$  is convex on  $(-\infty, \infty)$ , we see from [9, p. 22] that  $Dq(t)$  is a nondecreasing function. In particular,  $Dq(0) \leq Dq(1)$ . We obtain therefore from (2.3) that

$$\nabla g(v)(v_1) \leq \nabla g(v + v_1)(v_1) - k\|v_1\|^2.$$

Condition (b) follows immediately from this fact.

To establish (c), we take  $v_2 = 0$  and  $v_3 = v$  in (2.2). Since  $q(t)$  is convex on  $(-\infty, \infty)$ , we have once again from [9, p. 22] that  $Dq(t)$  is nondecreasing. In particular,  $Dq(0) \leq Dq(t)$  for  $t \geq 0$ . Consequently, we see from (2.3) that  $Dq(t) = \nabla g(tv)(v) - k\|v\|^2 t$  and therefore that

$$(2.4) \quad \nabla g(0)(v) \leq \nabla g(tv)(v) - k\|v\|^2 t.$$

Next, we set  $p(t) = g(tv)$ . Then  $p(t)$  is a differentiable and continuous function on  $(-\infty, \infty)$ , and using (1.0) we obtain that  $Dp(t) = \nabla g(tv)(v)$ . Also, from (2.4), we see that  $p(1) - p(0) = \int_0^1 Dp(t) dt \geq \int_0^1 [\nabla g(0)(v) + k\|v\|^2 t] dt$ . Consequently,

$$g(0) + \nabla g(0)(v) + k\|v\|^2 2^{-1} \leq g(v)$$

for  $v$  in  $V$ . Since  $\nabla g(0)$  is in  $\mathcal{B}(V, R^1)$  and  $k$  is a positive constant, (c) follows immediately from this last inequality. The proof of the lemma is therefore complete.

**3. Monotone convergence and convergence.** Let  $W = X \oplus Y$  be as in Theorem 1 and suppose that  $f$  is  $\mathcal{G} - \mathcal{F}$  differentiable on  $W$ . We shall say  $\nabla f(x + y)$  is locally pointwise Lipschitz in  $X$  for each  $y$  in  $Y$  if the following prevails:

(3.1) for every  $x$  in  $X$  and  $y$  in  $Y$ , positive constants  $M(x, y)$  and  $\delta(x, y)$  exist such that if  $\|x_1\| \leq \delta(x, y)$ , then

$$\|\nabla f(x + x_1 + y) - \nabla f(x + y)\| \leq M(x, y)\|x_1\|.$$

We next intend to prove the following theorem, which is to be compared with [5, Thm. 2].

**THEOREM 2.** *Let  $W$  be a reflexive Banach space with  $W = X \oplus Y$  where  $X$  and  $Y$  are closed subspaces of  $W$ . Suppose that  $f$  is a real-valued function defined on  $W$  which is continuous in the norm topology of  $W$  and has a  $\mathcal{G} - \mathcal{F}$  derivative, designated by  $\nabla f(w)$ , at each point  $w$  of  $W$ . Suppose, furthermore, that*

- (i) *there is a positive constant  $k$  such that  $\bar{D}^2 f(w, y_1) \geq k\|y_1\|^2$  for every  $w$  in  $W$  and  $y_1$  in  $Y$ ;*
- (ii) *for each  $y$  in  $Y$ ,  $f(x + y)$  is strictly concave on  $X$ ;*
- (iii) *there is a  $\tilde{y}$  in  $Y$  such that  $\lim_{\|x\| \rightarrow \infty} f(x + \tilde{y}) = -\infty$ ;*
- (iv)  *$\nabla f(x + y)$  is locally pointwise Lipschitz in  $X$  for each  $y$  in  $Y$ ;*
- (v) *there is an increasing sequence  $\{X_n\}_1^\infty$  of finite dimensional subspaces of  $X$  such that  $\cup_{n=1}^\infty X_n$  is dense in  $X$ .*

Set  $W_n = X_n \oplus Y$ . Then the following prevails:

- (a) *there is a unique  $w_n$  in  $W_n$  such that  $\nabla f(w_n)(w) = 0$  for  $w$  in  $W_n$ ;*
- (b) *there is a unique  $w_0$  in  $W$  such that  $\nabla f(w_0)(w) = 0$  for  $w$  in  $W$ ;*
- (c)  *$f(w_n) \leq f(w_{n+1})$  for  $n = 1, 2, \dots$ ;*
- (d)  *$\lim_{n \rightarrow \infty} f(w_n) = f(w_0)$ .*

It follows from Lemma 1 and condition (i) in the hypothesis of the above theorem that  $f(x + y)$  is strictly convex on  $Y$  for each  $x$  in  $X$  and also that  $\lim_{\|y\| \rightarrow \infty} f(x + y) = +\infty$  for each  $x$  in  $X$ . Consequently, it follows from Theorem 1 and the hypothesis of Theorem 2 that both (a) and (b) in the conclusion of Theorem 2 hold. In particular, with

$K(x, y)$ ,  $G(x)$ , and  $\phi(x)$  as in (1.1), (1.9), and (1.10) respectively, we see from (b) and (c) in Theorem 1 that

$$(3.2) \quad f(w_n) = \sup_{x \text{ in } X_n} G(x) = G(x_n)$$

and

$$(3.3) \quad w_n = x_n + \phi(x_n), \quad \text{where } G(x_n) = K[x_n, \phi(x_n)],$$

$n = 0, 1, 2, \dots$ , where for convenience we are now calling  $X = X_0$ . Since  $X_n \subset X_{n+1}$  for  $n = 1, 2, \dots$ , conclusion (c) in Theorem 2 follows immediately from (3.2).

It remains to establish (d) in the conclusion of Theorem 2. To accomplish this, we use the finite-dimensionality of  $X_n$  and select a sequence of points  $\{x'_n\}_{n=1}^\infty$  with  $x'_n$  in  $X_n$  such that  $\|x'_n\| = \inf \{\|x - x_0\| : x \text{ in } X_n\}$ . Then, it follows from condition (v) that

$$(3.4) \quad \lim_{n \rightarrow \infty} \|x'_n - x_0\| = 0.$$

Next, we observe from (3.2) that  $G(x'_n) \leq G(x_n) \leq G(x_0)$  for  $n = 1, 2, \dots$ . From (1.1), (3.2) and (3.3), we see therefore that

$$(3.5) \quad f[x'_n + \phi(x'_n)] \leq f(w_0) = f[x_0 + \phi(x_0)].$$

If we can show, using the norm topology in  $X$  and  $Y$ , respectively, that

$$(3.6) \quad \phi \text{ is a continuous mapping of } X \text{ into } Y,$$

then (d) in the conclusion of Theorem 2 will follow immediately from (3.4), (3.5) and the fact that  $f$  is continuous in the norm topology on  $W$ .

We now establish (3.6) by fixing a point  $x'$  in  $X$  and showing that  $\phi$  is continuous at  $x'$ . We recall from (1.1), (1.9), and (1.10) that  $\phi(x')$  is the unique point  $y'$  in  $Y$  such that  $\inf_{y \text{ in } Y} f(x' + y) = f[x' + \phi(x')]$ . Since  $f[x' + \phi(x')] \leq f(x' + y)$  for  $y$  in  $Y$ , we see from (1.0) that

$$\nabla f[x' + \phi(x')](y) = 0 \quad \text{for } y \text{ in } Y.$$

In particular, we obtain from this last stated fact that for  $x$  in  $X$ ,

$$\begin{aligned} 0 &= \nabla f[x' + x + \phi(x' + x)][\phi(x' + x) - \phi(x')] \\ &= \nabla f[x' + \phi(x')][\phi(x' + x) - \phi(x')]. \end{aligned}$$

This in turn tells us with  $y' = \phi(x')$  and

$$(3.7) \quad \Delta\phi(x', x) = \phi(x' + x) - \phi(x')$$

that

$$\begin{aligned} &\{\nabla f[x' + y'] - \nabla f[x' + x + y']\}[\Delta\phi(x', x)] \\ &= \{\nabla f[x' + x + y' + \phi(x' + x) - \phi(x')] - \nabla f[x' + x + y']\}[\Delta\phi(x', x)]. \end{aligned}$$

Let  $\|x\| \leq \delta(x', y')$ . Then we conclude from this last computation, (3.1), and condition (iv) that

$$(3.8) \quad \begin{aligned} &\{\nabla f[x' + x + y' + \Delta\phi(x', x)] - \nabla f[x' + x + y']\}[\Delta\phi(x', x)] \\ &\leq M(x', y')\|x\| \|\Delta\phi(x', x)\|. \end{aligned}$$

Next, using Lemma 1, we identify  $V$  with  $Y$ ,  $g(\cdot)$  with  $f(x' + x + \cdot)$ , and conclude



from condition (i) in the hypothesis of Theorem 2 and (3.8) that

$$k\|\Delta\phi(x', x)\|^2 \leq M(x', y)\|x\| \|\Delta\phi(x', x)\|.$$

We consequently obtain from this last fact and (3.7) that

$$(3.9) \quad \|\phi(x' + x) - \phi(x')\| \leq M(x', y)k^{-1}\|x\|$$

for  $\|x\| \leq \delta(x', y)$ . Since  $M(x', y)$  and  $k$  are positive constants, we conclude that  $\phi$  is continuous at  $x'$ . Condition (3.6) is therefore established, and the proof of Theorem 2 is complete.

Let  $W$  be as in Theorem 1 and suppose  $f$  is  $\mathcal{G} - \mathcal{F}$  differentiable in  $W$ . We shall say  $\nabla f$  is locally pointwise Lipschitz in  $W$  if the following prevails.

$$(3.10) \quad \text{for each } w \text{ in } W, \text{ there are positive constants } M(w) \text{ and } \delta(w) \text{ such that if } \|w_1\| \leq \delta(w) \text{ then}$$

$$\|\nabla f(w + w_1) - \nabla f(w)\| \leq M(w)\|w_1\|.$$

It is clear that if  $f$  is twice Fréchet differentiable in  $W$ , then  $\nabla f$  is locally pointwise Lipschitz on  $W$ . (See [3, p. 129].) Also, it is clear that this latter condition implies  $\nabla f(x + y)$  is pointwise Lipschitz on  $X$  for each  $y$  in  $Y$ , i.e., (3.10) implies (3.1).

Using Theorem 2, we shall next establish the following theorem.

**THEOREM 3.** *Let  $W$  be a reflexive Banach space with  $W = X \oplus Y$ , where  $X$  and  $Y$  are closed subspaces of  $W$ . Suppose that  $f$  is a real-valued function defined on  $W$  which is continuous in the norm topology of  $W$  and has a  $\mathcal{G} - \mathcal{F}$  derivative, designated by  $\nabla f(w)$ , at each point  $w$  of  $W$ . Suppose, furthermore, that*

- (i) *there is a positive constant  $k_1$  such that  $\bar{D}^2 f(w, y_1) \geq k_1\|y_1\|^2$  for every  $w$  in  $W$  and  $y_1$  in  $Y$ ;*
- (ii) *there is a positive constant  $k_2$  such that  $\underline{D}^2 f(w, x_1) \leq -k_2\|x_1\|^2$  for every  $w$  in  $W$  and  $x_1$  in  $X$ ;*
- (iii)  *$\nabla f(w)$  is locally pointwise Lipschitz in  $W$ ;*
- (iv) *there are two increasing sequences  $\{X_n\}_1^\infty$  and  $\{Y_n\}_1^\infty$  of finite dimensional subspaces of  $X$  and  $Y$  respectively such that  $\cup_{n=1}^\infty X_n$  is dense in  $X$  and  $\cup_{n=1}^\infty Y_n$  is dense in  $Y$ .*

Set  $W_n = X_n \oplus Y_n$ . Then the following prevails:

- (a) *there is a unique  $w_n$  in  $W_n$  such that  $\nabla f(w_n)(w) = 0$  for  $w$  in  $W_n$ ;*
- (b) *there is a unique  $w_0$  in  $W$  such that  $\nabla f(w_0)(w) = 0$  for  $w$  in  $W$ ;*
- (c)  *$\lim_{n \rightarrow \infty} f(w_n) = f(w_0)$ .*

To establish Theorem 3, we set

$$(3.11) \quad W_n^* = X_n \oplus Y \quad \text{and} \quad W_n^{**} = X \oplus Y_n.$$

Next, we notice that (i) in Theorem 3 is the same as (i) in Theorem 2; that (ii) in Theorem 3 and (a) and (c) of Lemma 1 imply that (ii) and (iii) of Theorem 2 holds; and that (iii) and (iv) of Theorem 3 imply that (iv) and (v) of Theorem 2 hold. Consequently, we see from Theorem 2 that

$$(3.12) \quad \text{there is a unique } w_n^* \text{ in } W_n^* \text{ such that } \nabla f(w_n^*)(w) = 0 \quad \text{for } w \text{ in } W_n^*.$$

Since (ii) in Theorem 3 and (c) of Lemma 1 imply  $\lim_{\|x\| \rightarrow \infty} f(x + y) = -\infty$  for every  $y$  in  $Y$ , we see also from Theorem 2 that both (a) and (b) of Theorem 3 hold. Also, we see

from Theorem 2 that

$$(3.13) \quad \begin{aligned} &\text{there is a unique } w_n^{**} \text{ in } W_n^{**} \text{ such that} \\ &\nabla f(w_n^{**})(w) = 0 \quad \text{for every } w \text{ in } W_n^{**}. \end{aligned}$$

Furthermore, we obtain from (d) in Theorem 2 that

$$(3.14) \quad \lim_{n \rightarrow \infty} f(w_n^*) = \lim_{n \rightarrow \infty} f(w_n^{**}) = f(w_0).$$

To establish (c) of Theorem 3, we see from (3.11) and (b) of Theorem 1 that  $f(w_n^*) = \sup_{x \text{ in } X_n} [\inf_{y \text{ in } Y} f(x + y)]$ ,  $f(w_n) = \sup_{x \text{ in } X_n} [\inf_{y \text{ in } Y_n} f(x + y)] = \inf_{y \text{ in } Y_n} [\sup_{x \text{ in } X_n} f(x + y)]$ , and  $f(w_n^{**}) = \inf_{y \text{ in } Y_n} [\sup_{x \text{ in } X} f(x + y)]$ . Consequently,

$$(3.15) \quad f(w_n^*) \leq f(w_n) \leq f(w_n^{**}).$$

But then (c) of Theorem 3 follows immediately from this last fact and (3.14), and the proof of Theorem 3 is complete.

**4. Convergence of the Ritz-Galerkin approximations.** In this section, we intend to extend Theorem 4 in [5]. The first theorem we prove in this direction is the following:

**THEOREM 4.** *Let  $W$  be a reflexive Banach space with  $W = X \oplus Y$ , where  $X$  and  $Y$  are closed subspaces of  $W$ . Suppose that  $f$  is a real-valued function defined on  $W$  which is continuous in the norm topology of  $W$  with a  $\mathcal{G} - \mathcal{F}$  derivative at each point  $w$  of  $W$ . Suppose, furthermore, that*

- (i) *there is a positive constant  $k_1$  such that  $\bar{D}^2 f(w, y_1) \geq k_1 \|y_1\|^2$  for every  $w$  in  $W$  and  $y_1$  in  $Y$ ;*
- (ii) *there is a positive constant  $k_2$  such that  $\underline{D}^2 f(w, x_1) \leq -k_2 \|x_1\|^2$  for every  $w$  in  $W$  and  $x_1$  in  $X$ ;*
- (iii)  *$\nabla f(w)$  is locally pointwise Lipschitz in  $W$ ;*
- (iv) *there is an increasing sequence  $\{X_n\}_1^\infty$  of finite-dimensional subspaces of  $X$  such that  $\cup_{n=1}^\infty X_n$  is dense in  $X$ .*

Set  $W_n = X_n \oplus Y$ . Then the following prevails:

- (a) *there is a sequence  $\{w_n\}_1^\infty$  and a  $w_0 = x_0 + y_0$  satisfying conditions (a), (b), (c), and (d) in the conclusion of Theorem 2;*
- (b) *for each  $n$ , choose  $x'_n$  in  $X_n$  such that*

$$\|x'_n - x_0\| = \inf \{\|x - x_0\| : x \text{ in } X_n\}.$$

*Then there exist constants  $c_1 > 0$  and  $c_2 > 0$  such that*

$$c_1 \|x_0 - x'_n\| \leq \|w_0 - w_n\| \leq c_2 \|x_0 - x'_n\|$$

*for all  $n$ .*

From conditions (i) and (ii) in the hypothesis of the above theorem and from (a) and (c) of Lemma 1, we see that

$$(4.1) \quad f(x + y) \text{ is strictly concave in } X \text{ (convex in } Y) \text{ for each } y \text{ in } Y \text{ (for each } x \text{ in } X);$$

$$(4.2) \quad \begin{aligned} \lim_{\|x\| \rightarrow \infty} f(x + y) &= -\infty \quad \text{for each } y \text{ in } Y, \\ \lim_{\|y\| \rightarrow \infty} f(x + y) &= +\infty \quad \text{for each } x \text{ in } X. \end{aligned}$$

In particular, we see that all the conditions in the hypothesis of Theorem 2 are met and

consequently (a) above in the conclusion of Theorem 4 holds. It remains to establish (b).

To accomplish this fact, we first of all see from (4.1), (4.2), and (b) of Theorem 1 that

$$(4.3) \quad f(w_0) = \sup_{x \text{ in } X} [\inf_{y \text{ in } Y} f(x + y)]$$

and for  $n = 1, 2, \dots$ ,

$$(4.4) \quad f(w_n) = \inf_{y \text{ in } Y} \sup_{x \text{ in } X_n} [f(x + y)].$$

Next, with  $w_n = x_n + y_n$ , we write

$$(4.5) \quad f(w_0) - f(w_n) = [f(x_0 + y_0) - f(x_n + y_0)] + [f(x_n + y_0) - f(x_n + y_n)]$$

and set

$$(4.6) \quad q_1(t) = f[x_n + y_n + t(y_0 - y_n)] - k_1 \|y_n - y_0\|^2 t^2 2^{-1}.$$

It follows from the hypothesis of Theorem 4 that  $q_1(t)$  is a differentiable function for  $-\infty < t < \infty$ . Setting  $Dq_1(t) = dq_1(t)/dt$ , we see that

$$(4.7) \quad Dq_1(t) = \nabla f[x_n + y_n + t(y_0 - y_n)](y_0 - y_n) - k_1 \|y_n - y_0\|^2 t.$$

Also, we see from (2.1) and (4.6) that

$$\bar{D}^2 q_1(t) = \bar{D}^2 f[x_n + y_n + t(y_0 - y_n), y_0 - y_n] - k_1 \|y_n - y_0\|^2.$$

Consequently, we see from condition (i) in the hypothesis of the current theorem that  $\bar{D}^2 q_1(t) \geq 0$  for  $-\infty < t < \infty$ . Since  $q_1$  is also a continuous function of  $t$ , it follows from [9, p. 23] that  $q_1$  is a convex function on  $(-\infty, \infty)$  and, therefore, from [9, p. 22] that  $Dq_1(t)$  is a nondecreasing function. In particular,

$$(4.8) \quad Dq_1(0) \leq Dq_1(t) \quad \text{for } 0 \leq t < \infty.$$

From (4.7), we see that  $Dq_1(0) = \nabla f(x_n + y_n)(y_0 - y_n)$ . Now,  $y_n - y_0$  is in  $Y$ , and therefore in  $W_n$  for each  $n$ . Consequently, it follows from (a) of Theorems 2 and 4 that  $Dq_1(0) = 0$ . We obtain, therefore, from (4.8) that  $q_1(0) \leq q_1(1)$ . This implies in turn from (4.6) that

$$(4.9) \quad f(x_n + y_n) + k_1 \|y_n - y_0\|^2 2^{-1} \leq f(x_n + y_0).$$

Next, we set

$$(4.10) \quad q_2(t) = f[x_0 + y_0 + t(x_n - x_0)] + k_2 \|x_n - x_0\|^2 t^2 2^{-1}$$

and observe that  $Dq_2(t) = \nabla f[x_0 + y_0 + t(x_n - x_0)](x_n - x_0) + k_2 \|x_n - x_0\|^2 t$ ,  $\underline{D}^2 q_2(t) = \underline{D}^2 f[x_0 + y_0 + t(x_n - x_0), x_n - x_0] + k_2 \|x_n - x_0\|^2$ , and  $Dq_2(0) = \nabla f(w_0)(x_n - x_0)$ . We conclude from condition (ii) in the hypothesis of Theorem 4 that  $q_2(t)$  is concave on  $(-\infty, \infty)$  and from (b) in Theorem 2 that  $Dq_2(0) = 0$ . Therefore (in a manner similar to that involving  $q_1$ ) we obtain  $q_2(0) \geq q_2(1)$ . Consequently, we see from (4.10) that

$$f(x_n + y_0) + k_2 \|x_n - x_0\|^2 2^{-1} \leq f(x_0 + y_0).$$

From (4.5), (4.9) and this last fact, we obtain that

$$(4.11) \quad [k_2 \|x_n - x_0\|^2 + k_1 \|y_n - y_0\|^2] 2^{-1} \leq f(w_0) - f(w_n).$$

We conclude from (a) of Theorem 4, (d) of Theorem 2 and (4.11) that

$$(4.12) \quad \lim_{n \rightarrow \infty} [\|x_n - x_0\| + \|y_n - y_0\|] = 0.$$

Next, we notice from (4.1), (4.2), (4.4) and (c) of Theorem 1 applied to  $W_n$  that

$$(4.13) \quad f(x'_n + y_n) \leq f(x_n + y_n)$$

since  $x'_n$  is in  $X_n$ . Likewise from (c) of Theorem 1 applied to  $W$  we have  $f(x_0 + y_0) \leq f(x_0 + y_n)$ . We consequently conclude from this last fact and (4.13) and (4.11) that

$$(4.14) \quad 0 \leq f(w_0) - f(w_n) \leq f(x_0 + y_n) - f(x'_n + y_n).$$

Now  $f(x'_n + y_n) = f[x_0 + y_n + (x'_n - x_0)]$ , and we obtain from the mean-value theorem applied to  $f[x_0 + y_n + t(x'_n - x_0)]$ , that there exists an  $s$  in the interval  $(0, 1)$  such that the right-hand side of the last inequality in (4.14) is equal to

$$\nabla f[x_0 + y_n + s(x'_n - x_0)](x_0 - x'_n).$$

Now  $\nabla f(x_0 + y_0)(x_0 - x'_n) = 0$  and we obtain, therefore, that

$$(4.15) \quad f(w_0) - f(w_n) \leq \sup_{0 \leq s \leq 1} \|\nabla f(w_0) - \nabla f[w_0 + (y_n - y_0) + s(x'_n - x_0)]\| \|x_0 - x'_n\|.$$

From condition (iii) in the hypothesis of Theorem 4, we next see that there are positive constants  $M(w_0)$  and  $\delta(w_0)$  such that

$$(4.16) \quad \|\nabla f(w_0 + w) - \nabla f(w_0)\| \leq M(w_0)\|w\| \quad \text{for } \|w\| \leq \delta(w_0).$$

Using (4.12), we choose  $N$  such that

$$(4.17) \quad \|x_n - x_0\| + \|y_n - y_0\| \leq \delta(w_0) \quad \text{for } n \geq N.$$

It follows from the definition of  $x'_n$  (given in (b) of Theorem 4) that  $\|x'_n - x_0\| \leq \|x_n - x_0\|$ . We conclude consequently from (4.15), (4.16), and (4.17) that for  $n \geq N$ ,

$$(4.18) \quad f(w_0) - f(w_n) \leq M(w_0)[\|x'_n - x_0\| + \|y_n - y_0\|]\|x_0 - x'_n\|.$$

Equation (4.11) in conjunction with this last inequality gives

$$(4.19) \quad k_2\|x_n - x_0\|^2 + k_1\|y_n - y_0\|^2 \leq 2M(w_0)[\|x'_n - x_0\| + \|y_n - y_0\|]\|x'_n - x_0\|$$

for  $n \geq N$ . Set  $k_3 = \min(k_1, k_2)$ . Then it follows from (4.19) that

$$\|x_n - x_0\| + \|y_n - y_0\| \leq 8M(w_0)k_3^{-1}\|x'_n - x_0\|$$

for  $n \geq N$ . This establishes the second inequality in (b) of Theorem 4.

To establish the first inequality in (b), we note that each  $w$  in  $W$  is uniquely representable in the form  $w = x + y$ , where  $x$  is in  $X$  and  $y$  is in  $Y$ . Consequently, it follows from the closed graph theorem, [6, p. 171], that the mapping  $\Lambda(w) = x$  defines a bounded linear transformation of  $W$  into  $X$ . It is apparent, therefore, that there is a positive constant  $c_4$  such that  $\|x_0 - x_n\| \leq c_4\|(x_0 - x_n) + (y_0 - y_n)\|$ . But from the definition of  $x'_n$ ,  $\|x_0 - x'_n\| \leq \|x_0 - x_n\|$ . Therefore,

$$c_4^{-1}\|x_0 - x'_n\| \leq \|(x_0 - x_n) + (y_0 - y_n)\|.$$

The first inequality in (b) is established (with  $c_1 = c_4^{-1}$ ), and the proof of Theorem 4 is therefore complete.

At this point, we would like to make two remarks concerning facts established in the proof of the above theorem.

*Remark 1.* With  $k_1$  and  $k_2$  the positive constants given in (i) and (ii) of Theorem 4 respectively, and with  $w_n = x_n + y_n$  and  $w_0 = x_0 + y_0$  defined in (a) of Theorem 4 the following fact was established (see 4.11):

$$[k_2\|x_n - x_0\|^2 + k_1\|y_n - y_0\|^2]2^{-1} \leq f(w_0) - f(w_n)$$

for  $n = 1, 2, \dots$ .

*Remark 2.* With  $M(w_0)$  the Lipschitz constant associated with  $\nabla f$  at  $w_0$  given by (iii) of Theorem 4, with  $w_n = x_n + y_n$  and  $w_0 = x_0 + y_0$  given in (a) of Theorem 4, and with  $x'_n$  defined in (b) of Theorem 4 the following fact holds (see (4.18)): there is a positive integer  $N$  such that for  $n \geq N$ ,

$$f(w_0) - f(w_n) \leq M(w_0)[\|x'_n - x_0\| + \|y_n - y_0\|]\|x_0 - x'_n\|.$$

Using these remarks, we next establish the following theorem.

**THEOREM 5.** *Let  $W$  be a reflexive Banach space with  $W = X \oplus Y$ , where  $X$  and  $Y$  are closed subspaces of  $W$ . Suppose that  $f$  is a real-valued function defined on  $W$  which is continuous in the norm topology of  $W$ , and has a  $\mathcal{G}-\mathcal{F}$  derivative at each point  $w$  of  $W$ . Suppose furthermore that conditions (i), (ii), (iii), and (iv) in the hypothesis of Theorem 3 hold. Set  $W_n = X_n \oplus Y_n$ . Then the following prevails:*

- (a) *There is a sequence  $\{w_n\}_1^\infty$  and  $w_0 = x_0 + y_0$  satisfying (a), (b), and (c) of Theorem 3.*
- (b) *For each  $n$ , choose  $x'_n$  in  $X_n$  and  $y'_n$  in  $Y_n$  such that*

$$\|x'_n - x_0\| = \inf \{\|x - x_0\| : x \text{ in } X_n\},$$

$$\|y'_n - y_0\| = \inf \{\|y - y_0\| : y \text{ in } Y_n\}.$$

*Then there are constants  $c_1 > 0$  and  $c_2 > 0$  such that for all  $n$ ,*

$$c_1[\|x_0 - x'_n\| + \|y_0 - y'_n\|] \leq \|w_0 - w_n\| \leq c_2[\|x_0 - x'_n\| + \|y_0 - y'_n\|].$$

Since the hypothesis of Theorem 5 is identical with the hypothesis of Theorem 3, (a) in the conclusion of Theorem 5 is immediate. It remains to prove (b). To do this we proceed as in the proof of Theorem 3 and define  $W_n^*$  and  $W_n^{**}$  as in (3.11). Next, we obtain  $w_n^* = x_n^* + y_n^*$  in  $W_n^*$  and  $w_n^{**} = x_n^{**} + y_n^{**}$  in  $W_n^{**}$  having the properties enumerated in (3.12) and (3.13) respectively. Also, it follows from Remark 1 that

$$(4.20) \quad [k_2\|x_n^* - x_0\|^2 + k_1\|y_n^* - y_0\|^2]2^{-1} \leq f(w_0) - f(w_n^*)$$

and

$$(4.21) \quad [k_2\|x_n^{**} - x_0\|^2 + k_1\|y_n^{**} - y_0\|^2]2^{-1} \leq f(w_n^{**}) - f(w_0),$$

where  $k_1$  and  $k_2$  are the positive constants in (i) and (ii) of Theorem 3.

Next, with  $w_n$  as in (a) of Theorem 3 we see that (3.15) holds. Furthermore, we see from Remark 1 that

$$(4.22) \quad [k_2\|x_n - x_n^{**}\|^2 + k_1\|y_n - y_n^{**}\|^2]2^{-1} \leq f(w_n^{**}) - f(w_n)$$

and

$$(4.23) \quad [k_2\|x_n - x_n^*\|^2 + k_1\|y_n - y_n^*\|^2]2^{-1} \leq f(w_n) - f(w_n^*).$$

From Remark 2, we see that there is a positive integer  $N$  such that

$$(4.24) \quad f(w_0) - f(w_n^*) \leq M(w_0)[\|x'_n - x_0\| + \|y_n^* - y_0\|]\|x_0 - x'_n\|,$$

$$(4.25) \quad f(w_n^{**}) - f(w_0) \leq M(w_0)[\|y'_n - y_0\| + \|x_n^{**} - x_0\|]\|y_0 - y'_n\|$$

for  $n \geq N$ , where  $M(w_0)$  is the Lipschitz constant associated with  $\nabla f$  at  $w_0$ .

Next, we set  $k = \min(k_1, k_2)$ , and obtain from (4.20), (4.24) and the definition of  $x'_n$  that

$$\begin{aligned} \|x_n^* - x_0\|^2 + \|y_n^* - y_0\|^2 &\leq 2k^{-1}M(w_0)[\|x'_n - x_0\| + \|y_n^* - y_0\|]\|x_0 - x'_n\| \\ &\leq 2k^{-1}M(w_0)[\|x_n^* - x_0\| + \|y_n^* - y_0\|]\|x_0 - x'_n\| \end{aligned}$$

for  $n \geq N$ , since  $x_n^*$  is also in  $X_n$ . We conclude in particular from this inequality that

$$(4.26) \quad \|y_n^* - y_0\| \leq 8k^{-1}M(w_0)\|x_0 - x'_n\|$$

for  $n \geq N$ . In a similar manner, we obtain from (4.21) and (4.25) that

$$(4.27) \quad \|x_n^{**} - x_0\| \leq 8k^{-1}M(w_0)\|y_0 - y'_n\|$$

for  $n \geq N$ .

Next, we see that  $\|w_n - w_0\|^2 \leq 4\|w_n^* - w_n\|^2 + 4\|w_n^* - w_0\|^2 \leq 16[\|x_n^* - x_n\|^2 + \|y_n^* - y_n\|^2 + \|x_n^* - x_0\|^2 + \|y_n^* - y_0\|^2]$ . Consequently, we obtain from (4.20) and (4.23) that

$$(4.28) \quad \|w_n - w_0\|^2 \leq 32k^{-1}[f(w_n) - f(w_n^*) + f(w_0) - f(w_n^*)].$$

From (3.15), we have that  $f(w_n) - f(w_n^*) \leq f(w_n^{**}) - f(w_n^*)$ . Using this fact in conjunction with (4.28), we in turn obtain

$$(4.29) \quad \|w_n - w_0\|^2 \leq 32k^{-1}\{f(w_n^{**}) - f(w_0) + 2[f(w_0) - f(w_n^*)]\}.$$

Next, we observe from (4.26) and (4.27) that the right-hand side of the inequalities in (4.24) and (4.25) are majorized by  $M(w_0)[1 + 8k^{-1}M(w_0)]\|x'_n - x_0\|^2$  and  $M(w_0)[1 + 8k^{-1}M(w_0)]\|y'_n - y_0\|^2$ , respectively. This in conjunction with (4.24), (4.25), and (4.29) tells us that

$$(4.30) \quad \|w_n - w_0\|^2 \leq 64k^{-1}M(w_0)[1 + 8k^{-1}M(w_0)][\|x'_n - x_0\|^2 + \|y'_n - y_0\|^2]$$

for  $n \geq N$ . The second inequality in (b) of Theorem 5 follows immediately from (4.30).

To establish the first inequality in (b) of Theorem 5, we note that every  $w$  in  $W$  is uniquely expressible in the form  $w = x + y$ , where  $x$  is in  $X$  and  $y$  is in  $Y$ . Consequently, it follows from the closed graph theorem (as in the proof of Theorem 4) that there is a positive constant  $c_5$  such that  $\|x_0 - x'_n\| \leq \|x_0 - x_n\| \leq c_5\|(x_0 - x_n) + (y_0 - y_n)\|$  for  $n = 1, 2, \dots$ . Likewise, there is a positive constant  $c_6$  such that  $\|y_0 - y'_n\| \leq c_6\|w_0 - w_n\|$  for  $n = 1, 2, \dots$ . Consequently,  $\|x_0 - x'_n\| + \|y_0 - y'_n\| \leq (c_5 + c_6)\|w_0 - w_n\|$  for  $n = 1, 2, \dots$ . The first inequality in (b) of Theorem 5 is therefore established, and the proof of Theorem 5 is complete.

**5. An example.** As an example of a nonlinear functional that satisfies the conditions in the hypothesis of Theorem 5, we use one associated with the generalized nonhomogeneous Dirichlet problem for the biharmonic operator (i.e.,  $\Delta^2 w = g(\xi, w)$  where  $\Delta$  is the usual Laplace operator) under zero boundary conditions (a type of problem which arises in the theory of elasticity, see [4, p. 288]). To be specific, let  $\Omega$  be a bounded open set in real Euclidean space  $R^n$ , and let  $W \equiv W_0^{2,2}(\Omega)$ , where the inner product  $\langle w_1, w_2 \rangle_2$ , in  $W$  is given by

$$\langle w_1, w_2 \rangle_2 = \int_{\Omega} \Delta w_1(\xi) \Delta w_2(\xi) d\xi.$$

(We use the standard notation of Sobolev spaces, e.g., see [1]). We shall designate the

usual inner product in  $L^2(\Omega)$  by  $\langle u_1, u_2 \rangle_0$ ; thus for  $u_1$  and  $u_2$  in  $L^2(\Omega)$ ,

$$\langle u_1, u_2 \rangle_0 = \int_{\Omega} u_1(\xi)u_2(\xi) d\xi.$$

Using Sobolev space theory, it is an easy matter to show that the operator  $T$ , with range in  $W_0^{2,2}(\Omega)$  defined by

$$\langle Tu, \phi \rangle_2 = \langle u, \phi \rangle_0 \quad \text{for } \phi \text{ in } W_0^{2,2}(\Omega),$$

is a linear bounded, self-adjoint, completely continuous operator mapping  $L^2(\Omega)$  into  $L^2(\Omega)$ . It is also an easy matter to show, using the standard Fredholm–Riesz–Schauder theory, that there exists a sequence of positive numbers  $\{\lambda_k\}_1^\infty$  such that  $\lambda_k \leq \lambda_{k+1}$  and  $\lim_{k \rightarrow \infty} \lambda_k = +\infty$  and a corresponding sequence  $\{\phi_k\}_1^\infty$  in  $W_0^{2,2}(\Omega)$  such that  $\Delta^2 \phi_k = \lambda_k \phi_k$  in the distribution sense in  $\Omega$  and such that  $\langle \phi_j, \phi_k \rangle_0 = \delta_j^k$  (the Kronecker-delta). Furthermore,  $\{\phi_k\}_1^\infty$  is complete in  $L^2(\Omega)$  and  $\{\phi_k \lambda_k^{-1/2}\}_1^\infty$  is a complete orthonormal system in  $W_0^{2,2}(\Omega)$ . Also, the following two facts hold for  $w$  in  $W_0^{2,2}(\Omega)$ :

$$(5.1) \quad \langle w, w \rangle_2 = \sum_{k=1}^\infty \lambda_k \langle \phi_k, w \rangle_0^2;$$

$$(5.2) \quad \langle w, w \rangle_0 = \sum_{k=1}^\infty \langle \phi_k, w \rangle_0^2.$$

We shall suppose that  $g(\xi, t)$  is a continuous function defined in  $\bar{\Omega} \times (-\infty, \infty)$ . Furthermore, we shall suppose that there are two positive constants  $\gamma_1$  and  $\gamma_2$  and a positive integer  $N$  such that the following holds:

$$(5.3) \quad \lambda_N < \gamma_1 < \gamma_2 < \lambda_{N+1};$$

$$(5.4) \quad \gamma_1 s \leq g(\xi, t+s) - g(\xi, t) \leq \gamma_2 s \quad \text{for } \xi \text{ in } \bar{\Omega}, t \text{ in } (-\infty, \infty), \text{ and } s \text{ in } (0, \infty).$$

For each  $\xi$  in  $\bar{\Omega}$ , we shall set

$$(5.5) \quad G(\xi, t) = \int_0^t g(\xi, r) dr$$

for  $t$  in  $(-\infty, \infty)$ .

The functional  $f$  on  $W$  that we shall deal with is the following:

$$(5.6) \quad f(w) = 2^{-1} \langle w, w \rangle_2 - \int_{\Omega} G[\xi, w(\xi)] d\xi.$$

We propose to show that  $f$  so defined on  $W$  meets the conditions in the hypothesis of Theorem 5. Also, we propose to show that the unique point  $w_0$  in  $W$  with the property that  $\nabla f(w_0)(w) = 0$  for every  $w$  in  $W$  is a distribution solution of  $\Delta^2 w(\xi) = g[\xi, w(\xi)]$  in  $\Omega$ . (The results that we obtain in this section are to be compared with [5, §7]. The results in this latter reference concerning the Laplace operator can be extended along the lines presented here.)

To accomplish this, we set

$$(5.7) \quad \sup_{\xi \text{ in } \Omega} |g(\xi, 0)| = K_1$$

and observe that  $K_1$  is finite (since  $g$  is in  $C[\bar{\Omega} \times (-\infty, \infty)]$ ). Next, we observe from (5.3), (5.4) and the fact that  $\lambda_N > 0$  that

$$(5.8) \quad |g(\xi, t)| \leq \gamma_2 |t| + K_1$$

for  $\xi$  in  $\bar{\Omega}$  and  $t$  in  $(-\infty, \infty)$ . This inequality in conjunction with (5.5) gives that

$$(5.9) \quad |G(\xi, t)| \leq \gamma_2 t^2 / 2 + K_1 |t| \quad \text{for } \xi \text{ in } \bar{\Omega} \text{ and } t \text{ in } (-\infty, \infty).$$

If  $w$  is in  $W$ , then  $w$  in particular is in  $L^2(\Omega)$  and, therefore, in  $L^1(\Omega)$  since  $\Omega$  is a bounded open set. We consequently conclude from (5.6) and (5.9) that  $f(w)$  is indeed well-defined for  $w$  in  $W$ .

To show that  $f$  is continuous in the norm topology on  $W$ , we observe from (5.4) and (5.5) that

$$(5.10) \quad |G(\xi, t_2) - G(\xi, t_1)| \leq 2^{-1} \gamma_2 |t_2 - t_1|^2 + |g(\xi, t_1)| |t_2 - t_1|$$

for  $\xi$  in  $\bar{\Omega}$  and  $t_1, t_2$  in  $(-\infty, \infty)$ . We consequently obtain from this fact in conjunction with (5.8) that

$$\int_{\Omega} |G[\xi, w_1(\xi) + w(\xi)] - G[\xi, w_1(\xi)]| d\xi$$

is dominated by

$$(5.11) \quad \int_{\Omega} [\gamma_2 |w(\xi)|^2 + \gamma_2 |w_1(\xi)| |w(\xi)| + K_1 |w(\xi)|] d\xi.$$

Now, as is well-known, there is a positive constant  $K_2$  such that

$$(5.12) \quad \langle w, w \rangle_0 \leq K_2 \langle w, w \rangle_2 \equiv K_2 \|w\|^2.$$

Consequently, the integral in (5.11) goes to zero as  $\|w\| \rightarrow 0$ , and we obtain from (5.6) that

$$\lim_{\|w\| \rightarrow 0} f(w_1 + w) = f(w_1)$$

for  $w_1$  in  $W$ .  $f$  is indeed continuous on  $W$ .

Next, we compute  $\nabla f(w_1)(w_2)$  for  $w_1$  and  $w_2$  in  $W$ . From (5.6), we see that

$$(5.13) \quad \begin{aligned} f(w_1 + tw_2) - f(w_1) &= t \langle w_1, w_2 \rangle_2 + t^2 \langle w_2, w_2 \rangle_2 2^{-1} \\ &\quad - \int_{\Omega} \{G[\xi, w_1 + tw_2] - G[\xi, w_1]\} d\xi. \end{aligned}$$

From (5.10), we see the absolute value of the integrand in this last expression is dominated by  $\gamma_2 t^2 |w_2|^2 + |t| |g(\xi, w_1)| |w_2|$ . We consequently see from (5.5), (5.8), (5.13) and the Lebesgue dominated convergence theorem that

$$(5.14) \quad \begin{aligned} \nabla f(w_1)(w_2) &= \lim_{t \rightarrow 0} [f(w_1 + tw_2) - f(w_1)] t^{-1} \\ &= \langle w_1, w_2 \rangle_2 - \int_{\Omega} g(\xi, w_1) w_2 d\xi \end{aligned}$$

for  $w_1$  and  $w_2$  in  $W$ .

To show that  $\nabla f$  is locally pointwise Lipschitz in  $W$ , we use (5.14) and observe that

$$(5.15) \quad \nabla f(w + w_1)(w_2) - \nabla f(w)(w_2) = \langle w_1, w_2 \rangle_2 - \int_{\Omega} [g(\xi, w + w_1) - g(\xi, w)] w_2 d\xi$$

for  $w_1, w_2$  and  $w$  in  $W$ . We consequently obtain from (5.4) and Schwarz's inequality that



the absolute value of the left-hand side of (5.15) is majorized by

$$\|w_1\| \|w_2\| + \gamma_2 \langle |w_1|, |w_2 \rangle_0.$$

We conclude from this fact and (5.12) that

$$\|\nabla f(w + w_1) - \nabla f(w)\| \leq (1 + \gamma_2 K_2^{1/2}) \|w_1\|$$

for  $w_1$  and  $w$  in  $W$ . Condition (iii) in the hypothesis of Theorem 5 (or 3) is therefore established.

Next, we define  $X = \{x: x = \sum_{j=1}^N c_j \phi_j\}$ ,  $Y_n = \{y: y = \sum_{j=1}^n c_j \phi_{j+N}\}$ , and  $Y = (\cup_{n=1}^\infty Y_n)^-$ , i.e.,  $Y$  is the closure in  $W$  of  $\cup_{n=1}^\infty Y_n$ . Since  $\{\phi_j \lambda_j^{-1/2}\}_1^\infty$  is a complete orthonormal system in  $W$ , it follows that  $W = X \oplus Y$ . Furthermore with  $X_n = X$  for  $n = 1, 2, \dots$ , it is clear that condition (iv) in the hypothesis of Theorem 5 (or 3) is met. It remains to show that conditions (i) and (ii) are also met.

To accomplish this, we first observe from (5.5) that

$$G(\xi, t+s) + G(\xi, t-s) - 2G(\xi, t) = \int_0^s [g(\xi, t+r) - g(\xi, t-r)] dr$$

for  $t$  in  $(-\infty, \infty)$  and  $s$  in  $(0, \infty)$ . Consequently, we see from (5.4) that

$$(5.16) \quad \gamma_1 s^2 \leq G(\xi, t+s) + G(\xi, t-s) - 2G(\xi, t) \leq \gamma_2 s^2$$

for  $t$  in  $(-\infty, \infty)$  and  $s$  in  $(0, \infty)$ .

Next, we use (5.6) and observe from the second inequality in (5.16) that

$$\begin{aligned} f(w + tw_1) + f(w - tw_1) - 2f(w) - t^2 \langle w_1, w_1 \rangle_2 \\ = - \int_\Omega [G(\xi, w + tw_1) + G(\xi, w - tw_1) - 2G(\xi, w)] d\xi \\ \geq -\gamma_2 t^2 \int_\Omega w_1^2 d\xi. \end{aligned}$$

We consequently obtain from (2.1) that

$$(5.17) \quad \underline{D}^2 f(w, w_1) \geq \langle w_1, w_1 \rangle_2 - \gamma_2 \langle w_1, w_1 \rangle_0$$

for  $w$  and  $w_1$  in  $W$ . In a similar manner, we obtain from (5.6) and the first inequality in (5.16) that

$$(5.18) \quad \bar{D}^2 f(w, w_1) \leq \langle w_1, w_1 \rangle_2 - \gamma_1 \langle w_1, w_1 \rangle_0$$

for  $w$  and  $w_1$  in  $W$ .

From the definition of  $X$  and from (5.1) and (5.2) we see that for  $x$  in  $X$ ,  $\langle x, x \rangle_0 = \sum_{j=1}^N \langle x, \phi_j \rangle_0^2$  and  $\langle x, x \rangle_2 = \sum_{j=1}^N \lambda_j \langle x, \phi_j \rangle_0^2$ . Consequently, we obtain from (5.18) [since  $\underline{D}^2 f(w, w_1) \leq \bar{D}^2 f(w, w_1)$ ] that  $\underline{D}^2 f(w, x) \leq \sum_{j=1}^N (\lambda_j - \gamma_1) \langle x, \phi_j \rangle_0^2$ . But then

$$(5.19) \quad -\underline{D}^2 f(w, x) \geq \sum_{j=1}^N (\gamma_1 \lambda_j^{-1} - 1) \lambda_j \langle x, \phi_j \rangle_0^2.$$

Setting  $k_2 = (\gamma_1 \lambda_N^{-1} - 1)$ , we see from (5.3) that  $k_2 > 0$  and from the monotonicity of  $\{\lambda_j\}_1^\infty$  in conjunction with (5.19) that

$$-\underline{D}^2 f(w, x) \geq k_2 \sum_{j=1}^N \lambda_j \langle x, \phi_j \rangle_0^2.$$

But the sum on the right-hand side of the above inequality is equal to  $\|x\|^2$ . Therefore,

$-D^2f(w, x) \geq k_2\|x\|^2$ , i.e.,  $D^2f(w, x) \leq -k_2\|x\|^2$ , and condition (ii) in the hypothesis of Theorem 5 (or 3) is established.

To show that condition (i) holds, we observe from (5.1), (5.2), and the definition of  $Y$  that

$$(5.20) \quad \begin{aligned} \langle y, y \rangle_0 &= \sum_{j=1}^{\infty} \langle y, \phi_{N+j} \rangle_0^2, \\ \langle y, y \rangle_2 &= \sum_{j=1}^{\infty} \lambda_{N+j} \langle y, \phi_{N+j} \rangle_0^2 \end{aligned}$$

for  $y$  in  $Y$ . We then obtain from (5.17) [since  $\bar{D}^2f(w, y) \geq \bar{D}^2f(w, y)$ ] that for  $w$  in  $W$  and  $y$  in  $Y$ ,

$$\begin{aligned} \bar{D}^2f(w, y) &\geq \langle y, y \rangle_2 - \gamma_2 \langle y, y \rangle_0 \\ &\geq \sum_{j=1}^{\infty} [\lambda_{N+j} - \gamma_2] \langle y, \phi_{N+j} \rangle_0^2 \\ &\geq \sum_{j=1}^{\infty} [1 - \lambda_{N+j}^{-1} \gamma_2] \lambda_{N+j} \langle y, \phi_{N+j} \rangle_0^2. \end{aligned}$$

Setting  $k_1 = (1 - \lambda_{N+1}^{-1} \gamma_2)$ , we see from (5.3) that  $k_1 > 0$  and from the last inequality established that  $\bar{D}^2f(w, y) \geq k_1 \sum_{j=1}^{\infty} \lambda_{N+j} \langle y, \phi_{N+j} \rangle_0^2$ . But then from (5.20) we have that  $\bar{D}^2f(w, y) \geq k_1 \|y\|^2$ , and condition (i) in the hypothesis of Theorem 5 (or 3) is established. Therefore, all the conditions in the hypothesis of Theorem 5 are established and we conclude that there exists a unique  $w_0$  in  $W_0$  such that  $\nabla f(w_0) = 0$ . Also this  $w_0$  can be found via the Ritz-Galerkin approximations in Theorem 5.

Let  $\phi$  be a function in  $C_0^\infty(\Omega)$ . Then  $\phi$  is in  $W$ , and we have from (5.14) (since  $\nabla f(w_0)(\phi) = 0$ ) that

$$(5.21) \quad \int_{\Omega} \Delta w_0(\xi) \Delta \phi(\xi) \, d\xi = \int_{\Omega} g[\xi, w_0(\xi)] \phi(\xi) \, d\xi.$$

But  $\int_{\Omega} \Delta w_0(\xi) \Delta \phi(\xi) \, d\xi = \int_{\Omega} w_0(\xi) \Delta^2 \phi(\xi) \, d\xi$ . This equality in conjunction with (5.21) tells that  $\Delta^2 w_0$  is indeed equal to  $g(\xi, w_0)$  in the distribution sense in  $\Omega$ , and consequently, all assertions made in this section about the functional  $f(w)$  in (5.6) are established.

REFERENCES

[1] R. A. ADAMS, *Sobolev spaces*, Academic Press, New York, 1975.  
 [2] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42-47.  
 [3] A. GOLDSTEIN, *Constructive real analysis*, Harper and Row, New York, 1967.  
 [4] L. V. KANTOROVICH AND V. I. KRYLOV, *Approximate methods of higher analysis*, Interscience, New York, 1958.  
 [5] A. C. LAZER, E. M. LANDESMAN AND D. R. MEYERS, *On saddle-point problems in the calculus of variations, the Ritz algorithm, and monotone convergence*, J. Math. Anal. Appl., 52 (1975), pp. 594-614.  
 [6] H. L. ROYDEN, *Real analysis*, Macmillan, New York, 1965.  
 [7] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.  
 [8] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1968.  
 [9] A. ZYGMUND, *Trigonometric Series*, Vol. 1, Cambridge University Press, Cambridge, England, 1959.

## THE GENERALIZED ABEL EQUATIONS FOR SCHWARTZ DISTRIBUTIONS\*

MARION ORTON†

**Abstract.** The class of equations  $aI^\alpha f + bK^\alpha f = g$  is analyzed for data  $g$  and solutions  $f$  belonging to the space  $\mathcal{D}'(\mathcal{R})$  of Schwartz distributions.  $f$  has compact support contained in  $[-1, 1]$ , and  $g$  is known only on  $(-1, 1)$ .  $I^\alpha$  and  $K^\alpha$  are the usual operators of fractional integration defined as convolutions. The equation is to be satisfied as an identity between distributions on the open interval  $(-1, 1)$ . The coefficients  $a$  and  $b$  are infinitely differentiable functions on  $(-1, 1)$  subject to certain growth conditions at the endpoints.

It is shown that in this setting the equation  $aI^\alpha f + bK^\alpha f = g$  is equivalent to a boundary value problem for functions analytic off the real axis. The class of analytic functions furnishing the solutions is characterized in terms of its growth rate at infinity and its limiting behavior at the real axis. Solutions to the generalized Abel equations are found explicitly for arbitrary distribution data and all complex values of  $\alpha$ .

**1. Introduction.** Various generalizations of Abel's integral equation have been studied in the past. Among them are the integral equations

$$(1.1) \quad S^\alpha f = g$$

involving the operator

$$(1.2) \quad S^\alpha f = a(x) \frac{1}{\Gamma(\alpha)} \int_c^x (x-t)^{\alpha-1} f(t) dt + b(x) \frac{1}{\Gamma(\alpha)} \int_x^d (t-x)^{\alpha-1} f(t) dt$$

on an interval  $(c, d)$  with  $-\infty \leq c < d \leq \infty$ . The class of integral operators  $S^\alpha$  defined by (1.2) includes the operators of fractional integration

$$(1.3) \quad I^\alpha f = \frac{1}{\Gamma(\alpha)} \int_c^x (x-t)^{\alpha-1} f(t) dt$$

$$(1.4) \quad K^\alpha f = \frac{1}{\Gamma(\alpha)} \int_x^d (t-x)^{\alpha-1} f(t) dt;$$

the Riesz potentials

$$(1.5) \quad R^\alpha f = \frac{1}{2} \sec\left(\frac{1}{2}\alpha\pi\right) \frac{1}{\Gamma(\alpha)} \int_c^d |t-x|^{\alpha-1} f(t) dt;$$

and the generalized Hilbert transforms

$$(1.6) \quad H^\alpha f = \frac{1}{2} \csc\left(\frac{1}{2}\alpha\pi\right) \frac{1}{\Gamma(\alpha)} \int_c^d |t-x|^{\alpha-1} \operatorname{sgn}(x-t) f(t) dt.$$

A further generalization of Abel's equation which includes (1.1) is defined by taking

$$(1.7) \quad T^\alpha f = \frac{1}{\Gamma(\alpha)} \int_c^d k(x, t) |t-x|^{\alpha-1} f(t) dt$$

in place of  $S^\alpha f$ , where  $k(x, t)$  may be discontinuous along the diagonal  $x = t$ .

\* Received by the editors December 6, 1978.

† Department of Mathematics, University of California, Irvine, California. Now at Schlumberger-Doll Research, Ridgefield, Connecticut 06877.

If neither  $a$  nor  $b$  is identically zero, then solutions of (1.1) can often be found using Cauchy integrals in terms of which (1.1) is transformed into a Hilbert boundary value problem for analytic functions. This method was first employed by Carleman [3], who inverted the operator  $R^\alpha$  on  $[0, 1]$  in this manner. It establishes a connection between operators of fractional integration and singular integral operators which has been used most successfully in the analysis of the operators  $S^\alpha$ ,  $R^\alpha$ ,  $H^\alpha$  and  $T^\alpha$  on  $L^p$ -spaces, especially in determining their ranges and indices. For  $L^p$ -spaces over bounded intervals results for  $S^\alpha$  are found in [26], [36]–[38]; for  $S^\alpha + A$ ,  $A$  compact, in [4], [27]. The operator  $S^\alpha$  on  $L^p(-\infty, \infty)$  is discussed in [29], [30];  $T^\alpha$  and its range are examined in [24], [25], [31], [32]. The properties of the particular operators  $R^\alpha$  and  $H^\alpha$  on  $L^p(-\infty, \infty)$  are discussed in [3], [9], [10], [15], [19], [20], [34]. The  $L^p$ -theory for the operators of fractional integration,  $I^\alpha$  and  $K^\alpha$ , is discussed in [1] and [11], for example.

In this article we consider equation (1.1) over a bounded interval with data in  $\mathcal{D}'(\mathcal{R})$ , the space of Schwartz distributions on the real line. In particular, and without loss of generality, we take  $(-1, 1)$  for the interval  $(c, d)$ . We take the coefficients  $a$  and  $b$  to be infinitely differentiable on  $(-1, 1)$ , and interpret the integrals  $I^\alpha f$  and  $K^\alpha f$  as convolutions of distributions. Equation (1.1) can then be viewed as an equality between the restrictions of distributions to an open set: given data  $g \in \mathcal{D}'(\mathcal{R})$  we seek  $f \in \mathcal{D}'(\mathcal{R})$  with support in  $[-1, 1]$  such that  $aI^\alpha f + bK^\alpha f = g$  on  $(-1, 1)$ .

It will be shown that equation (1.1) as formulated for distributions is equivalent to a Hilbert boundary value problem for analytic functions. We give a complete proof of this fact which follows from a Paley–Wiener type theorem for analytic representations of fractional integrals of compactly supported distributions. (No analogous result for fractional integrals of  $L^p$ -functions appears to be known).

Hilbert boundary value problems which involve analytic functions with boundary values in  $\mathcal{D}'(\mathcal{R})$  were treated in [22] in a general setting. We apply those results to prove the existence of distribution solutions to (1.1) for a large class of coefficients  $a$  and  $b$ , and arbitrary data  $g \in \mathcal{D}'(\mathcal{R})$ .

In preparation for our main result we discuss the extensions of the operators  $I^\alpha$  and  $K^\alpha$  from  $L^p(-1, 1)$  to the space of Schwartz distributions with support in  $[-1, 1]$ . (Extensions of the operators  $I^\alpha$  and  $K^\alpha$  from  $L^p(\mathcal{R})$  to subspaces of  $\mathcal{D}'(\mathcal{R})$  containing  $L^p(\mathcal{R})$  are treated in [5], [6], [17]). Following that we consider the analytic representations of fractional integrals.

**2. Formulation of the problem and statement of the results.** Let  $\mathcal{D}(\mathcal{R})$  denote the space of infinitely differentiable complex-valued functions with compact support defined on the real line  $\mathcal{R}$ . Its dual  $\mathcal{D}'(\mathcal{R})$  is the space of Schwartz distributions.  $\mathcal{D}(\mathcal{R})$  and  $\mathcal{D}'(\mathcal{R})$  carry the usual topologies [33]. For  $f \in \mathcal{D}'(\mathcal{R})$  and  $\varphi \in \mathcal{D}(\mathcal{R})$  we write  $\langle f, \varphi \rangle$  for the value of  $f$  at  $\varphi$ . Two distributions  $f$  and  $g$  agree on an open set  $\Omega \subset \mathcal{R}$ , if  $\langle f, \varphi \rangle = \langle g, \varphi \rangle$  for all  $\varphi \in \mathcal{D}(\mathcal{R})$  with support in  $\Omega$ . For the details of the theory we refer the reader to [8], [33].

Let  $\mathcal{D}'_l$  and  $\mathcal{D}'_r$  denote the subsets of distributions in  $\mathcal{D}'(\mathcal{R})$  whose supports are bounded on the left, and on the right, respectively.  $\mathcal{D}'_l$  and  $\mathcal{D}'_r$  are algebras under convolution, with the Dirac delta distribution  $\delta$  as the identity element and no divisors of zeros [33]. For  $f \in \mathcal{D}'_l$  and  $g \in \mathcal{D}'_r$ , the convolutions

$$(2.1) \quad I^\alpha f = \frac{x_+^{\alpha-1}}{\Gamma(\alpha)} * f, \quad K^\alpha g = \frac{x_-^{\alpha-1}}{\Gamma(\alpha)} * g$$

define the operators  $I^\alpha$  and  $K^\alpha$  of fractional integration. Here  $x_+^{\alpha-1}/\Gamma(\alpha)$  and

$x_-^{\alpha-1}/\Gamma(\alpha)$  are the distributions defined, for  $\text{Re } \alpha > 0$ , by the locally integrable functions

$$(2.2) \quad \frac{x_+^{\alpha-1}}{\Gamma(\alpha)} = \begin{cases} \frac{x^{\alpha-1}}{\Gamma(\alpha)}, & x > 0, \\ 0, & x < 0, \end{cases}$$

$$\frac{x_-^{\alpha-1}}{\Gamma(\alpha)} = \begin{cases} 0, & x > 0, \\ \frac{|x|^{\alpha-1}}{\Gamma(\alpha)}, & x < 0. \end{cases}$$

For  $\text{Re } \alpha \leq 0$ ,  $x_+^{\alpha-1}/\Gamma(\alpha)$  and  $x_-^{\alpha-1}/\Gamma(\alpha)$  are defined by analytic continuation in the parameter  $\alpha$ . This is possible, since they are entire functions of  $\alpha$  [8]. For  $n = 0, 1, 2, \dots$  analytic continuation yields

$$(2.3) \quad \left. \frac{x_+^{\alpha-1}}{\Gamma(\alpha)} \right|_{\alpha=-n} = \delta^{(n)}(x), \quad \left. \frac{x_-^{\alpha-1}}{\Gamma(\alpha)} \right|_{\alpha=-n} = (-1)^n \delta^{(n)}(x).$$

For  $\alpha, \beta \in \mathbb{C}$  one obtains

$$(2.4) \quad I^\alpha I^\beta f = I^{\alpha+\beta} f, \quad K^\alpha K^\beta g = K^{\alpha+\beta} g$$

so that  $I^\alpha$  and  $K^\alpha$  are invertible in  $\mathcal{D}'_i$  and  $\mathcal{D}'_r$ , with inverses  $I^{-\alpha}$  and  $K^{-\alpha}$ , respectively.

The domain on which the operators  $I^\alpha$  and  $K^\alpha$  will be studied here is the subspace of distributions which are zero on  $\mathcal{R} \setminus [-1, 1]$ . We will denote that space by  $\mathcal{D}'([-1, 1])$ . On  $\mathcal{D}'([-1, 1])$   $I^\alpha$  and  $K^\alpha$  are extensions of the integral operators given by

$$(2.5) \quad I^\alpha \varphi = \frac{1}{\Gamma(\alpha)} \int_{-1}^x (x-t)^{\alpha-1} \varphi(t) dt, \quad K^\alpha \varphi = \frac{1}{\Gamma(\alpha)} \int_x^1 (t-x)^{\alpha-1} \varphi(t) dt$$

for  $\varphi \in \mathcal{D}(\mathcal{R})$  and  $\text{Re } \alpha > 0$ .

Traditionally, equations (2.5) have been used as the defining equations for integral operators on  $L^p(-1, 1)$ . In particular, if  $0 < \alpha < 1$ , and  $1 \leq p \leq 1/\alpha$  then the integrals (2.5) define bounded operators mapping  $L^p(-1, 1)$  into itself. Every function  $f \in L^p(-1, 1)$ ,  $p \geq 1$ , is associated with a unique distribution in  $\mathcal{D}'([-1, 1])$  by means of the assignment

$$\langle f, \varphi \rangle = \int_{-1}^1 f(x)\varphi(x) dx$$

for  $\varphi \in \mathcal{D}(\mathcal{R})$ . Identifying  $f \in L^p(-1, 1)$  with the associated distribution in  $\mathcal{D}'([-1, 1])$  the convolutions  $I^\alpha f$  and  $K^\alpha f$  given by (2.1) are defined and yield elements in  $\mathcal{D}'(\mathcal{R})$ . The restrictions of  $I^\alpha f$  and  $K^\alpha f$  to the interval  $(-1, 1)$  belong to  $\mathcal{D}'((-1, 1))$ , which is the dual of  $\mathcal{D}((-1, 1))$ , the subset of  $\mathcal{D}(\mathcal{R})$  of functions with support in  $(-1, 1)$ . On the other hand, the integrals (2.5) belong to  $L^p(-1, 1)$  and thus define, in the obvious way, distributions in  $\mathcal{D}'((-1, 1))$  which agree with the restrictions of  $I^\alpha f$  and  $K^\alpha f$ . We conclude, that the maps from  $L^p(-1, 1)$  into  $L^p(-1, 1)$  defined by the integral operators (2.5) lift to maps from  $\mathcal{D}'([-1, 1])$  into  $\mathcal{D}'((-1, 1))$ . The latter are given by the convolution operators  $I^\alpha, K^\alpha : \mathcal{D}'([-1, 1]) \rightarrow \mathcal{D}'(\mathcal{R})$ , followed by the restriction map  $\iota : \mathcal{D}'(\mathcal{R}) \rightarrow \mathcal{D}'((-1, 1))$ .

In light of these remarks, it is to be expected that many identities in  $L^p(-1, 1)$  involving the integrals  $I^\alpha f, K^\alpha f$  for  $f \in L^p(-1, 1)$ , are special cases of more general statements for the restrictions to  $(-1, 1)$  of the convolutions  $I^\alpha f, K^\alpha f$  for  $f \in \mathcal{D}'([-1, 1])$ .

That this is indeed so is demonstrated in [23], where we discuss a number of important identities relating fractional integrals and Hilbert transforms of distributions. Therefore, we conclude that the generalized Abel equation (1.1) is properly reformulated as a problem in  $\mathcal{D}'([-1, 1])$  when stated as follows:

*Problem I.* Given data  $g \in \mathcal{D}'(\mathcal{R})$  and coefficients  $a$  and  $b$  which are infinitely differentiable on the interval  $(-1, 1)$ , find  $f \in \mathcal{D}'([-1, 1])$  which on  $(-1, 1)$  satisfies

$$(2.6) \quad aI^\alpha f + bK^\alpha f = g.$$

*Remarks.* 1. It is easily seen that every solution of (2.6) in the  $L^p$ -sense is a solution of Problem I in the distribution sense.

2. Problem I could be generalized by considering data  $g \in \mathcal{D}'((-1, 1))$ . This class of distributions contains elements which cannot be extended to distributions on the line and thus our methods below would have to be modified.

3. There is no loss of generality in assuming, as we shall do below, that  $g = 0$  on  $|x| > 1$ . Given any  $g \in \mathcal{D}'(\mathcal{R})$ , it is always possible to find  $g_0 \in \mathcal{D}'(\mathcal{R})$  with  $g_0 = g$  on  $|x| < 1$  and  $g_0 = 0$  on  $|x| > 1$ . To see this, recall that we can find an integer  $n$  and a continuous function  $G$  with support on an arbitrary neighborhood of  $[-1, 1]$  such that  $g = d^n G/dx^n$  on  $(-1, 1)$ . Let  $G_0$  be the locally integrable function  $\theta(1-x)\theta(1+x)G(x)$  where  $\theta$  is Heaviside's unit step function. Define  $g_0 = d^n G_0/dx^n$ . Then  $g_0$  has the desired properties.

The connection between Problem I and a boundary value problem for analytic functions is established in the  $L^p$ -case by the use of Cauchy integrals. In the case of distributions, Cauchy integrals are replaced by analytic representations. With every distribution  $f \in \mathcal{D}'(\mathcal{R})$  there is associated such an analytic representation; that is there exists a function  $F(z)$ , defined and analytic for  $\text{Im } z \neq 0$  for which the limits

$$(2.7) \quad F(x + i0) = \lim_{\varepsilon \downarrow 0} F(x + i\varepsilon), \quad F(x - i0) = \lim_{\varepsilon \downarrow 0} F(x - i\varepsilon)$$

exist in the sense of convergence in  $\mathcal{D}'(\mathcal{R})$  and satisfy

$$(2.8) \quad f(x) = F(x + i0) - F(x - i0).$$

Thus for  $\varphi \in \mathcal{D}(\mathcal{R})$ ,  $\langle f, \varphi \rangle$  can be represented as

$$(2.9) \quad \langle f, \varphi \rangle = \lim_{\varepsilon \downarrow 0} \int_{-\infty}^{\infty} [F(x + i\varepsilon) - F(x - i\varepsilon)] \varphi(x) dx.$$

If  $f$  has compact support, then  $f$  extends to a continuous linear functional on the space of  $C^\infty$ -functions on  $\mathcal{R}$  and we may take for its analytic representation the Cauchy "integral"

$$(2.10) \quad F(z) = \frac{1}{2\pi i} \langle f(t), (t - z)^{-1} \rangle, \quad \text{Im } z \neq 0.$$

The theory of analytic representations is developed in [2], [16], [35], among others.

It will be shown below that for distributions  $f$  of compact support there exists an analytic representation  $F_\alpha(z)$  of  $I^\alpha f$ ,  $\alpha \neq 1, 2, \dots$ , with boundary values  $F_\alpha(x + i0)$  and  $F_\alpha(x - i0)$  in  $\mathcal{D}'(\mathcal{R})$  such that

$$(2.11) \quad \begin{aligned} F_\alpha(x + i0) - F_\alpha(x - i0) &= I^\alpha f, \\ e^{i\alpha\pi} F_\alpha(x + i0) - e^{-i\alpha\pi} F_\alpha(x - i0) &= K^\alpha f, \\ |F_\alpha(z)z^{-\alpha}| &= O(|z|^{-1}) \quad \text{as } |z| \rightarrow \infty. \end{aligned}$$

We now recall that for a distribution  $f \in \mathcal{E}'(\mathcal{R})$  and  $g \in \mathcal{D}'(\mathcal{R})$

$$(2.12) \quad \text{supp } (f * g) \subset \text{supp } f + \text{supp } g.$$

Thus the supports of  $I^\alpha f$  and  $K^\alpha f$  are contained in the intervals  $[-1, \infty)$  and  $(-\infty, 1]$ , respectively. Thus we derive from Problem I the following Hilbert boundary value problem for  $F_\alpha(z)$ ,  $\alpha \neq 1, 2, \dots$ :

*Problem II.* Given data  $g \in \mathcal{D}'(\mathcal{R})$  and coefficients  $a$  and  $b$  which are infinitely differentiable on  $(-1, 1)$ , find the function  $F_\alpha(z)$  analytic for  $\text{Im } z \neq 0$  such that

(i) the limits  $F_\alpha(x + i0)$  and  $F_\alpha(x - i0)$  exist in the sense of convergence in  $\mathcal{D}'(\mathcal{R})$  and satisfy

$$\begin{aligned} F_\alpha(x + i0) - F_\alpha(x - i0) &= 0 \quad \text{on } (-\infty, -1), \\ (a + b e^{i\alpha\pi})F_\alpha(x + i0) - (a + b e^{-i\alpha\pi})F_\alpha(x - i0) &= g \quad \text{on } (-1, 1), \\ e^{i\alpha\pi}F_\alpha(x + i0) - e^{-i\alpha\pi}F_\alpha(x - i0) &= 0 \quad \text{on } (1, \infty); \end{aligned}$$

$$(ii) \quad |F_\alpha(z)z^{-\alpha}| = O(|z|^{-1}) \quad \text{as } |z| \rightarrow \infty.$$

*Remark.* Choosing  $0 < \arg z < 2\pi$ ,  $F_\alpha(z)z^{-\alpha}$  will be shown to be analytic for  $|z| > 1$ . Thus condition (ii) is well-posed.

Our aim is to prove the following:

**THEOREM 2.1.** *Problems I and II are equivalent. For  $\alpha \neq 1, 2, \dots$ , a one-to-one correspondence between their solutions is established by the equations*

$$(2.13) \quad \begin{aligned} f(x) &= I^{-\alpha}[F_\alpha(x + i0) - F_\alpha(x - i0)], \\ F_\alpha(z) &= \frac{\Gamma(1 - \alpha)}{2\pi i} \langle f(t), (t - z)^{\alpha-1} \rangle. \end{aligned}$$

Once Theorem 2.1 is established we find all solutions of the generalized Abel equation (1.1)—under certain restrictions on the growth of  $a$  and  $b$  at the endpoints of the interval  $(-1, 1)$ —in terms of the solutions of the Hilbert boundary value problem. Thus the question of existence and construction of solutions to Problem I is answered for a large class of coefficients  $a$  and  $b$ .

**3. Analytic representations of  $I^\alpha f$  and  $K^\alpha f$ .** For  $f \in L^p(-1, 1)$  with  $1 \leq p \leq 1/\alpha$ ,  $0 < \alpha < 1$ , let  $F_\alpha(z)$  denote the usual Cauchy integral of the convolution  $I^\alpha f \in L^p(\mathcal{R})$ . The limits  $F_\alpha(x + i0)$  and  $F_\alpha(x - i0)$  exist in  $L^p(\mathcal{R})$  and satisfy the Plemelj relations

$$(3.1) \quad \begin{aligned} F_\alpha(x + i0) - F_\alpha(x - i0) &= I^\alpha f, \\ i[F_\alpha(x + i0) + F_\alpha(x - i0)] &= H(I^\alpha f) \end{aligned}$$

where  $H$  denotes the Hilbert transform, that is convolution with the Cauchy principal value distribution  $-(1/\pi) \text{Pv } (1/x)$ . The following identities in  $L^p(\mathcal{R})$  are known to hold for  $f \in L^p(\mathcal{R})$  (and thus for  $f \in L^p(-1, 1) \cap \mathcal{D}'([-1, 1])$ )  $1 \leq p \leq 1/\alpha$ ,  $0 < \alpha < 1$ :

$$(3.2) \quad \begin{aligned} I^\alpha f &= \cos(\alpha\pi)K^\alpha f - \sin(\alpha\pi)H(K^\alpha f), \\ K^\alpha f &= \cos(\alpha\pi)I^\alpha f + \sin(\alpha\pi)H(K^\alpha f). \end{aligned}$$

(For these and other identities relating operators of fractional integration and singular integral operators see [12]–[14], [28], [35].) Equations (3.2) establish the connection between the generalized Abel equation and a singular integral equation which can be treated as a boundary value problem for the analytic function  $F_\alpha(z)$  by the methods of [7] or [18].

In this section we define and study functions  $F_\alpha(z)$  for complex values  $\alpha \neq 1, 2, \dots$  which are analytic representations of the distributions  $I^\alpha f$ , for  $f \in \mathcal{D}'([-1, 1])$ .  $F_\alpha(z)$  agrees with the usual Cauchy integral of  $I^\alpha f$  if  $I^\alpha f \in L^p(\mathcal{R})$ ,  $p \geq 1$ . The boundary values  $F_\alpha(x + i0)$  and  $F_\alpha(x - i0)$ , defined as limits in  $\mathcal{D}'(\mathcal{R})$ , are linear combinations of  $I^\alpha f$  and  $K^\alpha f$ , as are obtained by combining (3.1) and (3.2). Thereby the generalized Abel equation in its new setting, Problem I, again gives rise to a Hilbert boundary value problem for  $F_\alpha(z)$ .

Let  $f \in \mathcal{D}'([-1, 1])$ . For  $\text{Im } z \neq 0$  and  $t \in \mathcal{R}$  define  $(t - z)^{\alpha-1}$  by requiring that  $-\pi < \arg(t - z) < 0$  for  $\text{Im } z > 0$  and  $0 < \arg(t - z) < \pi$  for  $\text{Im } z < 0$ . For fixed  $z$  with  $\text{Im } z \neq 0$ ,  $(t - z)^{\alpha-1}$  is an infinitely differentiable function of  $t$ , and thus belongs to the space  $\mathcal{E}(\mathcal{R})$  of  $C^\infty$ -functions on  $\mathcal{R}$ . Since  $f \in \mathcal{D}'(\mathcal{R})$  and has compact support,  $f$  can be viewed as an element in  $\mathcal{E}'(\mathcal{R})$ . Thus, for  $\alpha \neq 1, 2, 3, \dots$  and  $\text{Im } z \neq 0$

$$(3.3) \quad F_\alpha(z) = \frac{\Gamma(1-\alpha)}{2\pi i} \langle f(t), (t-z)^{\alpha-1} \rangle$$

is well-defined. It is a simple exercise to show that  $F_\alpha(z)$  extends to an analytic function in the complex plane minus the real interval  $[a, \infty)$  where  $a = \inf \{x \in \mathcal{R} \wedge x \in \text{supp } f\}$ . Let  $y = \text{Im } z$ . Viewed as a distribution in the variable  $x$ ,  $F_\alpha(x + iy)$  is the convolution of a distribution  $f$  of compact support and the infinitely differentiable function  $(-x - iy)^{\alpha-1}$ . As  $y \rightarrow 0$  for  $y > 0$ , and  $y < 0$  respectively, the corresponding limits for  $F_\alpha(x + iy)$  exist in the sense of convergence in  $\mathcal{D}'(\mathcal{R})$ . Making use of the continuity of convolution with a distribution of compact support these limits are computed as

$$(3.4a) \quad \begin{aligned} F_\alpha(x + i0) &= \frac{\Gamma(1-\alpha)\Gamma(\alpha)}{2\pi i} [K^\alpha f + e^{-i(\alpha-1)\pi} I^\alpha f] \\ F_\alpha(x - i0) &= \frac{\Gamma(1-\alpha)\Gamma(\alpha)}{2\pi i} [K^\alpha f + e^{i(\alpha-1)\pi} I^\alpha f] \end{aligned}$$

for  $\alpha \neq 0, \pm 1, \pm 2, \dots$ . For  $\alpha = -n, n = 0, 1, 2, \dots$  we find

$$(3.4b) \quad \begin{aligned} F_{-n}(x + i0) &= \frac{1}{2} f^{(n)} - \frac{1}{2\pi i} \text{Pv} \frac{1}{x} * f^{(n)} \\ F_{-n}(x - i0) &= -\frac{1}{2} f^{(n)} - \frac{1}{2\pi i} \text{Pv} \frac{1}{x} * f^{(n)} \end{aligned}$$

where  $\text{Pv}(1/x)$  is the Cauchy principal value distribution. Using the identity  $\Gamma(\alpha)\Gamma(1-\alpha) = \pi/\sin(\alpha\pi)$  we obtain for  $\alpha \neq 1, 2, 3, \dots$  the analogue of (3.1) and (3.2)

$$(3.5) \quad \begin{aligned} F_\alpha(x + i0) - F_\alpha(x - i0) &= I^\alpha f, \\ e^{i\alpha\pi} F_\alpha(x + i0) - e^{-i\alpha\pi} F_\alpha(x - i0) &= K^\alpha f. \end{aligned}$$

LEMMA 3.1. Let  $f \in \mathcal{D}'([-1, 1])$  and  $F_\alpha(z)$  be defined by (3.3) for  $\alpha \neq 1, 2, 3, \dots$ . Define  $z^{-\alpha}$  by taking  $0 < \arg z < 2\pi$ . Then  $F_\alpha(z)z^{-\alpha}$  extends to an analytic function on  $|z| > 1$  and

$$(3.6) \quad |F_\alpha(z)z^{-\alpha}| = O(|z|^{-1}) \quad \text{as } |z| \rightarrow \infty.$$

Proof. The analyticity of  $F_\alpha(z)z^{-\alpha}$  follows from the observation that the boundary values of  $F_\alpha(x + iy)(x + iy)^{-\alpha}$  satisfy

$$e^{i\alpha\pi} [F_\alpha(x + i0)(x + i0)^{-\alpha} - F_\alpha(x - i0)(x - i0)^{-\alpha}] = \begin{cases} |x|^{-\alpha} I^\alpha f = 0, & x < -1, \\ |x|^{-\alpha} K^\alpha f = 0, & x > 1, \end{cases}$$



so that  $F_\alpha(z)z^{-\alpha}$ , as defined for  $\text{Im } z < 0$ , is the analytic continuation  $F_\alpha(z)z^{-\alpha}$ , as defined for  $\text{Im } z > 0$ , across the real intervals  $(-\infty, -1)$  and  $(1, \infty)$  [21].

To derive (3.6), we note that  $f$  has compact support, so that there exist an interval  $[-a, a]$ , a constant  $M$  and an integer  $n \geq 0$ , such that for  $|z| \geq 2a$

$$\begin{aligned} \left| \frac{1}{2\pi i} \langle f(t), (t-z)^{\alpha-1} \rangle \right| &\leq M \sup \left\{ \left| \frac{d^n}{dt^n} (t-z)^{\alpha-1} \right| : |t| \leq a \right\} \\ &= M |(\alpha-1)(\alpha-2) \cdots (\alpha-n)| |z|^{\alpha-n-1} \\ &\quad \cdot \sup \left\{ \left| \frac{t}{z} - 1 \right|^{\alpha-n-1} : |t| \leq a \right\} \leq C |z|^{\alpha-1} \end{aligned}$$

for some constant  $C$ .

**COROLLARY 3.1.** *If  $\text{Re } \alpha < 1$ , then  $F_\alpha(z)$  as defined by (3.3) is the only analytic representation of  $I^\alpha f$  which satisfies (3.6).*

*Proof.* Analytic representations are unique up to addition of an arbitrary entire analytic function. Any entire function satisfying (3.6) must be identically zero.

**THEOREM 3.1.** *Let  $f \in \mathcal{D}'([-1, 1])$ . For  $\alpha \in \mathbb{C}$  with  $\alpha \neq 1, 2, \dots$  let  $F_\alpha(z)$  be defined by (3.3). Then*

(a)  $F_\alpha(z)$  is analytic on the complement of  $\{z \in \mathbb{C} : \text{Im } z = 0 \wedge \text{Re } z \in \text{supp } I^\alpha f\}$ . As  $y \rightarrow 0$  in the upper or lower half-plane, respectively,  $F_\alpha(x + iy)$  has boundary values in  $\mathcal{D}'(\mathcal{R})$  satisfying (3.5) and

- (i)  $F_\alpha(x + i0) - F_\alpha(x - i0) = 0$  on  $(-\infty, -1)$
- (ii)  $e^{i\alpha\pi} F_\alpha(x + i0) - e^{-i\alpha\pi} F_\alpha(x - i0) = 0$  on  $(1, \infty)$
- (iii)  $|F_\alpha(z)| = O(|z|^{\alpha-1})$  as  $|z| \rightarrow \infty$ .

(b) If  $G_\alpha(z)$  is any other analytic representation of  $I^\alpha f$  satisfying (i)–(iii), then  $G_\alpha(z) = F_\alpha(z)$ . If  $\alpha \neq 0, \pm 1, \dots$  then conditions (i)–(iii) are equivalent to (i)–(ii). If  $\alpha = 0, -1, \dots$  then conditions (i)–(iii) are equivalent to (i) and (iii).

*Proof.* (a) Equations (i) and (ii) follow from (3.5) and the facts that by (2.4)  $I^\alpha f = 0$  on  $(-\infty, -1)$  and  $K^\alpha f = 0$  on  $(1, \infty)$ . (iii) was proved in Lemma 3.1.

(b) Let  $E = G_\alpha - F_\alpha$ . Then  $E$  is an entire analytic function, since  $G_\alpha$  and  $F_\alpha$  are analytic representations of the same distributions.

For  $\alpha \neq 0, \pm 1, \dots$  suppose  $G_\alpha$  also satisfies (ii). Then we have for  $E(x)$  (as a distribution in  $\mathcal{D}'(\mathcal{R})$ )

$$e^{i\alpha\pi} E(x) - e^{-i\alpha\pi} E(x) = 2 \sin(\alpha\pi) E(x) = 0 \quad \text{on } (1, \infty).$$

Since  $\sin(\alpha\pi) \neq 0$ , this implies that  $E(x) = 0$  on  $(1, \infty)$ . Since  $E$  is entire,  $E \equiv 0$ .

If  $\alpha = 0, -1, \dots$ , and  $G_\alpha$  satisfies (iii), then  $|E(z)| = O(|z|^{-1})$ , so that again,  $E \equiv 0$ . This completes the proof.

Theorem 3.1 states that for  $\alpha \neq 1, 2, \dots$  every solution  $f$  of the convolution equation in Problem I defines a solution  $F_\alpha(z)$ , given by (3.2), of the Hilbert boundary value problem, Problem II. The converse is also true as we shall see in the next section.

**4. The equivalence of the integral equation and the Hilbert problem.** In order to establish that every solution of the Hilbert Problem II yields a solution of the integral equation Problem I, we first characterize the class of analytic representations within which solutions of the Hilbert problem are sought, by means of a Paley–Wiener type theorem.

**THEOREM 4.1.** *Suppose  $\alpha \in \mathbb{C}$  and  $F_\alpha(z)$  is an analytic representation of a distribution in  $\mathcal{D}'(\mathcal{R})$  such that*

$$(4.1) \quad \begin{aligned} & \text{(i) if } f_\alpha = F_\alpha(x+i0) - F_\alpha(x-i0), \text{ then } f_\alpha = 0 \text{ on } (-\infty, -1) \\ & \text{(ii) if } g_\alpha = e^{i\alpha\pi}F_\alpha(x+i0) - e^{-i\alpha\pi}F_\alpha(x-i0), \text{ then } g_\alpha = 0 \text{ on } (1, \infty) \\ & \text{(iii) } |F_\alpha(z)z^{-\alpha}| = O(|z|^{-1}) \text{ as } |z| \rightarrow \infty. \end{aligned}$$

Then  $I^{-\alpha}f_\alpha = K^{-\alpha}g_\alpha$  and has compact support contained in  $[-1, 1]$ .

*Proof.* We distinguish three cases: (1)  $\alpha = n, n = 0, 1, 2, \dots$ , (2)  $\alpha = -n, n = 1, 2, 3, \dots$  and (3)  $\alpha \neq \pm n, n = 0, 1, 2, \dots$ .

(1)  $\alpha = n, n = 0, 1, 2, \dots$ : In this case,  $g_n = (-1)^n f_n$ , so that by definition

$$I^{-n}f_n = f_n^{(n)} = (-1)^n g_n^{(n)} = K^{-n}g_n.$$

Clearly,  $\text{supp } I^{-n}f_n \subset \text{supp } f_n^{(n)} \cap \text{supp } g_n^{(n)} \subset [-1, 1]$ .

(2)  $\alpha = -n, n = 1, 2, 3, \dots$ : Again,  $g_n = (-1)^n f_n$ , and has compact support. Thus

$$f_n = \frac{d^n}{dx^n}(I^n f_n) = (-1)^n \frac{d^n}{dx^n}(K^n f_n) = \frac{d^n}{dx^n}(K^n g_n)$$

which implies that for some polynomial  $p_{n-1}$  of degree  $n - 1$ ,

$$I^n f_n = K^n g_n + p_{n-1}.$$

If we can show that there exists  $f \in \mathcal{D}'([-1, 1])$  with  $f^{(n)} = f_n$ , we are done, since in that case,  $K^n g_n = f + q_{n-1}$ , where  $q_{n-1}$  is a polynomial of degree  $n - 1$ , which vanishes for  $x > 1$ , i.e.  $q_{n-1} \equiv 0$ , and  $K^n g_n = f$ . It follows that  $p_{n-1} = I^n f_n - f = 0$  for  $x < -1$ , and thus  $p_{n-1} \equiv 0$  as well, so that

$$I^n f_n = K^n g_n = f \in \mathcal{D}'([-1, 1]).$$

The existence of  $f$ , however, is an immediate consequence of the following argument:  $F_{-n}(z)$  is analytic for  $|z| > 1$  and  $|F_{-n}(z)| = O(|z|^{-n-1})$  as  $|z| \rightarrow \infty$ . Thus there exists  $F(z)$  analytic for  $|z| > 1$  and of order  $O(|z|^{-1})$  as  $|z| \rightarrow \infty$ , such that  $F^{(n)}(z) = F_{-n}(z)$ . Taking  $f = F(x+i0) - F(x-i0)$ , we obtain the distribution  $f$  as desired.

(3)  $\alpha \neq 0, \pm 1, \pm 2, \dots$ : Let us first prove that if  $I^{-\alpha}f_\alpha$  has compact support in  $[-1, 1]$ , then  $I^{-\alpha}f_\alpha = K^{-\alpha}g_\alpha$ . To do so, let  $h = I^{-\alpha}f_\alpha$ , and let  $H_\alpha(z)$  be the analytic representation of  $I^\alpha h$  defined by (3.12). Then  $H_\alpha - F_\alpha = E$  for some entire analytic function  $E$ . It suffices to show that  $E \equiv 0$ . Now (ii) and (3.5) imply that

$$e^{i\alpha\pi}E(x) - e^{-i\alpha\pi}E(x) = 2i \sin(\alpha\pi)E(x) = 0 \quad \text{for } x > 1.$$

Thus  $E = 0$  since  $\alpha \neq 0, \pm 1, \pm 2, \dots$ .

It remains to be shown that  $I^{-\alpha}f_\alpha$  has compact support in  $[-1, 1]$ . Since  $I^{-\alpha}f_\alpha = x_+^{-\alpha-1}/\Gamma(-\alpha) * f_\alpha$ , and  $\text{supp } f_\alpha \subset [-1, \infty)$ , it follows that  $\text{supp } (I^{-\alpha}f_\alpha) \subset [-1, \infty) + [0, \infty) = [-1, \infty)$ . Therefore the proof is completed, if we can show that  $I^{-\alpha}f_\alpha = 0$  on  $(1, \infty)$ , or, equivalently, that  $\langle I^{-\alpha}f_\alpha, \varphi \rangle = 0$  for all  $\varphi \in \mathcal{D}(\mathcal{R})$  with  $\text{supp } \varphi \subset (1, \infty)$ .

Using the definition of the convolution of two distributions we have for  $\varphi$  with  $\text{supp } \varphi \subset (1, \infty)$

$$(4.2) \quad \langle I^{-\alpha}f_\alpha, \varphi \rangle = \langle f_\alpha, K^{-\alpha}\varphi \rangle.$$

For  $\text{Im } z \neq 0$ , equation (3.3) reads, with  $f$  replaced by  $\varphi$  and  $\alpha$  by  $-\alpha$ :

$$(4.3) \quad \Phi_{-\alpha}(z) = \frac{\Gamma(1+\alpha)}{2\pi i} \int_{-\infty}^{\infty} \varphi(t)(t-z)^{-\alpha-1} dt.$$

Then  $|\Phi_{-\alpha}(z)| = O(|z|^{-\alpha-1})$  as  $|z| \rightarrow \infty$ .  $\Phi_{-\alpha}(x+i\varepsilon)$  and  $\Phi_{-\alpha}(x-i\varepsilon)$  converge to

infinitely differentiable limits  $\Phi_{-\alpha}(x+i0)$  and  $\Phi_{-\alpha}(x-i0)$  and the convergence is uniform on compact subsets of  $\mathcal{R}$ .

Let  $a \in \mathcal{R}, a > 1$  be sufficiently large so that  $\text{supp } \varphi \subset [-a, a]$ . Then using the analytic representation  $F_\alpha(z)$  of  $f_\alpha$  we have

$$(4.4) \quad \langle f_\alpha, K^{-\alpha} \varphi \rangle = \lim_{\varepsilon \downarrow 0} \int_{-a}^a [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] K^{-\alpha} \varphi(x) dx.$$

Choose  $\delta > 0$  so that  $\text{supp } \varphi \subset (1+\delta, \infty)$ . Define

$$(4.5) \quad \begin{aligned} I_1(\varepsilon) &= \int_{-a}^{1+\delta} [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] K^{-\alpha} \varphi(x) dx \\ &= \int_{-a}^{1+\delta} [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] [e^{-i\alpha\pi} \Phi_{-\alpha}(x+i0) - e^{i\alpha\pi} \Phi_{-\alpha}(x-i0)] dx. \end{aligned}$$

On  $(-\infty, 1+\delta]$ ,  $I^{-\alpha} \varphi = 0$ , i.e.  $\Phi_{-\alpha}(x+i0) = \Phi_{-\alpha}(x-i0)$ . Thus

$$\begin{aligned} I_1(\varepsilon) &= \int_{-a}^{1+\delta} F_\alpha(x+i\varepsilon) \Phi_{-\alpha}(x+i0) [e^{-i\alpha\pi} - e^{i\alpha\pi}] dx \\ &\quad + \int_{-a}^{1+\delta} F_\alpha(x-i\varepsilon) \Phi_{-\alpha}(x-i0) [e^{-i\alpha\pi} - e^{i\alpha\pi}] dx. \end{aligned}$$

Define

$$(4.6) \quad \begin{aligned} I_2(\varepsilon) &= \int_{1+\delta}^a [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] K^{-\alpha} \varphi(x) dx \\ &= \int_{1+\delta}^a [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] e^{-i\alpha\pi} \Phi_{-\alpha}(x+i0) dx \\ &\quad - \int_{1+\delta}^a [F_\alpha(x+i\varepsilon) - F_\alpha(x-i\varepsilon)] e^{i\alpha\pi} \Phi_{-\alpha}(x-i0) dx \end{aligned}$$

and

$$(4.7) \quad \begin{aligned} \tilde{I}_2(\varepsilon) &= \int_{1+\delta}^a F_\alpha(x+i\varepsilon) \Phi_{-\alpha}(x+i0) [e^{-i\alpha\pi} - e^{i\alpha\pi}] dx \\ &\quad - \int_{1+\delta}^a F_\alpha(x-i\varepsilon) \Phi_{-\alpha}(x-i0) [e^{-i\alpha\pi} - e^{i\alpha\pi}] dx. \end{aligned}$$

Then

$$\begin{aligned} I_2(\varepsilon) - \tilde{I}_2(\varepsilon) &= - \int_{1+\delta}^a [F_\alpha(x+i\varepsilon) e^{i\alpha\pi} - F_\alpha(x-i\varepsilon) e^{-i\alpha\pi}] \Phi_{-\alpha}(x+i0) dx \\ &\quad - \int_{1+\delta}^a [F_\alpha(x+i\varepsilon) e^{i\alpha\pi} - F_\alpha(x-i\varepsilon) e^{-i\alpha\pi}] \Phi_{-\alpha}(x-i0) dx \end{aligned}$$

and so

$$(4.8) \quad \lim_{\varepsilon \downarrow 0} [I_2(\varepsilon) - \tilde{I}_2(\varepsilon)] = 0.$$

Substitution of (4.5), (4.6) into (4.4) and application of (4.8) yields

$$\langle I^{-\alpha} f_{\alpha}, \varphi \rangle = \lim_{\varepsilon \downarrow 0} \left\{ \int_{-a}^a F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i0)[e^{-i\alpha\pi} - e^{i\alpha\pi}] dx - \int_{-a}^a F_{\alpha}(x-i\varepsilon)\Phi_{-\alpha}(x-i0)[e^{-i\alpha\pi} - e^{i\alpha\pi}] dx \right\}.$$

Now let  $\lambda(x) \equiv 1$  on  $[-a, a]$ ,  $\lambda \in \mathcal{D}(\mathcal{R})$ . Using the same arguments as above we obtain

$$\begin{aligned} \langle I^{-\alpha} f_{\alpha}, \varphi \rangle &= \lim_{\varepsilon \downarrow 0} [e^{-i\alpha\pi} - e^{i\alpha\pi}] \int_{-\infty}^{\infty} \lambda(x) \cdot [F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i0) - F_{\alpha}(x-i\varepsilon)\Phi_{-\alpha}(x-i0)] dx \\ &= [e^{-i\alpha\pi} - e^{i\alpha\pi}] \langle F_{\alpha}(x+i0)\Phi_{-\alpha}(x+i0) - F_{\alpha}(x-i0)\Phi_{-\alpha}(x-i0), \lambda(x) \rangle \\ &= [e^{-i\alpha\pi} - e^{i\alpha\pi}] \langle G_{\alpha}, \lambda \rangle \end{aligned}$$

where the products are well-defined, since  $\Phi_{-\alpha}(x+i0)$  and  $\Phi_{-\alpha}(x-i0)$  are infinitely differentiable. The distribution  $G_{\alpha}$  thus defined has  $F_{\alpha}(z)\Phi_{-\alpha}(z)$  as an analytic representation [21]. Therefore

$$\begin{aligned} \langle I^{-\alpha} f_{\alpha}, \varphi \rangle &= \lim_{\varepsilon \downarrow 0} \int_{-\infty}^{\infty} \lambda(x) [e^{-i\alpha\pi} - e^{i\alpha\pi}] \cdot [F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i\varepsilon) - F_{\alpha}(x-i\varepsilon)\Phi_{-\alpha}(x-i\varepsilon)] dx \end{aligned}$$

for every such  $\lambda$ . Thus the proof is completed if we can show that for  $\varepsilon > 0$

$$(4.9) \quad \int_{-\infty}^{\infty} F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i\varepsilon) dx = 0$$

and

$$(4.10) \quad \int_{-\infty}^{\infty} F_{\alpha}(x-i\varepsilon)\Phi_{-\alpha}(x-i\varepsilon) dx = 0.$$

To prove this, let  $\mathcal{C}_+$  denote the contour

$$\mathcal{C}_+ = \{z : (y = \varepsilon \wedge |x| < a) \vee (|z|^2 = \varepsilon^2 + a^2, y > \varepsilon)\}.$$

Since  $F_{\alpha}(z)\Phi_{-\alpha}(z)$  is analytic on the interior of  $\mathcal{C}_+$  we have

$$\begin{aligned} 0 &= \int_{\mathcal{C}_+} F_{\alpha}(z)\Phi_{-\alpha}(z) dz = \int_{-a}^a F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i\varepsilon) dx \\ &\quad + \int_{\theta=\theta_0}^{\pi-\theta_0} F_{\alpha}(R e^{i\theta})\Phi_{-\alpha}(R e^{i\theta})R e^{i\theta} d\theta \end{aligned}$$

where  $R = \sqrt{a^2 + \varepsilon^2}$ ,  $\theta_0 = \tan^{-1}(\varepsilon/a)$ . If  $a$  is sufficiently large, then  $|\int_{-a}^a F_{\alpha}(x+i\varepsilon)\Phi_{-\alpha}(x+i\varepsilon) dx| \leq C\pi a^{-1}$  for some constant  $C$  such that  $|F_{\alpha}(R e^{i\theta})/R^{\alpha-1} \cdot \Phi_{-\alpha}(R e^{i\theta})/R^{-\alpha-1}| \leq C$ . Since  $a$  was arbitrarily large (4.9) follows. Similarly, we prove (4.10). This completes the proof.

DEFINITION 4.1. For  $\alpha \in \mathbb{C}$ , let  $\mathcal{A}_{\alpha}$  denote the set of functions  $F_{\alpha}(z)$  analytic for  $\text{Im } z \neq 0$  with boundary values in  $\mathcal{D}'(\mathcal{R})$  satisfying

- (i)  $F_{\alpha}(x+i0) - F_{\alpha}(x-i0) = 0$  on  $(-\infty, -1)$ ;
- (ii)  $e^{i\alpha\pi}F_{\alpha}(x+i0) - e^{-i\alpha\pi}F_{\alpha}(x-i0) = 0$  on  $(1, \infty)$ ;
- (iii)  $|F_{\alpha}(z)z^{-\alpha}| = O(|z|^{-1})$  as  $|z| \rightarrow \infty$ .

From Theorems 3.1 and 4.1 follows

COROLLARY 4.1. For each  $\alpha \in \mathbb{C}, \alpha \neq 1, 2, 3, \dots$ , the map

$$(4.11) \quad f \mapsto \frac{\Gamma(1-\alpha)}{2\pi i} \langle f(t), (t-z)^{\alpha-1} \rangle$$

establishes a one-to-one correspondence between  $\mathcal{D}'([-1, 1])$  and  $\mathcal{A}_\alpha$ . The inverse of (4.11) given by

$$(4.12) \quad F_\alpha \mapsto I^{-\alpha} [F_\alpha(x+i0) - F_\alpha(x-i0)].$$

Corollary 4.1 is equivalent to Theorem 2.1:

COROLLARY 4.2. For  $\alpha \in \mathbb{C}, \alpha \neq 1, 2, 3, \dots$ , Problems I and II are equivalent. A one-to-one correspondence between their solutions is given by (4.11) and (4.12).

If  $\alpha = 1, 2, 3, \dots$  Theorem 2.1, or, equivalently, Corollary 4.1 are no longer true. However, a relationship between Problems I and II can still be exhibited.

DEFINITION 4.2. Let  $\tilde{\mathcal{A}}_n$  denote the set of analytic representations  $F_n(z)$  of distributions in  $\mathcal{D}'(\mathcal{R})$  whose boundary values satisfy

$$(4.13) \quad \begin{aligned} & \text{(i)} \quad F_n(x+i0) - F_n(x-i0) = 0 \quad \text{on } x < -1; \\ & \text{(ii)} \quad (-1)^n [F_n(x+i0) - F_n(x-i0)] + q_{n-1} = 0 \quad \text{on } x > 1; \\ & \text{(iii)} \quad |F_n(z)| = O(|z|^{n-1}); \end{aligned}$$

for some polynomial  $q_{n-1}(x)$  of degree  $\leq n-1$ .

If  $F_n \in \tilde{\mathcal{A}}_n$ , it follows immediately that  $(d^n/dx^n)[F_n(x+i0) - F_n(x-i0)]$  belongs to  $\mathcal{D}'([-1, 1])$  and yields the zero element in  $\mathcal{D}'([-1, 1])$  if and only if  $F_n(z)$  is a polynomial of degree  $\leq n-1$ . Thus, let  $\mathcal{P}_{n-1}$  denote the space of polynomials in  $z$  (over  $\mathbb{C}$ ) of degree  $\leq n-1$ . Form the quotient space  $\tilde{\mathcal{A}}_n \text{ mod } \mathcal{P}_{n-1}$  whose elements we write as  $F_n + \mathcal{P}_{n-1}$ . Then the map

$$F_n + \mathcal{P}_{n-1} \mapsto \frac{d^n}{dx^n} [F_n(x+i0) - F_n(x-i0)]$$

is well-defined and maps  $\tilde{\mathcal{A}}_n \text{ mod } \mathcal{P}_{n-1}$  into  $\mathcal{D}'([-1, 1])$ . As we shall see, this map is 1-1 and onto.

Let  $\mathbb{C}^*$  be the complex plane with the real interval  $[-1, 1]$  removed. Given  $z_0, z \in \mathbb{C}^*$  let  $\Gamma_0$  be any path in  $\mathbb{C}^*$  beginning at  $z_0$  and ending at  $z$  which does not intersect the positive real axis. Let  $\mathcal{C}(z_0, z)$  be the family of all paths in  $\mathbb{C}^*$  beginning at  $z_0$  and ending at  $z$  which are equivalent in  $\mathbb{C}^*$  to  $\Gamma_0$ . For  $f \in \mathcal{D}'([-1, 1])$  let  $F_0(z)$  be its Cauchy "Integral" given by (2.10) and define

$$(4.14) \quad A_n[f; z_0](z) = \frac{1}{(n-1)!} \int_{\Gamma \in \mathcal{C}(z_0, z)} F_0(\zeta) (z-\zeta)^{n-1} d\zeta.$$

LEMMA 4.1. The map

$$(4.15) \quad f \mapsto A_n[f; z_0] + \mathcal{P}_{n-1}$$

defines a one-to-one correspondence between  $\mathcal{D}'([-1, 1])$  and  $\tilde{\mathcal{A}}_n \text{ mod } \mathcal{P}_{n-1}$  whose inverse is given by

$$(4.16) \quad F_n + \mathcal{P}_{n-1} \mapsto \frac{d^n}{dx^n} [F_n(x+i0) - F_n(x-i0)].$$

*Proof.* Let  $G_n(z) = A_n[f; z_0](z) + p_{n-1}(z)$ , where  $p_{n-1} \in \mathcal{P}_{n-1}$ . From Definition (4.14) it follows that  $G_n(x+i0) - G_n(x-i0) = I^n f$  and that  $|G_n(z)| = O(|z|^{n-1})$ . Thus  $G_n$  satisfies (4.13) (i) and (iii). To prove (ii), let  $\Gamma$  be any closed path in  $\mathbb{C}^*$  which encircles the interval  $[-1, 1]$  once clockwise, and let  $\Gamma = \Gamma_+ + \Gamma_-$ , where  $\Gamma_+ = \{z \in \Gamma: \text{Im } z \geq 0\}$  and

$\Gamma_- = \{z \in \Gamma: \text{Im } z \leq 0\}$ . It is a simple exercise to show that

$$H_n(z) = \begin{cases} (-1)^n G_n(z) + \frac{1}{(n-1)!} \int_{\Gamma_+} F_0(\zeta)(\zeta - z)^{n-1} d\zeta & \text{for } \text{Im } z > 0, \\ (-1)^n G_n(z) - \frac{1}{(n-1)!} \int_{\Gamma_-} F_0(\zeta)(\zeta - z)^{n-1} d\zeta & \text{for } \text{Im } z < 0 \end{cases}$$

defines an analytic representation of  $K^n f$ . Now for  $z \in \mathbb{C}$

$$\frac{1}{(n-1)!} \int_{\Gamma} F_0(\zeta)(\zeta - z)^{n-1} d\zeta = \langle f(t), (t - z)^{n-1} / (n-1)! \rangle$$

is a polynomial of degree  $\leq n - 1$ . Thus we find that

$$(-1)^n [G_n(x + i0) - G_n(x - i0)] + \langle f(t), (t - x)^{n-1} / (n-1)! \rangle = K^n f$$

which shows that  $G_n \in \tilde{\mathcal{A}}_n$ . Since  $(d^n/dx^n)[G_n(x + i0) - G_n(x - i0)] = (d^n/dx^n)(I^n f) = f$ , the assertion is proved.

Lemma 4.2 relates Problems I and II in the following way:

**COROLLARY 4.3.** For  $\alpha = n, n = 1, 2, 3, \dots$  every solution  $f$  of Problem I is of the form

$$f = \frac{d^n}{dx^n} [F_n(x + i0) - F_n(x - i0)]$$

for some  $F_n \in \tilde{\mathcal{A}}_n$  such that

$$\begin{aligned} F_n(x + i0) - F_n(x - i0) &= 0 \quad \text{on } x < -1, \\ (a + (-1)^n b)[F_n(x + i0) - F_n(x - i0)] &= g + q_{n-1} \quad \text{on } |x| < 1, \\ (-1)^n [F_n(x + i0) - F_n(x - i0)] &= q_{n-1} \quad \text{on } x > 1 \end{aligned}$$

where  $q_{n-1}$  is some polynomial of degree  $\leq n - 1$ .

Conversely, every solution  $F_n(z)$  of Problem II is of the form

$$F_n(z) = A_n[f; z_0](z) + p_{n-1}(z)$$

for some  $z_0 \in \mathbb{C}^*$ ,  $p_{n-1} \in \mathcal{P}_{n-1}$  and  $f \in \mathcal{D}'([-1, 1])$  which is a solution of Problem I and satisfies, in addition,

$$\langle f(t), t^k \rangle = 0 \quad \text{for } k = 0, 1, 2, \dots, n - 1.$$

**5. Solution of the Hilbert problem.** We present here without proof an existence theorem for solutions of the Hilbert boundary value problem, Problem II. Formal proofs of all statements and results are contained in [22] where Hilbert problems for analytic representations of distributions are defined and discussed in detail. It is worthwhile, however, to repeat here the general arguments and definitions.

**DEFINITION 5.1.** The coefficients  $a$  and  $b$  for the generalized Abel equation are called *proper*, if

- (i)  $a$  and  $b$  are infinitely differentiable on  $(-1, 1)$ ;
- (ii)  $c_1 = a + b e^{i\alpha\pi}$  and  $c_2 = a + b e^{-i\alpha\pi}$  are nonzero on  $(-1, 1)$ ;
- (iii) there exist  $\beta_j, \gamma_j \in \mathcal{R}$  such that the limits

$$\lim_{x \downarrow -1} \frac{d^k}{dx^k} [c_j(x)(1+x)^{\beta_j}] \quad \text{and} \quad \lim_{x \uparrow 1} \frac{d^k}{dx^k} [c_j(x)(1-x)^{\gamma_j}]$$

exist for  $k = 0, 1, 2, \dots, j = 1, 2$  are nonzero for  $k = 0$ .

Let  $a$  and  $b$  be proper coefficients. Define

$$(5.1) \quad \chi(x) = \begin{cases} e^{-i2\alpha\pi}, & x > 1, \\ \frac{a(x) + b(x) e^{-i\alpha\pi}}{a(x) + b(x) e^{i\alpha\pi}}, & |x| < 1, \\ 1, & x < -1. \end{cases}$$

Then Problem II is equivalent to that of finding  $F_\alpha(z)$  of order  $O(|z|^{\alpha-1})$  such that, for  $g \in \mathcal{D}'([-1, 1])$

$$(5.2) \quad F_\alpha(x + i0) - \chi(x)F_\alpha(x - i0) = g(x)/[a(x) + b(x) e^{i\alpha\pi}]$$

on  $\mathcal{R} \setminus \{-1, 1\}$ . The assumption that  $a$  and  $b$  are proper coefficients, assures us that there exists a distribution  $v_0$  with support on  $[-1, 1]$ , which is infinitely differentiable on  $(-1, 1)$  and yields

$$(5.3) \quad \chi(x) = e^{v_0(x)} \quad \text{on } |x| < 1.$$

Define

$$(5.4) \quad V_0(z) = \frac{1}{2\pi i} \langle v_0(t), (t - z)^{-1} \rangle, \quad \text{Im } z \neq 0,$$

and

$$(5.5) \quad M(z) = \begin{cases} e^{-i\alpha\pi} (z - 1)^\alpha e^{V_0(z)}, & \text{Im } z > 0, \\ e^{i\alpha\pi} (z - 1)^\alpha e^{V_0(z)}, & \text{Im } z < 0. \end{cases}$$

Then  $M(z)$  and  $1/M(z)$  are analytic representations of distributions in  $\mathcal{D}'(\mathcal{R})$  and have infinitely differentiable boundary values on  $\mathcal{R} \setminus \{-1, 1\}$ . Furthermore  $|M(z)| = O(|z|^\alpha)$  as  $|z| \rightarrow \infty$ . We can now factor  $\chi$ :

$$(5.6) \quad \chi(x)[1/M(x + i0)] = 1/M(x - i0) \quad \text{for } |x| \neq 1.$$

Let  $k \in \mathcal{D}'([-1, 1])$  such that on  $(-1, 1)$

$$(5.7) \quad k(x)M(x + i0) = g(x)/[a(x) + b(x) e^{i\alpha\pi}].$$

Such  $k$  exists, if  $a$  and  $b$  are proper [22]. Let

$$(5.8) \quad K(z) = \frac{1}{2\pi i} \left\langle k(t), \frac{1}{t - z} \right\rangle, \quad \text{Im } z \neq 0.$$

Then every solution of Problem II is of the form

$$(5.9) \quad F_\alpha(z) = M(z) \left[ K(z) + \sum_{j=0}^N [a_j(z - 1)^{-j-1} + b_j(z + 1)^{-j-1}] \right].$$

We thus summarize as follows:

**THEOREM 5.1.** *For  $\alpha \in \mathbb{C}$ , let  $a$  and  $b$  be proper coefficients for the generalized Abel equation (1.1). Then Problem I and II have solutions for every  $g \in \mathcal{D}'([-1, 1])$ .*

*Every solution of the Hilbert boundary value problem is of the form (5.9) for some constants  $a_j, b_j, j = 0, 1, \dots, N$  and integer  $N \geq 0$ . If  $\alpha \neq 1, 2, \dots$  then every solution of the Abel equation is of the form  $f = I^{-\alpha}[F_\alpha(x + i0) - F_\alpha(x - i0)]$  for some solution  $F_\alpha(z)$  of the corresponding Hilbert problem. If  $\alpha = 1, 2, \dots$  then Problem I has particular solutions  $f_p = (d^n/dx^n)[F_n(x + i0) - F_n(x - i0)]$  which satisfy  $\langle f_p(x), x^k \rangle = 0$  for  $k = 0, 1, \dots, n - 1$ ; the general solution in this case is  $f = f_p + \sum_{j=0}^{n-1} d_j \delta^{(j)}(1 - x)$  for arbitrary constants  $d_j, j = 0, 1, \dots, n - 1$ .*

**6. Examples.** Let us consider here the general solution of Problem I in some special cases. The formulae in examples 1 and 2 are easily checked. The results in examples 3 and 4 are obtained following the technique outlined in the previous section. The computations are elementary, but cumbersome, and thus omitted.

As before, we shall take  $g \in \mathcal{D}'([-1, 1])$ , and seek solutions  $f \in \mathcal{D}'([-1, 1])$ , satisfying  $S^\alpha f = g$  on  $(-1, 1)$ .

(a)  $S^\alpha = I^\alpha$ ;  $\alpha \in \mathbb{C}$ : (*Abel's equation*). The general solution of  $I^\alpha f = g$  is

$$(6.1) \quad f(x) = \frac{d^n}{dx^n} [\theta(1-x)I^{-\alpha+n}g] + \sum_{j=0}^n \left[ a_j \theta(1-x) \frac{(x+1)_+^{-\alpha-j-1}}{\Gamma(-\alpha-j)} + b_j \delta^{(j)}(x-1) \right]$$

where  $(x+1)_+^{\beta-1}/\Gamma(\beta)|_{\beta=-k} = \delta^{(k)}(x+1)$  for  $k = 0, 1, \dots, n \geq 0$  is an integer such that  $I^{-\alpha+n}g$  is integrable in a neighborhood of  $x = 1$ .

(b)  $S^\alpha = K^\alpha$ ;  $\alpha \in \mathbb{C}$ . The general solution of  $K^\alpha f = g$  is

$$(6.2) \quad f(x) = (-1)^n \frac{d^n}{dx^n} [\theta(1+x)K^{-\alpha+n}g] + \sum_{j=0}^n \left[ a_j \theta(1+x) \frac{(x-1)_-^{-\alpha-j-1}}{\Gamma(-\alpha-j)} + b_j \delta^{(j)}(x+1) \right]$$

where  $(x-1)_-^{\beta-1}/\Gamma(\beta)|_{\beta=-k} = (-1)^k \delta^{(k)}(x-1)$  for  $k = 0, 1, \dots, n \geq 0$  is an integer such that  $K^{-\alpha+n}g$  is integrable near  $x = -1$ .

(c)  $S^\alpha = R^\alpha$ ,  $\alpha \in \mathbb{C}$ ,  $\alpha \neq 0, \pm 1, \pm 2, \dots$  (*Riesz potential*). By definition  $R^\alpha = \frac{1}{2} \sec(\frac{1}{2}\alpha\pi)(I^\alpha + K^\alpha)$ . Every distribution  $f \in \mathcal{D}'([-1, 1])$  satisfying  $R^\alpha f = g$  on  $(-1, 1)$  is of the form

$$(6.3) \quad f = I^{-\alpha} f_\alpha$$

where  $f_\alpha(x) = F_\alpha(x+i0) - F_\alpha(x-i0)$  for some  $F_\alpha(z)$  given by

$$(6.4) \quad F_\alpha(z) = e^{\mp i\alpha\pi} (z^2-1)^{(1/2)\alpha} \left\{ \frac{(1-z^2)^{-n}}{2\pi i} \langle g(t)(1-t^2)^{n-(1/2)\alpha}, (t-z)^{-1} \rangle + \sum_{j=0}^N [a_j(z-1)^{-j-1} + b_j(z+1)^{-j-1}] \right\} \quad \text{for } \text{Im } z \geq 0.$$

Here  $(z^2-1)^{(1/2)\alpha}$  is defined by taking that branch which is analytic in the complex plane with the real interval  $|x| < 1$  removed and  $0 < \arg(z^2-1) < \pi$  for  $\text{Im } z > 0$  and  $-\pi < \arg(z^2-1) < 0$  for  $\text{Im } z < 0$ . On  $(-1, 1)$   $f_\alpha$  is given by

$$(6.5) \quad f_\alpha(x) = \cos(\frac{1}{2}\alpha\pi)g(x) - \sin(\frac{1}{2}\alpha\pi)(1-x^2)^{(1/2)\alpha-n}H[(1-x^2)^{n-(1/2)\alpha}g] - 2i \sin(\frac{1}{2}\alpha\pi)(1-x^2)^{(1/2)\alpha} \sum_{j=0}^N [a_j(x-1)^{-j-1} + b_j(x+1)^{-j-1}].$$

Here  $n$  is an integer such that  $n \geq 0$  and  $\text{Re}(n - \frac{1}{2}\alpha) > \max(n_1, n_2)$ , where  $n_1, n_2$  are the orders of  $g$  at  $-1$  and  $1$ , respectively.  $H$  stands for the Hilbert transform: if  $k \in \mathcal{D}'([-1, 1])$ , then  $Hk = -(1/\pi) \text{Pv } 1/x * k$ .

Let us note that if the integral makes sense, particular solutions  $f_\alpha$  in (6.5) are given by

$$(6.6) \quad f_\alpha(x) = \cos(\frac{1}{2}\alpha\pi)g(x) + \sin(\frac{1}{2}\alpha\pi) \frac{1}{\pi} \text{Pv} \int_{-1}^1 \left( \frac{1-x^2}{1-t^2} \right)^{(1/2)\alpha} \frac{g(t)}{x-t} dt.$$

(d)  $S^\alpha = H^\alpha$ ;  $\alpha \in \mathbb{C}$ ,  $\alpha \neq 0, \pm 1, \pm 2, \dots$  (*generalized Hilbert transform*). The operator  $H^\alpha$  is defined as  $H^\alpha = \frac{1}{2} \csc(\frac{1}{2}\alpha\pi)(I^\alpha - K^\alpha)$ . Note that for  $\alpha = 0$ ,  $H^\alpha$  reduces to the Hilbert transform in the sense that for  $\varphi \in \mathcal{D}(\mathcal{R})$ ,  $H^\alpha \varphi$  converges to  $H\varphi$



uniformly on compact subsets  $\mathcal{R}$ . (In fact,  $H^\alpha \varphi$  converges to  $H\varphi$  in the topology of  $\mathcal{E}(\mathcal{R})$ ). Every distribution  $f \in \mathcal{D}'([-1, 1])$  satisfying  $H^\alpha f = g$  on  $(-1, 1)$  is given by

$$(6.7) \quad f = I^{-\alpha} g$$

where  $f_\alpha(x) = F_\alpha(x + i0) - F_\alpha(x - i0)$  with  $F_\alpha(z)$  of the form

$$(6.8) \quad F_\alpha(z) = e^{\mp i\alpha\pi} (z^2 - 1)^{(1/2)\alpha - 1/2} \left\{ \frac{(1 - z^2)^{-n}}{2\pi i} \langle g(t)(1 - t^2)^{n - (1/2)\alpha + 1/2}, (t - z)^{-1} \rangle + \sum_{j=0}^N [a_j(z - 1)^{-j-1} + b_j(z + 1)^{-j-1}] \right\} \quad \text{for } \text{Im } z \geq 0$$

with branch cuts as in example (c). On  $(-1, 1)$ ,  $f_\alpha$  is given by

$$(6.9) \quad f_\alpha(x) = \sin\left(\frac{1}{2}\alpha\pi\right)g(x) - \cos\left(\frac{1}{2}\alpha\pi\right)(1 - x^2)^{(1/2)\alpha - 1/2 - n} H[(1 - t^2)^{n - (1/2)\alpha + 1/2} g](x) - \cos\left(\frac{1}{2}\alpha\pi\right)(1 - x^2)^{(1/2)\alpha - 1/2} \sum_{j=0}^N [a_j(x - 1)^{-j-1} + b_j(x + 1)^{-j-1}]$$

where  $n \geq 0$  is an integer such that  $\text{Re}(n - \frac{1}{2}\alpha + \frac{1}{2}) > \max(n_1, n_2)$ ,  $n_1, n_2$  and  $H$  are as in (c). If we may take  $n = 0$ , and if the integral below makes sense, then particular solutions are given by

$$(6.10) \quad f_\alpha(x) = \sin\left(\frac{1}{2}\alpha\pi\right)g(x) + \cos\left(\frac{1}{2}\alpha\pi\right) \frac{1}{\pi} \text{Pv} \int_{-1}^1 \left(\frac{1 - x^2}{1 - t^2}\right)^{(1/2)\alpha - 1/2} \frac{g(t)}{x - t} dt.$$

REFERENCES

[1] P. L. BUTZER AND W. TREBELS, *Hilbert Transformation, gebrochene Integration and Differentiation*, Forsch.-ber. des Landes Nordrh.-Westf. Nr. 1889, 1968.  
 [2] H. BREMERMAN, *Distributions, Complex Variables, and Fourier Transforms of Distributions*, Addison-Wesley, Reading, MA, 1965.  
 [3] T. CARLEMAN, *Über die Abelsche Integralgleichung mit konstanten Integrationsgrenzen*, Math. Z., 15 (1922), pp. 111-120.  
 [4] F. V. CHUMAKOV, *General theory of integral equations with power law kernel*, Differential'nye Uravnenija, 2 (1966), 544-559. (In Russian.)  
 [5] A. ERDÉLYI, *Fractional integrals of generalized functions*, J. Austral. Math. Soc., 14 (1972), pp. 30-37.  
 [6] A. ERDÉLYI AND A. C. MCBRIDE, *Fractional integrals of distributions*, this Journal, 1 (1970), pp. 547-557.  
 [7] F. D. GAKHOV, *Boundary Value Problems*, Pergamon Press, London, 1966.  
 [8] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, vol. I, Academic Press, New York, 1964.  
 [9] P. HEYWOOD, *On a modification of the Hilbert transform*, J. London Math. Soc., 42 (1967), pp. 641-645.  
 [10] ———, *On the inversion of fractional integrals*, Ibid., 3 (1971), no. 2, pp. 531-538.  
 [11] E. HILLE AND R. S. PHILIPS, *Functional Analysis and Semigroups*, American Mathematical Society Colloquium Publications, vol. 31, Providence, RI, 1957.  
 [12] R. K. JUBERG, *On the boundedness of certain singular integral operators*, Coll. Math. Soc. János Bolyai, 5 (1970), pp. 305-318.  
 [13] ———, *Finite Hilbert transforms in  $L^p$* , Bull. Amer. Math. Soc., 78 (1972), pp. 435-438.  
 [14] ———, *The spectra for operators of a basic collection*, Ibid., 79 (1973), pp. 821-824.  
 [15] H. KOBER, *A modification of Hilbert transforms, the Weyl integral, and functional equations*, J. London Math. Soc., 42 (1967), pp. 42-50.  
 [16] A. MARTINEAU, *Distributions et valeurs aux bords des fonctions holomorphes*, Proc. Int. Summer Conf. (Lisbon) 1964, pp. 196-326.  
 [17] A. C. MCBRIDE, *A theory of fractional integration for generalized functions*, this Journal, 6 (1975), 583-599.  
 [18] N. J. MUSKHELISHVILI, *Singular Integral Equations*, Noordhoff, Groningen, the Netherlands, 1958.

- [19] G. O. OKIKIOLU, *A generalization of the Hilbert transform*, J. London Math. Soc., 40 (1965), pp. 27–30.
- [20] ———, *Fourier transforms and the operators  $H_\alpha$* , Proc. Cambridge Philos. Soc., 62 (1966), pp. 73–78.
- [21] M. ORTON, *Hilbert transforms, Plemelj relations, and Fourier transforms of distributions*, this Journal, 4 (1973), 656–670.
- [22] ———, *Hilbert boundary value problems—a distributional approach*, Proc. Royal Soc. Edinburgh, 76A (1977), pp. 193–208.
- [23] ———, *Functional integrals and Hilbert transforms of distributions*, Ibid., to appear.
- [24] B. S. RUBIN, *On operators of potential type on a segment of the real line*, Izv. Vysš. Učebn. Zaved. Mat., 114 (1971), pp. 71–76. (In Russian.)
- [25] ———, *On operators of potential type in weight spaces on an arbitrary contour*, Dokl. Akad. Nauk SSSR, 207 (1972), pp. 300–303; (In Russian) = Soviet Math. Dokl., 13 (1972), pp. 1530–1534.
- [26] K. D. SAKALYUK, *A generalized Abel equation*, Dokl. Akad. Nauk SSSR, 131 (1960), pp. 748–751; (In Russian) = Soviet Math. Dokl., 1 (1960), pp. 332–335.
- [27] S. G. SAMKO, *Solution of generalized Abel equation by means of an equation with Cauchy kernel*, Dokl. Acad. Nauk SSSR, 176 (1967), pp. 1019–1022; (In Russian) = Soviet Math. Dokl., 8 (1967), pp. 1259–1262.
- [28] ———, *A generalized Abel equation and fractional integration operators*, Differencial'nye Uravnenija, 4 (1968), pp. 298–314. (In Russian.)
- [29] ———, *Noether's theory for the generalized Abel equation*, Differencial'nye Uravnenija, 4 (1968), pp. 315–326. (In Russian.)
- [30] ———, *Abel's generalized equation, Fourier transform, and convolution type equations*, Dokl. Nauk SSSR, 187 (1969); (In Russian) = Soviet Math. Dokl., 10 (1969), pp. 942–946.
- [31] ———, *Operators of potential type*, Dokl. Akad. Nauk SSSR, 196 (1971), pp. 299–301; (In Russian) = Soviet Math. Dokl., 12 (1971), pp. 125–128.
- [32] ———, *The space  $I^\alpha(L_p)$  of fractional integrals and operators of potential type*, Izv. Akad. Nauk Armjan. SSR Ser. Mat., 8 (1973), pp. 359–383, 425. (In Russian.)
- [33] L. SCHWARTZ, *Théorie des Distributions*, Hermann, Paris, 1966.
- [34] E. M. STEIN AND A. ZYGMUND, *On the fractional differentiability of functions*, Proc. London Math. Soc. Ser. A, 14 (1965), no. 3, pp. 249–264.
- [35] H. G. TILLMANN, *Darstellung der Schwartzschen Distributionen durch analytische Funktionen*, Math. Z., 77 (1961), pp. 106–124.
- [36] L. VON WOLFERSDORF, *Über eine Beziehung zwischen Integralen nichtganzer Ordnung*, Ibid., 90 (1965), pp. 24–28.
- [37] ———, *Abelsche Integralgleichungen und Randwertprobleme für die verallgemeinerte Tricomi-Gleichung*, Math. Nachr., 29 (1965), pp. 161–178.
- [38] ———, *Zur Lösung der verallgemeinerten Abelschen Integralgleichung mit konstanten Koeffizienten*, Z. Angew. Math. Mech., 49 (1969), pp. 759–761.

**MULTIPLE SOLUTIONS OF  
 TWO-POINT BOUNDARY VALUE PROBLEMS  
 OF NEUMANN TYPE WITH A SMALL PARAMETER\***

MASAYASU MIMURA,<sup>†</sup> MASAHISA TABATA<sup>‡</sup> AND YUZO HOSONO<sup>§</sup>

**Abstract.** This paper studies two-point boundary value problems for two-component systems with a small parameter  $\epsilon$ . The boundary conditions are of Neumann type. First it is shown that the reduced problem ( $\epsilon = 0$ ) has multiple solutions. With the aid of this result, the singular perturbation method is used for constructing large amplitude solutions of the original problem ( $\epsilon > 0$ ), which possess transition layers. As an application, a model system of prey-predator interaction with diffusion is considered.

**1. Introduction.** For asymptotic behaviors of solutions of interaction-diffusion equations, there are several interesting phenomena such as oscillation, wave propagation, stationary states and so on (see Fife [5] and its bibliography, for instance). By putting a restriction on systems in a bounded domain under zero flux boundary conditions, the asymptotic states may be classified into four categories; “spatially homogeneous stationary state or oscillation”, “spatially inhomogeneous stationary state”, “spatially inhomogeneous oscillation” and “spatio-temporal chaos” ([1], [15], [2], [8], for instance).

In this paper, we are interested in the second asymptotic state of a two component interaction-diffusion system in one dimensional space. The system considered here is of the form

$$(1.1a)_\epsilon \quad \begin{aligned} 0 &= \epsilon^2 \frac{d^2}{dx^2} u + f(u, v), \\ 0 &= \frac{d^2}{dx^2} v + g(u, v) \end{aligned} \quad (x \in I = (0, l)),$$

subject to zero flux boundary conditions

$$(1.1b)_\epsilon \quad \frac{d}{dx} u(0) = \frac{d}{dx} u(l) = 0 \quad \text{and} \quad \frac{d}{dx} v(0) = \frac{d}{dx} v(l) = 0.$$

It is known that the bifurcation theory can be used to study small amplitude solutions branching from constant solutions. On the other hand, when large amplitude solutions are studied, the asymptotic analysis is useful when  $\epsilon$  is small or large enough (see, for instance, [4], [7], [14]). We consider here the problem (1.1) <sub>$\epsilon$</sub>  when  $\epsilon$  is zero or a sufficiently small number.

The nonlinearities of  $f$  and  $g$  are assumed to be

- (i)  $f(u, v) = 0$  has at least two different solutions  $u = h_i(v)$  for some intervals  $J_i^*$  for  $i = 0, 1$  such that  $h_0(v) < h_1(v)$  in  $J_0^* \cap J_1^* (\neq \emptyset)$ , and
- (ii) there exist  $J_i (\subset J_i^*)$  for  $i = 0, 1$  such that  $g(h_1(v), v) > 0 > g(h_0(v), v)$  in  $J_0 \cup J_1$  and  $(d/dv)g(h_i(v), v) < 0$  in  $J_i$ .

The motivation of studying the problem (1.1) <sub>$\epsilon$</sub>  lies in the analysis of spatial patterns in morphogenetic and population dynamics models (for instance, [6], [13]). For an

\* Received by the editors March 9, 1979.

<sup>†</sup> Department of Applied Mathematics, Konan University, Kobe, Japan.

<sup>‡</sup> Department of Mathematics, Kyoto University, Kyoto, Japan.

<sup>§</sup> Department of Computer Sciences, Kyoto Sangyo University, Kyoto, Japan.

illustrative example, let us show a model system of prey-predator interaction with diffusion,

$$(1.2)_\epsilon \quad \begin{aligned} 0 &= \epsilon^2 \frac{d^2}{dx^2} u + \{f_0(u) - kv\}u, \\ 0 &= \frac{d^2}{dx^2} v - \{g_0(v) - ku\}v, \end{aligned}$$

where  $k$  is a positive constant. This system models a “plant-herbivore” like interaction. The variables  $u$  and  $v$  are the population densities of a prey and its predator, respectively. The functional forms of  $f_0$  and  $g_0$  are drawn in Fig. 1. It is shown in [13]

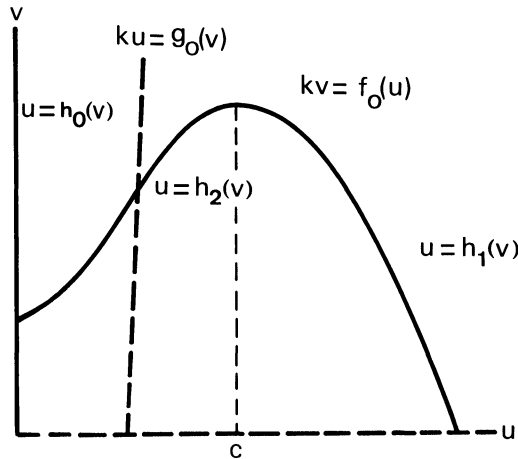


FIG. 1. Functional forms of  $(f_0(u) - kv)u = 0$  and  $(g_0(v) - ku)v = 0$  in  $(1.2)_\epsilon$ .

that  $(1.2)_\epsilon$  exhibits striking patchiness under suitable  $f_0$  and  $g_0$  when  $\epsilon$  is sufficiently small.

When  $\epsilon$  is zero, the problem  $(1.1)_0$  reduces to a boundary value problem for a single equation involving discontinuous nonlinearities, according to circumstances. To give an example we consider the system  $(1.2)_0$ . From the first equation, three different relations, say  $u = h_0(v)$ ,  $h_1(v)$  and  $h_2(v)$ , are obtained (we assume  $h_0 < h_2 < h_1$ , for simplicity). Using the function  $h_i(v)$ , we have the boundary value problem from  $(1.2)_0$ ,

$$(1.3) \quad \frac{d^2}{dx^2} v + G(v) = 0,$$

where  $G(v) = \{g_0(v) - kh_i(v)\}v$ . If two functions  $h_0$  and  $h_1$  are used,  $G(v)$  is constructed as a discontinuous function as follows: Taking one separating point, say  $\beta$ , in  $J_0 \cap J_1$ , we can define  $G(v)$  by

$$G(v) = \begin{cases} -\{g_0(v) - kh_0(v)\}v & (v \in J_0 \cap \{v < \beta\}), \\ -\{g_0(v) - kh_1(v)\}v & (v \in J_1 \cap \{v > \beta\}). \end{cases}$$

Thus, we find that the function  $G(v)$  has a point of discontinuity  $\beta$ . For this reason, we must seek a weak solution of  $(1.1)_0$ ,  $(U, V) \in L^2(I) \times H^1(I)$  which satisfies

$$\begin{aligned} f(U, V) &= 0 \quad \text{almost everywhere in } \bar{I}, \\ (V_x, \phi_x) &= (g(U, V), \phi) \quad \text{for all } \phi \in H^1(I), \end{aligned}$$

where  $(\cdot, \cdot)$  denotes the inner product in  $L^2(I)$ . The nonlinearity considered here is not included in ones treated by the monotone operator theory (see [10]). Stuart [18] has discussed an equation like (1.3) under Dirichlet boundary conditions. His nonlinearities are also different from ours.

When  $\varepsilon$  is not zero, Fife [3] argued the system  $(1.1a)_\varepsilon$  subject to Dirichlet boundary conditions under the assumption that a solution of the reduced problem  $(1.1a)_0$  exists, showing that this system exhibits boundary and interior transition layer phenomena. A solution of  $(1.1a)_0$  plays a lowest order approximation to a solution of  $(1.1a)_\varepsilon$  with sufficiently small  $\varepsilon$ .

In § 2, our main results will be shown.<sup>1</sup> One result concerns the existence of countably infinite number of nontrivial solutions of the problem  $(1.1)_0$ . It is seen that the first component of the solution is discontinuous, that is, it shows striking heterogeneity. The other is the existence of solutions of the problem  $(1.1)_\varepsilon$ . For this purpose, we follow Fife's arguments [3]. The result shows that interior transition layers appear in the first component. We shall apply these results to the specific ecological model  $(1.2)_\varepsilon$  and prove the existence of nonnegative solutions of  $(1.2)_\varepsilon$  subject to zero flux boundary conditions. Proofs are stated in §§ 3, 4 and 5. In § 6, we conclude with some remarks on the results obtained here.

We shall use the following notation throughout the paper:

$C^p(\bar{I})$  = the space of  $p$ -times continuously differentiable functions on  $\bar{I}$  with the norm

$$\|u\|_{C^p} = \sum_{k=0}^p \max_{x \in I} \left| \left( \frac{d}{dx} \right)^k u(x) \right|.$$

$C_0^p(\bar{I})$  = the subspace of  $C^p(\bar{I})$  with  $(d/dx)u(0) = u(l) = 0$ , ( $p \geq 1$ ).

$H^p(I)$  = the Sobolev space with the norm

$$\|u\|_{H^p(I)} = \left( \sum_{k=0}^p \int_0^l \left| \left( \frac{d}{dx} \right)^k u \right|^2 dx \right)^{1/2}.$$

$H^0(I)$  is usually denoted by  $L^2(I)$ .

$\mathcal{L}(X, Y)$  = the totality of continuous linear operators from  $X$  into  $Y$  equipped with the usual norm, where  $X$  and  $Y$  are Banach spaces.

For a positive number  $\varepsilon$ ,

$C_\varepsilon^p(\bar{I})$  = the space of  $p$ -times continuously differentiable functions on  $\bar{I}$  with the norm

$$\|u\|_{C_\varepsilon^p} = \sum_{k=0}^p \max_{x \in \bar{I}} \left| \left( \varepsilon \frac{d}{dx} \right)^k u(x) \right|.$$

$C_{\varepsilon_0}^p(\bar{I})$  = the subspace of  $C_\varepsilon^p(\bar{I})$  with  $(d/dx)u(0) = u(l) = 0$ , ( $p \geq 1$ ).

**2. Main results.** We study the boundary value problem for  $(u, v) = (u(x; \varepsilon), v(x; \varepsilon))$  in  $x \in I = (0, 1)$

$$(2.1a)_\varepsilon \quad \begin{aligned} 0 &= \varepsilon^2 \frac{d^2}{dx^2} u + f(u, v), \\ 0 &= \frac{1}{\sigma} \frac{d^2}{dx^2} v + g(u, v) \end{aligned}$$

<sup>1</sup> Part of them was reported at the meeting on Mathematics in Biology, at RIMS, Kyoto University, 1977, [12], [19].

with zero flux boundary conditions

$$(2.1b)_\varepsilon \quad \frac{d}{dx} u(0) = \frac{d}{dx} u(1) = 0 \quad \text{and} \quad \frac{d}{dx} v(0) = \frac{d}{dx} v(1) = 0,$$

where  $\varepsilon$  and  $\sigma$  are both positive constants. The length of the interval is normalized as unity. We shall consider two cases where (1)  $\varepsilon$  is zero and (2)  $\varepsilon$  and  $\sigma$  are both sufficiently small. We first impose the following assumptions on the nonlinearities of  $f$  and  $g$ :

- (A1) (i) The equation  $f(u, v)$  has at least two real distinct roots  $u = h_i(v)$  defined in intervals  $J_i^*$ , ( $i = 0, 1$ ).
- (ii)  $h_i(v) \in C^2(J_i^*)$  satisfies  $h_0(v) < h_1(v)$  in  $J_0^* \cap J_1^* \neq \emptyset$ , ( $i = 0, 1$ ).
- (A2) There exist two subintervals  $J_i = [c_i, d_i] \subset J_i^*$  ( $i = 0, 1$ ) such that
  - (i)  $G_i(v) \in C^1(J_i)$  and  $(d/dv)G_i(v) < 0$  in  $J_i$ , ( $i = 0, 1$ ),
  - (ii)  $G_1(v) > 0 > G_0(v)$  in  $(c_0, d_1)$ ,
 where  $J_0 \cap J_1 \neq \emptyset$  and  $G_i(v) = g(h_i(v), v)$ , ( $i = 0, 1$ ).

As seen in the introduction, from (2.1)<sub>0</sub>, we have a boundary value problem with discontinuous nonlinearities. Accordingly, in the case when  $\varepsilon$  is zero, we need to define the solution in a weak sense. We call  $(U(x), V(x))$  a solution of the problem (2.1)<sub>0</sub> if  $(U, V)$  satisfies

$$\begin{aligned} (U, V) &\in L^2(I) \times H^1(I), \\ f(U, V) &= 0 \quad \text{almost everywhere in } I, \text{ and} \\ \left( \frac{d}{dx} V, \frac{d}{dx} \phi \right) &= (\sigma g(U, V), \phi) \quad \text{for all } \phi \in H^1(I). \end{aligned}$$

For an arbitrarily fixed number  $\beta \in J_0 \cap J_1$ , we define  $G^\beta(v)$  by

$$G^\beta(v) = \begin{cases} G_0(v) & \text{for } v \in \{v < \beta\} \cap J_0, \\ G_1(v) & \text{for } v \in \{v > \beta\} \cap J_1. \end{cases}$$

Here  $G^\beta(\beta)$  remains undefined. As will be seen, however, this does not affect the construction of solutions except the trivial one (see Theorem 1). It is found that  $G^\beta(v)$  has a discontinuity of the first kind at  $v = \beta$ . Thus, the problem (2.1)<sub>0</sub> can be reduced to the two-point boundary value problem for a single equation:

$$(2.2a) \quad \frac{d^2}{dx^2} V + \sigma G^\beta(V) = 0 \quad (x \in I),$$

$$(2.2b) \quad \frac{d}{dx} V(0) = \frac{d}{dx} V(1) = 0.$$

For this problem, a solution  $V$  is defined by

$$(2.3a) \quad V \in H^1(I),$$

$$(2.3b) \quad \left( \frac{d}{dx} V, \frac{d}{dx} \phi \right) = (\sigma G^\beta(V), \phi) \quad \text{for all } \phi \in H^1(I),$$

$$(2.3c) \quad c_0 < V(x) < d_1 \quad (x \in \bar{I}).$$

We now have

**THEOREM 1.** *Suppose the assumption (A2) and fix  $\beta \in J_0 \cap J_1$  arbitrarily. Then there exists a positive integer  $n_0$  depending on  $\sigma G^\beta(V)$  such that a family of periodic solutions*

$$\{V_{n,i}^\beta(x)\}_{n \geq n_0, i=0,1} \subset C^1(I)$$

of the problem (2.2) exists, where  $n$  is the mode number.<sup>2</sup> When  $G^\beta(\beta) \neq \beta$ , there is no other solutions except the above, and when  $G^\beta(\beta) = \beta$ ,  $V = \beta$  must be added to the solutions.

**Remark 2.1.**  $V_{n,i}^\beta(x)$  crosses the line  $V = \beta$  at  $n$  points  $0 < x_1^i < x_3^i < \dots < x_{2n-1}^i < 1$ .  $V_{n,1}^\beta(x)$  is an appropriate reflection version of  $V_{n,0}^\beta(x)$ . These properties become clear in the proof of Theorem 1.

Thus, we find that there exists a countably infinite number of solutions  $(U_{n,i}^\beta(x), V_{n,i}^\beta(x))$  of the problem (2.1)<sub>0</sub>, where

$$U_{n,i}^\beta(x) = \begin{cases} h_0(V_{n,i}^\beta(x)) & \text{for } V_{n,i}^\beta(x) < \beta, \\ h_1(V_{n,i}^\beta(x)) & \text{for } V_{n,i}^\beta(x) > \beta. \end{cases}$$

**Remark 2.2.** If  $J_0 \cap J_1 \neq \emptyset$ , we can show that there exist other solutions except  $(U_{n,i}^\beta(x), V_{n,i}^\beta(x))$  for the full problem (2.1)<sub>0</sub> (see [14]).

For treating the problem (2.1) <sub>$\epsilon$</sub>  with  $\epsilon \neq 0$ , we make the following assumption in addition to (A1) and (A2):

- (A3) (i)  $(\partial/\partial u)f(h_i(v), v) < 0$  in  $J_i, (i = 0, 1)$ .
- (ii) Define  $\mathcal{F}(\beta)$  by  $\mathcal{F}(\beta) = \int_{h_0(\beta)}^{h_1(\beta)} f(s, \beta) ds$  for  $\beta \in J_0 \cap J_1$ .  $\mathcal{F}(\beta)$  has a zero at  $\beta = \beta^*$  and  $(d/d\beta)\mathcal{F}(\beta^*) \neq 0$ .
- (iii) There exists a constant  $\gamma \in (h_0(\beta^*), h_1(\beta^*))$  such that

$$\int_{h_0(\beta^*)}^k f(s, \beta^*) ds < 0 \quad \text{for all } k \in (h_0(\beta^*), \gamma),$$

$$\int_{h_1(\beta^*)}^k f(s, \beta^*) ds < 0 \quad \text{for all } k \in (\gamma, h_1(\beta^*)).$$

**Remark 2.3.** It is easily checked that conditons (i) and (iii) of (A3) imply (ii) when the curve  $f(u, v) = 0$  is of S or S-like shaped (see, for example, Fig. 1).

The assumption (A3) is the one introduced by Fife [3]. Under (A1) ~ (A3), we have

**THEOREM 2.** *Suppose (A1) ~ (A3). Let  $(U(x), V(x))$  be any solution  $(U_{n,i}^{\beta^*}(x), V_{n,i}^{\beta^*}(x))$  of the reduced problem (2.2). Then there exist some positive constant  $\epsilon_0$  and  $\sigma_0$  such that for each fixed  $\sigma \in (0, \sigma_0)$  a family of solutions  $(u(x; \epsilon), v(x; \epsilon))$  of the problem (2.1) <sub>$\epsilon$</sub>  exists for  $0 < \epsilon < \epsilon_0$ , which satisfies*

$$(2.4) \quad \lim_{\epsilon \rightarrow 0} u(x; \epsilon) = U(x) \quad \text{uniformly in } x \in \bar{I} - \bigcup_{j=1}^n [x_{2j-1} - \kappa, x_{2j-1} + \kappa],$$

$$\lim_{\epsilon \rightarrow 0} v(x; \epsilon) = V(x) \quad \text{uniformly in } x \in \bar{I},$$

for any  $\kappa > 0$ , where  $x_{2j-1}$  are the points stated in Remark 2.1.

<sup>2</sup> We call a function  $u(x) \in C^1[0, 1]$  a periodic function with  $n$  ( $\geq 1$ ) mode if  $u(x)$  satisfies the following conditions: (i)  $(d/dx)u(0) = (d/dx)u(1/n) = 0$ , (ii)  $u(x)$  is monotone on  $[0, 1/n]$ , (iii)  $u(x) = u((2/n) - x)$  on  $[1/n, 2/n]$ , (iv)  $u(x + (2/n)) = u(x)$  on  $[2/n, 1]$ .

As an application, we consider the prey-predator model system exemplified in the introduction:

$$(2.5)_\epsilon \quad \begin{aligned} 0 &= \epsilon^2 \frac{d^2}{dx^2} u + \{f_0(u) - kv\}u, \\ 0 &= \frac{1}{\sigma} \frac{d^2}{dx^2} v - \{g_0(v) - ku\}v. \end{aligned}$$

We impose the following assumptions on the nonlinearities of  $f_0$  and  $g_0$  from an ecological viewpoint (see, for instance, [11], [17]).

(A4)  $f_0(u) \in C^2(\mathbb{R}^+)$  satisfies

- (i)  $f_0(0) \geq 0$  and
- (ii)  $\frac{d}{du} f_0(u) \begin{cases} > 0 & (0 \leq u < c), \\ = 0 & (u = c), \\ < 0 & (u > c), \end{cases}$

where  $c$  is a positive constant.

*Remark 2.4.* Such a nonlinearity as (ii) is usually called the Allee effect in ecology.

Then it is found that  $f(u, v) = \{f_0(u) - kv\}u = 0$  has triple roots, say  $u = h_0(v) = 0$ ,  $u = h_1(v)$  for  $u > c$  and  $u = h_2(v)$  for  $0 < u < c$  (see Fig. 1).

(A5)  $g_0(v) \in C^2(\mathbb{R}^+)$  satisfies

- (i)  $g_0(0) > 0$ ,
- (ii)  $\frac{d}{dv} g_0(v) \geq 0 \quad (v \geq 0)$  and
- (iii)  $\frac{d}{dv} G_1(v) < 0 \quad \left( v \in \left( \beta^*, \frac{f_0(c)}{k} \right) \right)$ ,

where  $G_1(v) = \{g_0(v) - kh_1(v)\}v$  and  $\beta^* \in (f_0(0)/k, f_0(c)/k)$  is given by

$$\int_0^{h_1(\beta^*)} \{f_0(u) - k\beta^*\}u \, du = 0.$$

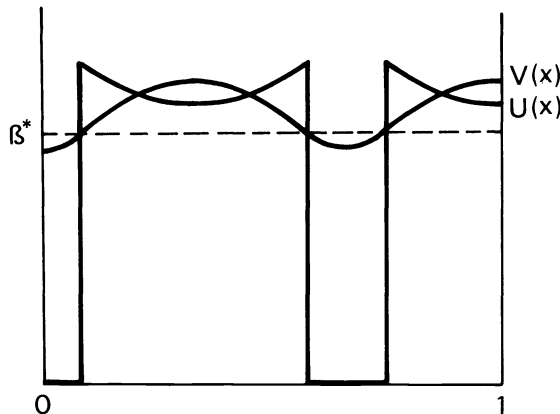


FIG. 2. Spatial patterns of  $(U(x), V(x))$  of  $(2.5)_0$  where the mode number is 3.



Thus we have

**THEOREM 3.** *Under (A4) and (A5), there exist some positive constant  $\epsilon_0$  and  $\sigma_0$  such that for each fixed  $\sigma \in (0, \sigma_0)$  a family of nonnegative solutions  $\{(u(x; \epsilon), v(x; \epsilon))\}_{\epsilon \in (0, \epsilon_0)}$  exists, which satisfies the same limiting process as (2.4) in Theorem 2.*

From this theorem, it turns out that the habitat of the prey and its predator is composed of two quantitatively different regions, in one region the prey is alive ( $u = h_1(v) > 0$ ) and in the other it is dead ( $u = h_0(v) \equiv 0$ ) (see Figs. 2 and 3).

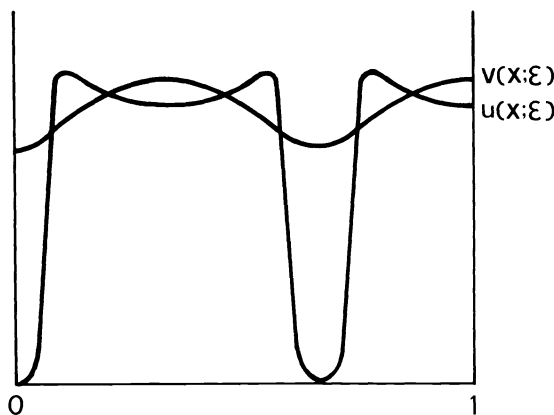


FIG. 3. Spatial patterns of  $(u(x; \epsilon), v(x; \epsilon))$  of  $(2.5)_\epsilon$  with sufficiently small  $\epsilon > 0$ .

**3. Proof of Theorem 1.** Here we give the proof of Theorem 1. We begin by considering an initial value problem

$$\begin{aligned}
 (3.1) \quad & \frac{d^2}{dx^2} V = -\sigma \tilde{G}_0(V) \quad (x > 0), \\
 & \frac{d}{dx} V(0) = 0, \\
 & V(0) = k,
 \end{aligned}$$

where  $k \in (c_0, d_0)$  and  $\tilde{G}_0 \in C^1(c_0, +\infty)$  is an extension of  $G_0 \in C^1(c_0, d_0)$  in a way that  $\tilde{G}'_0 (= (d/dV)\tilde{G}_0) < 0$  in  $(c_0, +\infty)$ . By the general theory of ordinary differential equations, we obtain the following result.

**LEMMA 3.1.** (i) *The problem (3.1) has a unique solution  $V(x; k)$  which is strictly increasing in  $x$  and satisfies that  $V(x; k) \uparrow +\infty$  as  $x \uparrow +\infty$ . The derivatives,  $(\partial/\partial x)V$ ,  $(\partial/\partial k)V$ ,  $(\partial^2/\partial x \partial k)V$  and  $(\partial^2/\partial k \partial x)V$  are continuous in  $(0, +\infty) \times (c_0, d_0)$ .*

(ii)  $\xi(x; k) = (\partial/\partial x)V(x; k)$  is a solution of the problem

$$\begin{aligned}
 (3.2) \quad & \frac{d^2}{dx^2} \xi = -\sigma \tilde{G}'_0(V(x; k))\xi \quad (x > 0), \\
 & \frac{d}{dx} \xi(0) = -\sigma \tilde{G}_0(k), \\
 & \xi(0) = 0,
 \end{aligned}$$

and  $\xi(x; k) > 0$  for all  $(x, k) \in (0, +\infty) \times (c_0, d_0)$ .

(iii)  $\eta(x; k) = (\partial/\partial k)V(x; k)$  is a solution of the problem

$$(3.3) \quad \begin{aligned} \frac{d^2}{dx^2} \eta &= -\sigma \tilde{G}'_0(V(x; k))\eta \quad (x > 0), \\ \frac{d}{dx} \eta(0) &= 0, \\ \eta(0) &= 1, \end{aligned}$$

and  $\eta(x; k) > 1$  for all  $(x, k) \in (0, +\infty) \times (c_0, d_0)$ .

We omit the proof of Lemma 3.1 since it is easily obtained by using the property of  $\tilde{G}_0$ .

Now we define a mapping  $l$  from  $D = \{(k, \beta); c_0 < k < \beta < d_0\}$  into  $R^1$  by  $V(l; k) = \beta$ . The mapping  $l$  is well-defined since  $V(0; k) < \beta$  and  $V(x; k) \uparrow +\infty$  as  $x \uparrow +\infty$  by (iii) of Lemma 3.1. Letting  $\bar{l}_0(\beta) = \limsup_{k \downarrow c_0} l(k, \beta)$ , we have

LEMMA 3.2. For every fixed  $\beta \in (c_0, d_0)$ ,  $l$  is a diffeomorphism of class  $C^1$  of  $(c_0, \beta)$  onto  $(0, \bar{l}_0(\beta))$  and satisfies

$$(3.4) \quad \frac{\partial}{\partial k} l(k, \beta) = -\frac{\eta(l(k, \beta); k)}{\xi(l(k, \beta); k)}.$$

*Proof.* By differentiating  $V(l(k, \beta); k) = \beta$  with respect to  $k$ , (3.4) is obtained. Since the right-hand side is negative by Lemma 3.1, we find that  $l$  is monotone decreasing in  $k$ . This property yields the diffeomorphism of  $l$ . Q.E.D.

For every fixed  $\beta \in (c_0, d_0)$ , we denote the inverse of  $l(k, \beta)$  by  $k(l, \beta)$ . Obviously it holds that

$$(3.5) \quad \frac{\partial}{\partial l} k(l, \beta) = -\frac{\xi(l; k(l, \beta))}{\eta(l; k(l, \beta))},$$

$$(3.6) \quad \frac{\partial}{\partial \beta} k(l, \beta) = -\frac{\partial k}{\partial l} \frac{\partial l}{\partial \beta} = \frac{1}{\eta(l(k, \beta); k)}.$$

LEMMA 3.3. Fix  $\beta \in (c_0, d_0)$  and  $l \in (0, \bar{l}_0(\beta))$  arbitrarily. Then,  $V_0(x; l, \beta) = V(x; k(l, \beta))$  is a unique solution of

$$(3.7) \quad \begin{aligned} \frac{d^2}{dx^2} V &= -\sigma G_0(V) \quad (0 < x < l), \\ \frac{d}{dx} V(0) &= 0, \\ V(l) &= \beta, \\ c_0 < V(x) &\leq \beta \quad (0 \leq x \leq l). \end{aligned}$$

And,  $V_0$  and  $\frac{\partial}{\partial x} V_0$  are continuously differentiable with respect to  $l$  and  $\beta$ .

*Proof.* It is easy to check that  $V_0$  is a unique solution of (3.7). We prove that  $(\partial/\partial x)V_0$  is continuously differentiable with respect to  $\beta$ . By definition, it holds that

$$\frac{\partial}{\partial x} V_0(x; l, \beta) = \frac{\partial}{\partial x} V(x; k(l, \beta)) = \xi(x; k(l, \beta)).$$

Differentiating both sides by  $\beta$ , we have from (3.6)

$$\frac{\partial^2}{\partial\beta\partial x} V_0(x; l, \beta) = \frac{\partial\xi}{\partial k} \frac{\partial k}{\partial\beta} = \frac{1}{\eta} \frac{\partial\xi}{\partial k}(x; k(l, \beta))$$

which implies that  $(\partial^2/\partial\beta\partial x)V_0$  is continuous. The others are obtained similarly. Q.E.D.

Hitherto we considered about  $G_0$ . Repeating the same argument about  $G_1$ , we obtain the following result, where  $\bar{l}_1$  is the counterpart of  $\bar{l}_0$ .

LEMMA 3.4. Fix  $\beta \in (c_1, d_1)$  and  $l \in (0, \bar{l}_1(\beta))$  arbitrarily. Then, the problem

$$\begin{aligned} \frac{d^2}{dx^2} V &= -\sigma G_1(V) & (0 < x < l), \\ \frac{d}{dx} V(0) &= 0, \\ V(l) &= \beta, \\ \beta &\leq V(x) < d_1 & (0 \leq x \leq l), \end{aligned} \tag{3.8}$$

has a unique solution  $V_1(x; l, \beta)$ .  $V_1$  and  $(\partial/\partial x)V_1$  are continuously differentiable with respect to  $l$  and  $\beta$ .

LEMMA 3.5. Fix  $\beta \in J_0 \cap J_1$  arbitrarily. Then  $\psi_i(l) = (\partial/\partial x)V_i(l; l, \beta)$ ,  $(i = 0, 1)$ , satisfy

$$0 < \psi_0(l) \leq -\sigma G_0(\beta - 0)l \quad (0 < l < \bar{l}_0(\beta)), \tag{3.9}$$

$$-\sigma G_1(\beta + 0)l \leq \psi_1(l) < 0 \quad (0 < l < \bar{l}_1(\beta)), \tag{3.10}$$

$$(-1)^i \frac{d}{dl} \psi_i(l) > 0 \quad (0 < l < \bar{l}_i(\beta), i = 0, 1). \tag{3.11}$$

*Proof.* We prove the results only for  $i = 0$ . By definition we have  $\psi_0(l) = (\partial/\partial x)V_0(l; k(l, \beta))$ . Integrating the first equation of (3.7) from  $x = 0$  to  $l$ , we obtain (3.9) immediately.

Now we prove (3.11). Noting  $\psi_0(l) = \xi(l; k(l, \beta))$  and using (3.5) and Lemma 3.1, we have

$$\begin{aligned} \frac{d}{dl} \psi_0(l) &= \frac{\partial\xi}{\partial x} + \frac{\partial\xi}{\partial k} \frac{\partial k}{\partial l} \\ &= \frac{\partial\xi}{\partial x} - \frac{\partial\xi}{\partial k} \frac{\xi}{\eta} \\ &= \frac{1}{\eta} \left[ \frac{\partial\xi}{\partial x} \eta - \frac{\partial\eta}{\partial x} \xi \right] (l, k(l, \beta)) \\ &= \frac{1}{\eta(l, k(l, \beta))} \left[ \frac{\partial\xi}{\partial x} \eta - \frac{\partial\eta}{\partial x} \xi \right] (0, k(l, \beta)) \\ &= \frac{\sigma G_0(k(l, \beta))}{\eta(l, k(l, \beta))} \\ &> 0. \end{aligned}$$

Here we used the fact that the Wronskian  $(\partial\xi/\partial x)\eta - (\partial\eta/\partial x)\xi$  is constant. Q.E.D.

*Proof of Theorem 1.* Set  $\bar{\psi}_i(\beta) = \lim_{l \rightarrow \bar{l}_i(\beta)} \psi_i(l)$ , ( $i = 0, 1$ ). Lemma 3.5 implies that  $\psi_0$  is a diffeomorphism of  $(0, \bar{l}_0(\beta))$  onto  $(0, \psi_0(\beta))$  and that  $\psi_1$  is also a diffeomorphism of  $(0, \bar{l}_1(\beta))$  onto  $(\bar{\psi}_1(\beta), 0)$ . By putting  $s(\alpha; \beta) = \psi_0^{-1}(\alpha) + \psi_1^{-1}(-\alpha)$ ,  $\bar{\alpha}(\beta) = \min(\bar{\psi}_0(\beta), -\bar{\psi}_1(\beta))$  and  $\bar{l}(\beta) = \lim_{\alpha \rightarrow \alpha(\beta)} s(\alpha; \beta)$ , the function  $s$  is obviously a homeomorphism of  $(0, \bar{\alpha}(\beta))$  onto  $(0, \bar{l}(\beta))$ . Let  $n_0(\beta)$  be the smallest positive integer greater than  $1/\bar{l}(\beta)$ . Choosing  $\alpha_n$  as  $s(\alpha_n; \beta) = 1/n$ , ( $n \geq n_0(\beta)$ ), we construct periodic functions  $V_{n,i}^\beta \in C^1[0, 1]$ , ( $i = 0, 1$ ) with period  $2s(\alpha_n)$  by

$$(3.12) \quad \begin{aligned} &V_i(x; \psi_i^{-1}((-1)^i \alpha_n), \beta) \quad (0 \leq x \leq \psi_i^{-1}((-1)^i \alpha_n)), \\ &V_{1+i}(s(\alpha_n) - x; \psi_{i+1}^{-1}((-1)^{i+1} \alpha_n), \beta) \quad (\psi_i^{-1}((-1)^i \alpha_n) < x \leq s(\alpha_n)), \\ &V_{n,i}^\beta(2s(\alpha_n) - x) \quad (s(\alpha_n) < x \leq 2s(\alpha_n)), \end{aligned}$$

where  $V_2 = V_0$  and  $\psi_2^{-1} = \psi_0^{-1}$ . It is not difficult to observe that  $V_{n,i}^\beta(x)$  satisfy (2.3) for  $i = 0, 1$  and  $n \geq n_0(\beta)$ . To complete the proof of Theorem 1, it suffices to show that there exist no other solutions of (2.3). Let  $V(\neq \beta)$  be a solution of (2.3). Substituting  $\phi = 1$  into (2.3b), we have

$$(3.13) \quad \int_0^1 G^\beta(V(x)) dx = 0.$$

We first show that there exists a point  $a_0 \in (0, 1)$  such that

$$(3.14) \quad V(a_0) = \beta \quad \text{and} \quad \frac{d}{dx} V(a_0) \neq 0.$$

Taking a point  $z_0$  such that  $V(z_0) \neq \beta$ , we let  $a_0$  be the nearest point to  $z_0$  satisfying  $V(a_0) = \beta$ . Such a point  $a_0$  is well-defined since the closed set  $\{x; x \in [0, 1], V(x) = \beta\}$  is not empty by (3.13). Without loss of generality, we may assume that

$$a_0 < z_0 \quad \text{and} \quad V(z_0) < V(a_0) (= \beta).$$

From (2.3a) and (2.3b) we observe  $V$  satisfies (2.2a) in  $(a_0, z_0)$ . Choose a point  $y_0 \in (a_0, z_0)$  satisfying  $(d/dx)V(y_0) < 0$ . Integrating (2.2a) from  $a_0$  to  $y_0$ , we have

$$\begin{aligned} \frac{d}{dx} V(a_0) &= \frac{d}{dx} V(y_0) + \sigma \int_{a_0}^{y_0} G^\beta(V(s)) ds \\ &\leq \frac{d}{dx} V(y_0) \\ &< 0. \end{aligned}$$

Hence  $a_0$  satisfies (3.14). Now we set  $\alpha = -(d/dx)V(a_0) > 0$ . While  $V$  is lying in  $(c_0, \beta)$ ,  $V$  satisfies (2.2a). There  $V$  can be extended until the graph  $(x, V(x))$  reaches  $V = \beta$  or  $x = 1$ . In the former case there exists a point  $a_1 (= a_0 + 2\psi_0^{-1}(\alpha)) \in (0, 1)$  where it holds that

$$V(a_1) = \beta \quad \text{and} \quad \frac{d}{dx} V(a_1) = \alpha.$$

Since  $V \in C^1(0, 1)$  and  $\alpha > 0$ ,  $V(x)$  traverses the line  $V = \beta$ . While  $V$  is lying in  $(\beta, d_1)$ ,  $V$  satisfies (2.2a). Hence  $V$  can be extended until the graph  $(x, V(x))$  reaches  $V = \beta$  or  $z = 1$ . In the former case there exists a point  $a_2 (= a_1 + 2\psi_1^{-1}(-\alpha)) \in (0, 1)$ , where it holds that

$$V(a_2) = \beta \quad \text{and} \quad \frac{d}{dx} V(a_2) = -\alpha.$$

Repeating this process on both sides of  $a_0$ , and noting the boundary condition, we find that  $\alpha$  must be equal to some  $\alpha_n$  and that  $V = V_{n,0}^\beta$  or  $V_{n,1}^\beta$ . This completes the proof on Theorem 1.

*Remark 3.6.* The estimate (3.11) of Lemma 3.5 plays an important role in proving Theorem 2. To obtain this estimate we assumed that  $G_i$  are continuously differentiable. Theorem 1, however, can be proved under the weaker condition that  $G_i$  are Lipschitz continuous (cf. [19]).

**4. Proof of Theorem 2.** We construct solutions of the problem  $(2.1)_\epsilon$  with the aid of solutions of the reduced problem (2.2) obtained in the previous section. The method used here is almost the same as the one for the Dirichlet boundary conditions studied by P. C. Fife [3]. Therefore, we only state the outline.

We choose a solution  $(U_{n,i}^{\beta^*}(x), V_{n,i}^{\beta^*}(x))$  of the problem  $(2.1)_0$  for an arbitrarily fixed mode number  $n$ . Here, omitting the superscript and the subscripts, we denote it simply by  $(U(x), V(x))$ . For simplicity we consider the problem  $(2.1)_\epsilon$  only on the subinterval  $[x_0(=0), x_2]$  such that  $V(x)$  satisfies  $c_0 < V(x) < \beta^*$  on  $[x_0, x_1]$ ,  $\beta^* < V(x) < d_1$  on  $(x_1, x_2]$ ,  $(d/dx)V(x_0) = 0$ ,  $(d/dx)V(x_2) = 0$ , and  $V(x_1) = \beta^*$ . Let  $V_0(x; \delta, \omega)$  be a solution of the boundary value problem

$$\begin{aligned}
 (4.1) \quad & \frac{d^2}{dx^2} V + \sigma G_0(V) = 0 \quad (x \in (x_0, x_1 + \delta)), \\
 & \frac{d}{dx} V(x_0) = 0, \\
 & V(x_1 + \delta) = \beta^* + \omega,
 \end{aligned}$$

where  $\delta$  and  $\omega$  are small parameters to be determined later. Also, let  $V_1(x; \delta, \omega)$  be a solution of

$$\begin{aligned}
 (4.2) \quad & \frac{d^2}{dx^2} V + \sigma G_1(V) = 0 \quad (x \in (x_1 + \delta, x_2)), \\
 & \frac{d}{dx} V(x_2) = 0, \\
 & V(x_1 + \delta) = \beta^* + \omega.
 \end{aligned}$$

**LEMMA 4.1.** *Let  $V(x)$  be defined as above. Under (A2), there exist positive constants  $\delta_0$  and  $\omega_0$  such that, for all  $|\delta| < \delta_0$  and  $|\omega| < \omega_0$ , (4.1) (resp. (4.2)) has a unique monotone solution  $V_0(x; \delta, \omega)$  (resp.  $V_1(x; \delta, \omega)$ ) which satisfies*

- (i)  $(\partial/\partial x)V_i(x; \delta, \omega)$  is continuous uniformly in  $\delta$  and  $\omega$  ( $i = 0, 1$ ),
- (ii)  $\|V(x) - V_0(x; \delta, \omega)\|_{C^1[x_0, x_1 + \delta]} + \|V(x) - V_1(x; \delta, \omega)\|_{C^1[x_1 + \delta, x_2]} \rightarrow 0$  as  $\delta$  and  $\omega$  tend to zero, and

$$(iii) \quad \frac{d}{d\delta} \left( \frac{\partial}{\partial x} V_0(x_1 + \delta; \delta, \omega) \right) > 0, \quad (|\delta| < \delta_0, |\omega| < \omega_0).$$

$$\frac{d}{d\delta} \left( \frac{\partial}{\partial x} V_1(x_1 + \delta; \delta, \omega) \right) < 0,$$

This lemma is a direct consequence of Lemma 3.5, so we omit the proof. Here, we consider two problems:

$$(4.3a) \quad \begin{aligned} \varepsilon^2 \frac{d^2}{dx^2} u_0 + f(u_0, v_0) &= 0, \\ \frac{d^2}{dx^2} v_0 + \sigma g(u_0, v_0) &= 0, \end{aligned} \quad (x \in (x_0, x_1 + \delta)),$$

with the boundary conditions

$$(4.3b) \quad \begin{aligned} \frac{d}{dx} u_0(x_0) &= 0, & u_0(x_1 + \delta) &= \gamma^*, \\ \frac{d}{dx} v_0(x_0) &= 0, & v_0(x_1 + \delta) &= \beta^* + \omega, \end{aligned}$$

and

$$(4.4a) \quad \begin{aligned} \varepsilon^2 \frac{d^2}{dx^2} u_1 + f(u_1, v_1) &= 0, \\ \frac{d^2}{dx^2} v_1 + \sigma g(u_1, v_1) &= 0, \end{aligned} \quad (x \in (x_1 + \delta, x_2)),$$

with

$$(4.4b) \quad \begin{aligned} \frac{d}{dx} u_1(x_2) &= 0, & u_1(x_1 + \delta) &= \gamma^*, \\ \frac{d}{dx} v_1(x_2) &= 0, & v_1(x_1 + \delta) &= \beta^* + \omega, \end{aligned}$$

where  $\gamma^* = \frac{1}{2}(h_0(\beta^*) + h_1(\beta^*))$ .

To solve these problems, we first consider the boundary layer equations derived from the first of (4.3a) and (4.4a)

$$(4.5a) \quad \frac{d^2}{d\eta^2} \hat{z}_i + f(h_i(\beta) + \hat{z}_i, \beta) = 0 \quad (0 < \eta < +\infty, i = 0, 1).$$

The boundary conditions are assumed to be

$$(4.5b) \quad \hat{z}_i(0) = \gamma - h_i(\beta) \quad \text{and} \quad \hat{z}_i(\pm\infty) = 0,$$

where  $\gamma = \gamma(\beta) = \frac{1}{2}(h_0(\beta) + h_1(\beta))$ .

LEMMA 4.2 (Fife [3]). *Consider the problem (4.5) under (i) and (iii) of (A3); there exists a positive constant  $\omega_1$  such that for all  $|\beta - \beta^*| < \omega_1$  (4.5) has a unique monotone solution  $\hat{z}_i(\eta; \beta)$  satisfying*

$$(4.6) \quad |\hat{z}_i(\eta; \beta)|, \quad \left| \frac{d}{d\eta} \hat{z}_i(\eta; \beta) \right| \leq C e^{-\kappa\eta}$$

for some positive constants  $\kappa$  and  $C$  independent of  $\beta$ .

Let  $|\delta| < \delta_0$  and  $|\omega| \leq \min(\omega_0, \omega_1)$ . Then, by Lemma 4.2 we obtain a correction term  $z_i(x; \varepsilon, \delta, \omega)$  to  $h_i(V_i(x; \delta, \omega))$ .  $z_0(x; \varepsilon, \delta, \omega)$  is constructed by  $\zeta((x_1 + \delta - x)/(x_1 +$

$\delta) \hat{z}_0((x_1 + \delta - x)/\varepsilon, \beta^* + \omega)$ , where  $\zeta(x)$  is a  $C^\infty$ -cutoff function defined by

$$\zeta(x) = \begin{cases} 1 & (x \in [0, \frac{1}{4}]), \\ 0 & (x \in [\frac{1}{2}, 1]). \end{cases}$$

$z_1(x; \varepsilon, \delta, \omega)$  is also constructed in a similar way. We set

$$(4.7) \quad U_i(x; \varepsilon, \delta, \omega) = h_i(V_i(x; \delta, \omega)) + z_i(x; \varepsilon, \delta, \omega) \quad (i = 0, 1).$$

Hereafter we restrict our argument to the case only for  $i = 0$ , so we omit the subscript  $i$ . Let us seek a solution  $(u(x; \varepsilon, \delta, \omega), v(x; \varepsilon, \delta, \omega))$  of the problem (4.3) which takes the form

$$(4.8) \quad \begin{aligned} u(x; \varepsilon, \delta, \omega) &= U(x; \varepsilon, \delta, \omega) + r(x; \varepsilon, \delta, \omega), \\ v(x; \varepsilon, \delta, \omega) &= V(x; \delta, \omega) + s(x; \varepsilon, \delta, \omega), \end{aligned}$$

where  $(r, s)$  is an unknown remainder. Since  $(U, V)$  satisfies the boundary condition (4.3b), the solvability of the problem (4.3) is reduced to finding a solution  $t(\varepsilon) = (r(\varepsilon), s(\varepsilon)) \in X_\varepsilon = (C^2_{\varepsilon_0} \times (H^2 \cap C^1_0))[0, x_1]$  of

$$(4.9) \quad T(t; \varepsilon; \delta, \omega, \sigma) = (R(r, s; \varepsilon; \delta, \omega, \sigma), S(r, s; \varepsilon; \delta, \omega, \sigma)) = 0,$$

where

$$(4.10) \quad \begin{aligned} R(r, s; \varepsilon; \delta, \omega, \sigma) &= \varepsilon^2 \frac{d^2}{dx^2} (U + r) + p(\delta)f(U + r, V + s), \\ S(r, s; \varepsilon; \delta, \omega, \sigma) &= \frac{d^2}{dx^2} (V + s) + \sigma p(\delta)g(U + r, V + s) \end{aligned}$$

for  $p(\delta) = (1 + (\delta/x_1))^2$ . Here, we transformed  $x$  into  $(x_1/(x_1 + \delta))x$ . We notice that  $U(x; \varepsilon, \delta, \omega)$  and  $V(x; \delta, \omega)$  also depend on a parameter  $\sigma$ .

LEMMA 4.3. *Let  $\sigma$  be a sufficiently small positive number. Under (A1) ~ (A3), there exist positive constants  $\delta_0, \omega_0$  and  $\varepsilon_0$  such that, for each  $|\delta| < \delta_0, |\omega| < \omega_0$ , and  $0 < \varepsilon < \varepsilon_0$ , it follows:*

(i)  $T(t; \varepsilon; \delta, \omega)$  is a continuously differentiable mapping from  $X_\varepsilon$  into  $Y = C^0[0, x_1] \times L^2(0, x_1)$ . Moreover, there exists a constant  $K_1$  independent of  $\varepsilon, \delta$  and  $\omega$  such that

$$\|T_r(t_2; \varepsilon; \delta, \omega) - T_r(t_1; \varepsilon; \delta, \omega)\|_{\mathcal{L}(X_\varepsilon, Y)} \leq K_1 \|t_2 - t_1\|_{X_\varepsilon}$$

for  $t_1, t_2 \in X_\varepsilon$ , where  $T_r$  is the Fréchet derivative of  $T$ ,

(ii)  $T_r(0; \varepsilon; \delta, \omega)$  has an inverse satisfying

$$\|T_r^{-1}(0; \varepsilon; \delta, \omega)\|_{\mathcal{L}(Y, X_\varepsilon)} \leq K_2,$$

where  $K_2$  is a constant independent of  $\varepsilon, \delta$  and  $\omega$ , and

(iii)  $\|T(0; \varepsilon; \delta, \omega)\|_Y \rightarrow 0, (\varepsilon \rightarrow 0)$  uniformly in  $\delta$  and  $\omega$ .

*Proof.* Noting that every function  $t(x)$  of (4.10) is extended to  $\tilde{t}(x)$  by reflection, namely

$$\tilde{t}(x) = \begin{cases} t(x) & (x \in [0, x_1]), \\ t(-x) & (x \in [-x_1, 0]), \end{cases}$$

which is also twice continuously differentiable; we find that the problem (4.9) is reduced to the zero Dirichlet boundary value problem. Then, the statement (i) and (ii) are verified by similar arguments to [3, Lemma 3.2].

For the proof of (iii), we first show that

$$\|T(0; \varepsilon; 0, 0)\|_Y \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

From (4.10) it holds that

$$\begin{aligned} R(0, 0; \varepsilon; 0, 0) &= \varepsilon^2 \frac{d^2}{dx^2} U(\varepsilon) + f(U(\varepsilon), V) \\ &= \varepsilon^2 \frac{d^2}{dx^2} h(V(x)) + \varepsilon^2 \frac{d^2}{dx^2} z(x; \varepsilon) + f(h(V(x)) + z(x; \varepsilon), V(x)), \end{aligned}$$

and

$$\begin{aligned} S(0, 0; \varepsilon; 0, 0) &= \frac{d^2}{dx^2} V + \sigma g(U(\varepsilon), V) \\ &= \frac{d^2}{dx^2} V(x) + \sigma g(h(V(x)) + z(x; \varepsilon), V(x)). \end{aligned}$$

We divide  $[0, x_1]$  into two intervals  $[0, x_1 - \sqrt{\varepsilon}]$  and  $[x_1 - \sqrt{\varepsilon}, x_1]$ . On the first interval we have

$$\begin{aligned} R(0, 0; \varepsilon; 0, 0) &= O(\varepsilon^2) + O(\varepsilon) - \zeta \left( \frac{x_1 - x}{x_1 - x_0} \right) f(h(\beta^*) + z(x; \varepsilon), \beta^*) \\ &\quad + f(h(V(x)) + z(x; \varepsilon), V(x)) \\ &\rightarrow 0 \quad \text{uniformly on } [0, x_1 - \sqrt{\varepsilon}], \quad (\varepsilon \rightarrow 0), \end{aligned}$$

and

$$\begin{aligned} S(0, 0; \varepsilon; 0, 0) &= -\sigma g(h(V(x)), V(x)) + \sigma g(h(V(x)) + z(x; \varepsilon), V(x)) \\ &\rightarrow 0 \quad \text{uniformly on } [0, x_1 - \sqrt{\varepsilon}], \quad (\varepsilon \rightarrow 0), \end{aligned}$$

since  $z(x; \varepsilon)$  converge to 0 uniformly on  $[0, x_1 - \sqrt{\varepsilon}]$  as  $\varepsilon \rightarrow 0$  by (4.6). On the second interval we have

$$\begin{aligned} R(0, 0; \varepsilon; 0, 0) &= O(\varepsilon^2) - f(h(\beta^*) + z(x; \varepsilon), \beta^*) + f(h(V(x)) + z(x; \varepsilon), V(x)) \\ &= O(\varepsilon^2) + \left\{ \frac{\partial}{\partial u} f(h(\theta) + z(x; \varepsilon), \theta) \frac{d}{dV} h(\theta) + \frac{\partial}{\partial v} f(h(\theta) + z(x; \varepsilon), \theta) \right\} \\ &\quad \times (V(x) - \beta^*) \\ &\rightarrow 0 \quad \text{uniformly on } [x_1 - \sqrt{\varepsilon}, x_1], \quad (\varepsilon \rightarrow 0), \end{aligned}$$

where  $\theta$  is an intermediate value between  $\beta^*$  and  $V(x)$ . Since  $S(0, 0; \varepsilon; 0, 0)$  are bounded uniformly in  $\varepsilon$ , we have

$$\int_{x_1 - \sqrt{\varepsilon}}^{x_1} |S(0, 0; \varepsilon; 0, 0)|^2(x) dx \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

Thus we obtain

$$\|R(0, 0; \varepsilon; 0, 0)\|_{C^0[x_0, x_1]}, \|S(0, 0; \varepsilon; 0, 0)\|_{L^2(x_0, x_1)} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

A slight modification of the above arguments leads to (iii). Q.E.D.



By the above lemma, we have the following

LEMMA 4.4. Under (A1) ~ (A3), there exist some positive constants  $\epsilon_0$  and  $\sigma_0$  such that the problem (4.3) has a solution  $(u_0, v_0)$  for  $0 < \epsilon < \epsilon_0$  and  $0 < \sigma < \sigma_0$ , which satisfies

- (i)  $\|u_0(\epsilon, \delta, \omega) - h_0(V_0)\|_{C^0[0, x_1 + \delta - \kappa']} \rightarrow 0, (\epsilon \rightarrow 0),$
- (ii)  $\|v_0(\epsilon, \delta, \omega) - V_0\|_{C^1[0, x_1 + \delta]} \rightarrow 0, (\epsilon \rightarrow 0),$
- (iii)  $\lim_{\epsilon \rightarrow 0} (\epsilon (d/dx)u_0(x_1 + \delta; \epsilon, \delta, \omega))^2 = -2 \int_{h_0(\beta^* + \omega)}^{\gamma(\beta^* + \omega)} f(s, \beta^* + \omega) ds,$

uniformly for  $\delta$  and  $\omega$  small enough, where  $\kappa'$  is an arbitrarily fixed positive constant. Similar results hold for the solution  $(u_1, v_1)$  of (4.4).

*Proof.* Consider the nonlinear operator  $T$  from  $X_\epsilon$  into  $Y$ . Then, by Lemma 4.3, we can apply to the equation  $T(t; \epsilon; \delta, \omega) = 0$  the implicit function theorem [3, Thm. 3.4]. Thus, we find the existence of a solution  $(r(\epsilon; \delta, \omega), s(\epsilon; \delta, \omega)) \in X_\epsilon$  of (4.9) such that  $\|r(\epsilon; \delta, \omega)\|_{C^2_0} + \|s(\epsilon; \delta, \omega)\|_{H^2} \rightarrow 0$  as  $\epsilon \rightarrow 0$ . From this result, (i) and (ii) are proved at the same time. Condition (iii) is easily obtained by integrating the equation (4.5). Q.E.D.

Thus, we find that a solution of the problem (4.3) (resp. (4.4)) exists for  $[0, x_1 + \delta]$  (resp.  $[x_1 + \delta, x_2]$ ). Here, we must notice that these two solutions  $(u_0, v_0)$  and  $(u_1, v_1)$  do not match at  $x = x_1 + \delta$  in the  $C^1$ -sense. Therefore, in order to complete the proof of Theorem 2, we determine two parameters  $\delta$  and  $\omega$  depending on  $\epsilon$  so that  $(d/dx)u_0(x_1 + \delta; \epsilon, \delta, \omega) = (d/dx)u_1(x_1 + \delta; \epsilon, \delta, \omega)$  and  $(d/dx)v_0(x_1 + \delta; \epsilon, \delta, \omega) = (d/dx)v_1(x_1 + \delta; \epsilon, \delta, \omega)$ .

Following the arguments in the proof of [3, Thm. 4.1], we define  $\Phi(\epsilon, \delta, \omega)$  and  $\Psi(\epsilon, \delta, \omega)$  by

$$\begin{aligned} \Phi(\epsilon, \delta, \omega) &= \left( \epsilon \frac{d}{dx} u_1(x_1 + \delta; \epsilon, \delta, \omega) \right)^2 - \left( \epsilon \frac{d}{dx} u_0(x_1 + \delta; \epsilon, \delta, \omega) \right)^2, \\ \Psi(\epsilon, \delta, \omega) &= \frac{d}{dx} v_1(x_1 + \delta; \epsilon, \delta, \omega) - \frac{d}{dx} v_0(x_1 + \delta; \epsilon, \delta, \omega). \end{aligned} \tag{4.11}$$

Noting that  $(\epsilon (d/dx)u_i)^2$  and  $(d/dx)v_i, (i = 0, 1)$  are uniformly continuous in  $\epsilon, \delta$  and  $\omega$ , we can extend them continuously to be defined for  $\epsilon \rightarrow 0$ . Setting  $\epsilon = 0$  in (4.11), we have

$$\begin{aligned} \Phi(0, \delta, \omega) &= 2 \int_{h_0(\beta^* + \omega)}^{\gamma(\beta^* + \omega)} f(s, \beta^* + \omega) ds - 2 \int_{h_1(\beta^* + \omega)}^{\gamma(\beta^* + \omega)} f(s, \beta^* + \omega) ds = 2 \mathcal{J}(\beta^* + \omega), \\ \Psi(0, \delta, \omega) &= \frac{d}{dx} V_1(x_1 + \delta; \delta, \omega) - \frac{d}{dx} V_0(x_1 + \delta; \delta, \omega), \end{aligned}$$

and, from (ii) of (A3) and Lemma 4.1, we know that

$$\begin{aligned} \Phi(0, 0, 0) &= \Psi(0, 0, 0) = 0, \\ \frac{\partial}{\partial \omega} \Phi(0, 0, \omega) \Big|_{\omega=0} &= 2 \frac{d}{d\beta} \mathcal{J}(\beta^*) \neq 0 \quad \text{and} \\ \frac{\partial}{\partial \delta} \Psi(0, \delta, 0) \Big|_{\delta=0} &= \frac{d}{d\delta} \left( \frac{\partial}{\partial x} V_1(x_1; 0, 0) \right) - \frac{d}{d\delta} \left( \frac{\partial}{\partial x} V_0(x_1; 0, 0) \right) \neq 0. \end{aligned} \tag{4.12}$$

Therefore, it is shown that  $\Phi(0, 0, \omega)$  and  $\Psi(0, \delta, 0)$  have an isolated zero at  $\omega = 0$  and  $\delta = 0$ , respectively. Moreover, it is easily found that

$$\Phi(0, \delta, \omega) \equiv \Phi(0, 0, \omega). \tag{4.13}$$

In view of (4.12) and (4.13), we can apply another implicit function theorem (see [3, Thm. 4.3]) to  $\Phi = \Psi = 0$ . We conclude that for sufficiently small  $\epsilon (> 0)$  there exist  $\delta(\epsilon)$

and  $\omega(\varepsilon)$  such that

$$\Phi(\varepsilon, \delta(\varepsilon), \omega(\varepsilon)) = \Psi(\varepsilon, \delta(\varepsilon), \omega(\varepsilon)) = 0,$$

and

$$\lim_{\varepsilon \rightarrow 0} \delta(\varepsilon) = \lim_{\varepsilon \rightarrow 0} \omega(\varepsilon) = 0.$$

Now, defining  $(u(x; \varepsilon), v(x; \varepsilon))$  by

$$u(x; \varepsilon) = \begin{cases} u_0(x; \varepsilon, \delta(\varepsilon), \omega(\varepsilon)) & (x \in [x_0, x_1 + \delta(\varepsilon)]), \\ u_1(x; \varepsilon, \delta(\varepsilon), \omega(\varepsilon)) & (x \in [x_1 + \delta(\varepsilon), x_2]) \end{cases}$$

and

$$v(x; \varepsilon) = \begin{cases} v_0(x; \varepsilon, \delta(\varepsilon), \omega(\varepsilon)) & (x \in [x_0, x_1 + \delta(\varepsilon)]), \\ v_1(x; \varepsilon, \delta(\varepsilon), \omega(\varepsilon)) & (x \in [x_1 + \delta(\varepsilon), x_2]) \end{cases}$$

we find that it matches at  $x = x_1 + \delta(\varepsilon)$  in the  $C^1$ -sense, which leads to a solution of (2.1). This completes the proof of Theorem 2.

**5. Proof of Theorem 3.** In order to prove the existence of solutions of the problem (2.5) under zero flux boundary conditions, except their nonnegativity, it suffices to show that the nonlinearities of  $f$  and  $g$  satisfy the assumptions (A1) ~ (A3).

We consider two distinct roots of  $f(u, v) = 0$ , that is,  $u = h_0(v) (\equiv 0)$  defined in  $J_0^* = (0, +\infty)$  and  $u = h_1(v) \geq c$  defined in  $J_1^* = (0, f_0(c)/k)$  which satisfies  $kv - f_0(h_1(v)) = 0$  (see Fig. 1). Then, the assumption (A1) is obviously satisfied.

We note that

$$\mathcal{F}(\beta) = \int_{h_0(\beta)}^{h_1(\beta)} f(s, \beta) ds = \int_0^{h_1(\beta)} (f_0(s) - k\beta)s ds$$

is defined in  $(f_0(0)/k, f_0(c)/k)$  and satisfies the inequalities  $\mathcal{F}(f_0(0)/k) < 0 < \mathcal{F}(f_0(c)/k)$  by (A4). Using the fact

$$\begin{aligned} \frac{d}{d\beta} \mathcal{F} &= (f_0(h_1(\beta)) - k\beta)h_1(\beta) - \frac{d}{dv} h_1(\beta) - \int_0^{h_1(\beta)} ks ds \\ &= -\frac{1}{2}k(h_1(\beta))^2 \leq -\frac{1}{2}kc^2 < 0, \end{aligned}$$

we find that  $\mathcal{F}(\beta)$  has a unique zero  $\beta^* \in (f_0(0)/k, f_0(c)/k)$  which implies that the assumption (ii) of (A3) is satisfied.

Next, we consider whether or not (A2) is satisfied. From (i) and (ii) of (A5), it follows

$$\begin{aligned} g(h_0(v), v) &= -g_0(v)v \leq -g_0(0)v < 0 & (v \in (0, +\infty)), \\ g(h_1(v), v) &= -(g_0(v) - kh_1(v))v > 0 & \left(v \in \left(0, \frac{f_0(0)}{k}\right)\right) \end{aligned}$$

and

$$\frac{d}{dv} G_0(v) = \frac{d}{dv} g(h_0(v), v) = \left(\frac{d}{dv} g_0(v)\right)v - g_0(v) < 0 \quad (v \in [0, +\infty)).$$

From (iii) of (A5), we also have

$$\frac{d}{dv} G_1(v) = \frac{d}{dv} g(h_1(v), v) < 0 \quad \left(v \in \left(\beta^*, \frac{f_0(c)}{k}\right)\right).$$

Taking  $J_0$  and  $J_1$  as  $J_0 = (f_0(0)/k, \beta^* + \mu)$  and  $J_1 = (\beta^* - \mu, f_0(c)/k)$  for a positive constant  $\mu$  respectively, we know that (A2) is satisfied.

Finally, we show that (i) and (iii) of (A3) are satisfied; we have

$$\frac{\partial}{\partial u} f(h_0(v), v) = \frac{\partial}{\partial u} f(0, v) = f_0(0) = -kv < 0 \quad \text{on } J_0,$$

and

$$\begin{aligned} \frac{\partial}{\partial u} f(h_1(v), v) &= f_0(h_1(v)) - kv + \left( \frac{d}{du} f_0(h_1(v)) \right) h_1(v) \\ &= \left( \frac{d}{du} f_0(h_1(v)) \right) h_1(v) < 0 \quad \text{on } J_1. \end{aligned}$$

This implies (i) of (A3). By setting  $\mathcal{F}_i(y) = \int_{h_i(\beta^*)}^y f(s, \beta^*) ds$ , it follows that

$$\frac{d}{dy} \mathcal{F}_i = f(y, \beta^*) = (f_0(y) - k\beta^*)y \quad (i = 0, 1)$$

has only one zero in  $(h_0(\beta^*), h_1(\beta^*))$  and that

$$\mathcal{F}_i(h_0(\beta^*)) = \mathcal{F}_i(h_1(\beta^*)) = 0 \quad (i = 0, 1).$$

Hence, we obtain  $\mathcal{F}_i(y) > 0$  for all  $y \in (h_0(\beta^*), h_1(\beta^*))$ ,  $(i = 0, 1)$ , which implies that (iii) of (A3) is satisfied. Therefore we find that (A1) ~ (A3) are all satisfied, so we can construct solutions  $(u(x; \varepsilon), v(x; \varepsilon))$  of  $(2.5)_\varepsilon$  from Theorem 2.

Now, from an ecological viewpoint, we must verify the nonnegativity of the solution since  $u$  and  $v$  represent the population densities of two species, a prey and its predator. According to § 4, we set  $U(x) = h_0(V(x))$  on  $[x_0, x_1]$  and  $U(x) = h_1(V(x))$  on  $[x_1, x_2]$ . From Theorem 1, there exists a positive constant  $q_1$  such that

$$\frac{f_0(0)}{k} + q_1 \leq V(x) \leq \frac{f_0(c)}{k} - q_1 \quad (x \in [x_0, x_2]).$$

Similarly, from Theorem 2, there exists a sufficiently small positive constant  $\varepsilon_1$  such that  $\|v(\cdot; \varepsilon) - V\|_{C^0[x_0, x_2]} < q_1/2$  for all  $0 < \varepsilon < \varepsilon_1$ . Therefore, we have

$$(5.1) \quad 0 < \frac{f_0(0)}{k} + \frac{q_1}{2} < v(x; \varepsilon) < \frac{f_0(c)}{k} - \frac{q_1}{2} \quad (x \in [x_0, x_2]).$$

In the proof of Theorem 2, we obtained  $u(x; \varepsilon)$  as

$$u(x; \varepsilon) = \begin{cases} h_0(V(x)) + z_0(x; \varepsilon) + r_0(x; \varepsilon) & (x \in [x_0, x_1 + \delta(\varepsilon)]), \\ h_1(V(x)) + z_1(x; \varepsilon) + r_1(x; \varepsilon) & (x \in [x_1 + \delta(\varepsilon), x_2]), \end{cases}$$

where  $z_i$ ,  $(i = 0, 1)$  are the boundary layer correction terms satisfying  $0 \leq z_0(x; \varepsilon) \leq \bar{\gamma}$  and  $-\bar{\gamma} \leq z_1(x; \varepsilon) \leq 0$  where  $\bar{\gamma} = \frac{1}{2}(h_1(\beta^*) - h_0(\beta^*)) = \frac{1}{2}h_1(\beta^*)$ . Noting that  $r_i(x; \varepsilon)$ ,  $(i = 0, 1)$  converge to zero uniformly on each interval as  $\varepsilon$  tends to zero and that they are equal to zero at the end point of each interval, we find that, for any positive  $q_2$ , there exists a positive constant  $\varepsilon_2$  such that  $u(x; \varepsilon) > -q_2$  on  $[x_0, x_1 + \delta(\varepsilon)]$  and  $u(x; \varepsilon) > 0$  on  $[x_1 + \delta(\varepsilon), x_2]$  for all  $0 < \varepsilon < \varepsilon_2$ . Now, suppose that there exists a point  $\zeta$  in an interval  $[x_0, x_1 + \delta(\varepsilon)]$  such that  $u(\zeta; \varepsilon) < 0$ . Then there exists some interval  $I_0 = (\zeta_1, \zeta_2)$  contained in  $[x_0, x_1 + \delta(\varepsilon)]$  such that

$$(5.2) \quad \begin{aligned} u(x; \varepsilon) &< 0 \quad \text{in } I_0, \\ \frac{d}{dx} u(\zeta_1; \varepsilon) &\leq 0, \quad u(\zeta_2; \varepsilon) = 0, \quad \frac{d}{dx} u(\zeta_2; \varepsilon) \geq 0. \end{aligned}$$

On the other hand, by virtue of (5.1), we have  $kv(x; \varepsilon) - f_0(0) > k(q_1/2)$ , and therefore, if  $q_2$  is chosen sufficiently small,

$$kv(x; \varepsilon) - f_0(u(x; \varepsilon)) > k\frac{q_1}{4} \quad (x \in I_0).$$

Hence, we have

$$\varepsilon^2 \frac{d^2}{dx^2} u = -(f_0(u) - kv)u < k\frac{q_1}{4} u < 0 \quad (x \in I_0).$$

By integrating the above inequality, it holds that

$$\begin{aligned} \varepsilon^2 \frac{d}{dx} u(\xi_2; \varepsilon) &< \varepsilon^2 \frac{d}{dx} u(\xi_1; \varepsilon) + k \int_{\xi_1}^{\xi_2} \frac{q_1}{4} u(x; \varepsilon) dx \\ &\leq \frac{kq_1}{4} \int_{\xi_2}^{\xi_1} u(x; \varepsilon) dx < 0, \end{aligned}$$

which contradicts to (5.2). Thus, the proof of Theorem 3 is completed.

**6. Concluding remarks.** We have considered the two point boundary value problem with two nonnegative parameters  $\varepsilon$  and  $\sigma$ ,

$$\begin{aligned} 0 &= \varepsilon^2 \frac{d^2}{dx^2} u + f(u, v), \\ 0 &= \frac{1}{\sigma} \frac{d^2}{dx^2} v + g(u, v) \end{aligned}$$

subject to zero flux boundary conditions and have shown the existence of large amplitude solutions when  $\varepsilon$  and  $\sigma$  are both sufficiently small. We note here that the result remains valid when  $\varepsilon$  and  $(\partial/\partial u)g(h(V(x)), V(x))$  are both sufficiently small. Smallness of  $\sigma$  is replaced by the latter condition. Applying this to the prey-predator model (2.5) <sub>$\varepsilon$</sub> , it is found that there exist nonnegative solutions of (2.5) <sub>$\varepsilon$</sub>  for any fixed  $\sigma > 0$  if  $\varepsilon$  and the interaction rate  $k$  are sufficiently small.

For an application of our results, we have given one model which describes a prey-predator system. In addition to this, we mention two examples. One is the model proposed by Gierer and Meinhardt [6], which is

$$\begin{aligned} 0 &= d_a \frac{d^2}{dx^2} a + \rho\rho_0 + \frac{c\rho a^2}{h(1 + \kappa a^2)} - \mu a, \\ 0 &= d_h \frac{d^2}{dx^2} h + c'\rho' a^2 - \nu h, \end{aligned}$$

where  $0 < d_a \ll d_h$  and  $\rho, \rho', \rho_0, c, c', \kappa, \mu$  and  $\nu$  are all positive constants. Another is a model substrate-inhibition reaction diffusion system

$$\begin{aligned} 0 &= d_1 \frac{d^2}{dx^2} u + j_1 - u - \beta r(u, v), \\ 0 &= d_2 \frac{d^2}{dx^2} v + j_2 - \gamma r(u, v), \end{aligned}$$

where  $r(u, v) = uv/(1 + u + v + Ku^2)$  and  $j_1, j_2, \beta, \gamma$  and  $K$  are positive constants. This model without diffusion was originally proposed by Seelig [16].

We have not as yet been able to discuss the stability of the solutions obtained here. The difficulty is caused by multiplicity of solutions. We note, however, that Fife's heuristic argument concerning the stability of large amplitude solutions [4] is worthy of attention.

We may construct approximations of  $(2.1)_\varepsilon$  to any desired degree of accuracy, which have power series of  $\varepsilon$ .

**Acknowledgment.** The authors wish to thank Professor Masaya Yamaguti for valuable suggestions and many criticisms.

#### REFERENCES

- [1] E. CONWAY, D. HOFF AND J. SMOLLER, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.
- [2] T. ERNEUX AND M. HERSCHKOWITZ-KAUFMAN, *Rotating waves as asymptotic solutions of a model chemical reaction*, J. Chem. Phys., 66 (1977), pp. 248–250.
- [3] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.
- [4] ———, *Stationary patterns for reaction-diffusion equations*, Research Notes in Mathematics, vol. 14, Pitman, London, 1977, pp. 81–121.
- [5] ———, *Asymptotic states for equations of reaction and diffusion*, Bull. Amer. Math. Soc., 84 (1978), pp. 693–726.
- [6] A. GIERER AND H. MEINHARDT, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.
- [7] J. P. KEENER, *Activators and inhibitors in pattern formation*, Studies in Appl. Math., 59 (1978), pp. 1–23.
- [8] Y. KURAMOTO, *Diffusion-induced chaos in reaction systems*, Progr. Theoret. Phys, 64 (1978), pp.
- [9] S. A. LEVIN, *Population dynamics models in heterogeneous environments*, Ann. Rev., Ecol. Syst., 7 (1976), pp. 287–310.
- [10] J. L. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites Nonlinéaires*, Dunod, Paris, 1969.
- [11] R. M. MAY, *Stability and complexity in model ecosystems*, Monograph on Population Biology 6, Princeton University Press, Princeton, NJ, 1973.
- [12] M. MIMURA, *Spatial structures in nonlinear interaction-diffusion systems*, Kokyuroku 317, RIMS, Kyoto Univ., Kyoto, Japan, 1978 pp. 17–29.
- [13] M. MIMURA AND J. D. MURRAY, *On a diffusive prey-predator model which exhibits patchiness*, J. Theoret. Biol., 75 (1979), pp. 249–262.
- [14] M. MIMURA AND Y. NISHIURA, *Spatial patterns for interaction-diffusion equations in biology*, Proc. International Symposium on Mathematical Topics in Biology, 1978, pp. 136–146.
- [15] M. MIMURA, Y. NISHIURA AND M. YAMAGUTI, *Some diffusive prey and predator systems and their bifurcation problems*, Ann. New York Acad. of Sciences, 316 (1979), pp. 490–510.
- [16] F. F. SEELIG, *Chemical oscillations by substrate inhibition. A parametrically universal oscillator type in homogeneous catalysis by metal complex formation*, Z. Naturforsch, 31a (1976), pp. 731–738.
- [17] J. M. SMITH, *Models in Ecology*, Cambridge University Press, London, 1974.
- [18] C. A. STUART, *Differential equations with discontinuous nonlinearities*, Arch. Rational Mech. Anal., 63 (1976), pp. 59–75.
- [19] M. TABATA, *Two-point boundary value problems with a discontinuous semilinear term*, Kokyuroku 317, RIMS, Kyoto Univ., Kyoto, Japan, 1978, pp. 93–101.

## ON THE EXISTENCE OF WEAK-SOLUTIONS TO AN $n$ -DIMENSIONAL STEFAN PROBLEM WITH NONLINEAR BOUNDARY CONDITIONS\*

J. R. CANNON<sup>†</sup> AND EMMANUELE DiBENEDETTO<sup>†</sup>

**Abstract.** The weak formulation of an  $n$ -dimensional Stefan problem with nonlinear flux assigned on the fixed boundary is studied. An existence theorem is proved by using Galerkin procedure, monotonicity methods and trace theorems.

**Introduction.** This paper is concerned with a free boundary Stefan-like problem in several space variables associated with a parabolic operator of the form

$$(A) \quad Lu = \alpha(u) \frac{\partial u}{\partial t} - \operatorname{div} \{K(u) \nabla_x u + \mathbf{b}(x, t, u)\} + c(x, t, u).$$

A precise formulation is given in § 1. The solution is required to vanish on the free boundary  $\Gamma$  while the condition

$$(B) \quad [K_1(u_1) \nabla_x u_1 + \mathbf{b}_1(x, t, u_1) - K_2(u_2) \nabla_x u_2 - \mathbf{b}_2(x, t, u_2)] \cdot \mathbf{n} = \nu \frac{\partial \Phi}{\partial t} |\operatorname{grad} \Phi|^{-1}$$

on  $\Gamma$  is required, where the subscripts 1 and 2 refer to the two sides of  $\Gamma$ ,  $\mathbf{n}$  is the spatial normal to  $\Gamma$  at the time  $t$  and  $\Phi(x, t)$  is a continuously differentiable function which implicitly determines the free boundary  $\Gamma$  in the domain where (A) is defined.

The problem can be viewed as a model of a solid-liquid phase change at a prescribed temperature.

An example situation is an ice-water mix contained in a fixed region  $G$  whose initial nonpositive temperature is specified. It should be noticed that we were unable to prove uniqueness for this system. To obtain stability and uniqueness one has to come off the boundary with the trace theorem which involves  $|\nabla u|$ .

This rules out any application of the methods of Cannon-Hill [1] and Friedman [4].

The plan of the paper is as follows. Section 1 contains the classical formulation of the problem while a generalized formulation and our concept of weak-solutions are introduced and discussed in § 2. Assumptions and the statement of the existence theorem are given in § 3. The theorem is demonstrated in § 4 by using a Galerkin-type argument. Finally in § 5 the Dirichlet problem for (A)-(B) is briefly discussed.

**1. Classical formulation of the problem.** Let  $G$  denote a bounded domain in  $\mathbb{R}^n$  with smooth boundary  $\partial G$ . If  $t \in [0, T]$ ,  $T > 0$ , then let  $G(t) = G \times \{t\}$ ,  $\partial G(t) = \partial G \times \{t\}$ ,  $\Omega_t = \bigcup_{0 \leq \tau \leq t} G(\tau)$ . We assume that the domain  $\Omega_T$  is divided into  $\Omega_1$  and  $\Omega_2$  by the free boundary  $\Gamma \equiv \Gamma_T = \bigcup_{0 \leq \tau \leq T} \Gamma(\tau)$ , where  $\Gamma(t)$  is a hypersurface in  $G(t)$  determined by  $\Phi(x, t) = 0$ . The function  $\Phi \in C^1(\bar{\Omega}_T)$ ,  $\Phi < 0$  in  $\Omega_1$ ,  $\Phi > 0$  in  $\Omega_2$  and  $\nabla_x \Phi(x, t) \neq 0$  on  $\Gamma$  where  $\nabla_x$  denotes the gradient with respect to the  $x \equiv (x_1, x_2, \dots, x_n)$  variables only. The set  $\Gamma(0)$  divides the initial region  $G(0)$  into two regions  $G_1(0)$  and  $G_2(0)$ . With respect to the boundary  $S_T$  of  $\Omega_T$ , let  $S_i = \bar{\Omega}_i \cap S_T$ .

Consider the problem of determining real valued functions  $u_i$ ,  $i = 1, 2$ , defined in

\* Received by the editors October 18, 1978, and in revised form June 25, 1979.

<sup>†</sup> Department of Mathematics, The University of Texas at Austin, Austin, Texas 78712. This research was supported in part by the Consiglio Nazionale delle Ricerche d'Italia.

$\bar{\Omega}_i$ , which satisfy

$$(1.1) \quad \alpha_i(u_i) \frac{\partial u_i}{\partial t} = \operatorname{div} \{K_i(u_i) \nabla_x u_i + \mathbf{b}_i(x, t, u_i)\} + c_i(x, t, u_i) \quad \text{in } \Omega_i, \quad i = 1, 2,$$

$$(1.2) \quad u_i = h_i(x) \quad \text{in } G_i(0), \quad i = 1, 2, \quad h_1(x) > 0, \quad h_2(x) < 0,$$

$$(1.3) \quad K_i(u_i) \frac{\partial u_i}{\partial \mathbf{n}} + \mathbf{b}_i(x, t, u_i) \cdot \mathbf{n} = g_i(x, t, u_i) \quad \text{on } S_i - \Gamma,$$

$$(1.4) \quad u_i = 0 \quad \text{on } \Gamma, \quad i = 1, 2,$$

$$(1.5) \quad [K_1(u_1) \nabla_x u_1 + \mathbf{b}_1(x, t, u_1) - K_2(u_2) \nabla_x u_2 - \mathbf{b}_2(x, t, u_2)] \cdot \nabla_x \Phi = \nu \frac{\partial \Phi}{\partial t},$$

where  $\nu$  is a positive constant,  $\mathbf{n}$  is the outer normal to  $S_i$ , the div-operator is with respect to the  $x$  variable only, the  $\alpha_i$ ,  $K_i$ ,  $c_i$ ,  $h_i$  and  $g_i$ ,  $i = 1, 2$  are known functions of their arguments, and the  $\mathbf{b}_i$  are vector valued functions mapping  $\Omega_T \times \mathbb{R}$  into  $\mathbb{R}^n$ . Here  $\alpha_i(u)$  and  $K_i(u)$  are sufficiently smooth functions defined for  $(-1)^i u \leq 0$ , which for some constants  $\gamma_0$  and  $\gamma_1$ , satisfy

$$(1.6) \quad 0 < \gamma_0 < \alpha_i(u), \quad K_i(u) < \gamma_1, \quad i = 1, 2.$$

**2. Generalized formulation of the problem.** Consider smooth test functions  $\varphi$  in  $\mathbb{R}^{n+1}$  such that

$$(2.1) \quad \varphi = 0 \quad \text{on } G(T).$$

Using formally identical arguments to those contained in [1], [4], we see that for all such  $\varphi$ , a classical solution to (1.1)–(1.5) must satisfy

$$(2.2) \quad \int \int_{\Omega_T} \left\{ \alpha(u) \frac{\partial \varphi}{\partial t} - \nabla_x k(u) \cdot \nabla_x \varphi - \mathbf{b}(x, t, u) \cdot \nabla_x \varphi + c(x, t, u) \varphi \right\} dx dt + \int_{G(0)} \varphi \alpha(h) dx + \int_{S_T} \varphi g(x, t, u) d\sigma = 0,$$

where

$$(2.3) \quad u = \begin{cases} u_1 & \text{in } \Omega_1, \\ u_2 & \text{in } \Omega_2, \end{cases} \quad h = \begin{cases} h_1 & \text{in } G_1(0), \\ 0 & \text{on } \Gamma(0), \\ h_2 & \text{in } G_2(0), \end{cases}$$

$$(2.4) \quad \alpha(u) = \begin{cases} \int_0^u \alpha_1(\xi) d\xi, & u > 0, \\ \int_0^u \alpha_2(\xi) d\xi - \nu, & u < 0, \\ [-\nu, 0] & \text{for } u = 0, \end{cases}$$

$$(2.5) \quad k(u) = \begin{cases} \int_0^u K_1(\xi) d\xi, & u > 0, \\ 0, & u = 0, \\ \int_0^u K_2(\xi) d\xi, & u < 0, \end{cases}$$

and  $\mathbf{b}$ ,  $c$  and  $g$  are defined in a similar manner. The following definitions and the discussion below will motivate our concept of weak-solution.

Let  $L_{q,r}(\Omega_T)$  denote the Banach space of those measurable functions mapping  $\Omega_T \rightarrow \mathbb{R}$  with norm defined by

$$(2.6) \quad \|u\|_{q,r,\Omega_T}^r = \int_0^T \|u\|_{q,G(\tau)}^r d\tau,$$

where

$$\|u\|_{q,G(\tau)}^q = \int_{G(\tau)} |u(x, \tau)|^q dx.$$

When  $q = r = 2$ ,  $L_{2,2}(\Omega_T)$  coincides with the Hilbert space  $L_2(\Omega_T)$  whose inner product  $(\cdot, \cdot)_{2,\Omega_T}$  generates the norm  $\|\cdot\|_{2,\Omega_T} = \|\cdot\|_{2,2,\Omega_T}$ . Let  $W_2^{1,0}(\Omega_T)$  denote the Hilbert space with inner product

$$(2.7) \quad (u, v)_{W_2^{1,0}(\Omega_T)} = (u, v)_{2,\Omega_T} + \sum_{i=1}^n \left( \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \right)_{2,\Omega_T}$$

while  $W_2^{1,1}(\Omega_T)$  denotes the Hilbert space with inner product

$$(2.8) \quad (u, v)_{W_2^{1,1}(\Omega_T)} = (u, v)_{W_2^{1,0}(\Omega_T)} + \left( \frac{\partial u}{\partial t}, \frac{\partial v}{\partial t} \right)_{2,\Omega_T}.$$

Here  $\partial u/\partial x_i$  and  $\partial u/\partial t$  denote generalized derivatives. Also let  $V_2(\Omega_T) \subset W_2^{1,0}(\Omega_T)$  denote the Banach space with norm

$$(2.9) \quad |u|_{V_2(\Omega_T)} = \text{ess sup}_{0 \leq t \leq T} \|u(\cdot, t)\|_{2,G(t)}^2 + \|\nabla_x u\|_{2,\Omega_T}^2,$$

where

$$(2.10) \quad \|\nabla_x u\|_{2,\Omega_T}^2 = \sum_{i=1}^n \left( \frac{\partial u}{\partial x_i}, \frac{\partial u}{\partial x_i} \right)_{2,\Omega_T}.$$

Finally we let  $V_2^{1,0}(\Omega_T) \subset V_2(\Omega_T)$  denote the Banach space of functions such that the map  $t \rightarrow u$  is continuous with respect to  $\|\cdot\|_{2,G}$  and the norm  $|\cdot|_{V_2^{1,0}(\Omega_T)}$  is that of (2.9) with the ess deleted.

Our definition of weak-solution will depend crucially on the notion of relation on a Cartesian product  $X \times Y$  of linear spaces  $X, Y$ .

A relation  $F$  on  $X \times Y$  is a subset of  $X \times Y$ . The domain of  $F$  is  $\{x \in X : Fx \neq \Phi\}$  and the image of  $x$  by  $F$  is the set  $Fx = \{y \in Y : [x, y] \in F\}$ . The range of  $F$  is  $\mathcal{R} = \cup\{Fx : x \in X\}$ . We identify  $F$  with its graph which is defined as

$$\text{graph } F \equiv \{[x, y] \in X \times Y : [x, y] \in F\}.$$

This permits the definition of inverse of  $F$ . The inverse of  $F$  is the relation  $F^{-1}$  on  $Y \times X$ , whose graph is symmetric with respect to the graph of  $F$ .

The graph of every function from a subset of  $X$  into  $Y$  is a relation on  $X \times Y$ . Therefore, it is natural to identify functions as relations. A relation  $F$  is a function:  $X \rightarrow Y$  if and only if the set  $Fx$  is a singleton for every  $x$  in the domain of  $F$ .

From (2.4) follows that  $\alpha(\cdot)$  is a relation in  $\mathbb{R} \times \mathbb{R}$ , whose inverse  $\alpha^{-1}(\cdot)$  is a function.



DEFINITION. By a *weak solution* of (1.1)–(1.5), we mean a function  $u \in V_2^{1,0}(\Omega_T)$  defined by

$$(2.11) \quad u = \alpha^{-1}(v),$$

where  $v$  is a function defined in  $\Omega_T$  such that

$$(2.12) \quad v \subset \alpha(\cdot),$$

the inclusion being intended in the sense of the graphs, and  $v$  and  $u = \alpha^{-1}(v)$  satisfy

$$(2.13) \quad \int \int_{\Omega_T} \left\{ v \frac{\partial \varphi}{\partial t} - \nabla_x k(\alpha^{-1}(v)) \cdot \nabla_x \varphi - \mathbf{b}(x, t, \alpha^{-1}(v)) \cdot \nabla_x \varphi + c(x, t, \alpha^{-1}(v)) \varphi \right\} dx dt + \int_{G(0)} \varphi \alpha(h) dx + \int_{S_T} \varphi g(x, t, \alpha^{-1}(v)) d\sigma = 0$$

for all the  $\varphi \in W_2^{1,1}(\Omega_T)$  which satisfy (2.1).

*Remark 1.* The  $\int_{G(0)} \varphi \alpha(h) dx$  is well defined if  $h \neq 0$  a.e. in  $G(0)$ .

*Remark 2.* Every function in  $V_2^{1,0}(\Omega_T)$  has trace in  $L_2(S_T)$  (see [6]), so every term in (2.13) is well-defined modulo basic assumptions on the data that will be specified below.

*Remark 3.* Each classical solution of (1.1)–(1.5) generates a weak-solution. Standard arguments [1], [4], imply that any sufficiently smooth weak solution whose level set  $\{u = 0\}$  is a smooth surface on which  $\nabla_x u \neq 0$  is a classical solution.

**3. Assumptions and statement of results.** The assumption (1.6) together with the definitions (2.4)–(2.5) imply that  $k(u)$  is Lipschitz continuous for all  $u$ , while  $\alpha(u)$  is Lipschitz continuous for all  $u \neq 0$ ; moreover for  $u \neq 0$  we have

$$(3.1) \quad 0 < \gamma_0 \leq \alpha'(u), \quad k'(u) \leq \gamma_1$$

and

$$(3.2) \quad \gamma_0 |u| \leq |\alpha(u)|, \quad k(u) \leq \gamma_1 |u|.$$

With respect to the data functions  $h, \mathbf{b}, c,$  and  $g$ , we have the following two assumptions.

(A1) The functions  $\mathbf{b}, c,$  and  $g$  are continuous functions of their arguments over the Cartesian product of their respective space-time domains with  $\{-\infty < u < +\infty\}$ , and satisfy respectively the following growth conditions:

$$(3.3) \quad \begin{aligned} |\mathbf{b}(x, t, u)| &\leq K_1 |u|, \\ |c(x, t, u)| &\leq K_0 + K_1 |u|, \\ |g(x, t, u)| &\leq K_0 + K_1 |u|, \end{aligned}$$

where  $|\mathbf{b}|$  denotes the Euclidean norm of  $\mathbf{b}$  as a vector in  $\mathbb{R}^n$  and  $K_0$  and  $K_1$  are positive constants.

*Remark 4.* The requirement that  $\mathbf{b}(x, t, 0) = 0$  exhibits the interaction of this term with the lack of uniform parabolicity exhibited by the vanishing of  $\alpha^{-1}(u)$  for certain values of  $u$ .

(A2) The function  $h \in L_2(G)$  and  $h \neq 0$ , a.e.

**THEOREM.** *Under the assumptions (1.6), (A1) and (A2), there exists a weak-solution of (1.1)–(1.5).*

**4. Proof of the theorem.** Let  $\varphi \in C_0^\infty(-1, 1)$  such that  $\int_{-1}^1 \varphi(y) dy = 1$  and set

$$\alpha_m(u) = m \int_{u-(1/m)}^{u+(1/m)} \bar{\alpha}(y)\varphi[m(u-y)] dy,$$

where

$$\bar{\alpha}(y) = \begin{cases} \alpha(y), & y \neq 0, \\ 0, & y = 0. \end{cases}$$

If  $u > 2/m$  we have

$$(4.1) \quad \alpha'_m(u) = m \int_{1/m}^\infty \bar{\alpha}'(y)\varphi(m(u-y)) dy \leq \gamma_1.$$

An analogous formula holds for  $u < -(2/m)$ ; therefore we see that

$$(4.2) \quad \gamma_0 \leq \alpha'_m(u) \leq \gamma_1 \quad \text{for } |u| > \frac{2}{m}.$$

From (4.1) we deduce also that

$$(4.3) \quad \gamma_0 \leq \alpha'_m(u) \leq Cm \quad \text{for } |u| \leq \frac{2}{m},$$

where  $C$  is a constant which is independent of  $m$ . The  $\alpha_m(\cdot)$  are invertible and

$$\alpha_m^{-1}(u) \rightarrow \alpha^{-1}(u)$$

uniformly on compact sets. We have also

$$(4.4) \quad 0 < \frac{1}{Cm} \leq \alpha_m^{-1'}(u) \leq \gamma_0^{-1}.$$

By defining  $k_m(u)$  in a similar way, we see that

$$k_m(u) \rightarrow k(u)$$

uniformly on compact sets, and that

$$(4.5) \quad \gamma_0 \leq k'_m(u) \leq \gamma_1.$$

We consider now the problem of finding for each  $m$  a function  $v_m \in V_2^{1,0}(\Omega_T)$  which satisfies

$$(4.6) \quad \int \int_{\Omega_T} \left\{ v_m \frac{\partial \varphi}{\partial t} - \nabla_x k_m(\alpha_m^{-1}(v_m)) \cdot \nabla_x \varphi - \mathbf{b}(x, t, \alpha_m^{-1}(v_m)) \cdot \nabla_x \varphi + c(x, t, \alpha_m^{-1}(v_m)) \varphi \right\} dx dt + \int_{G(0)} \varphi \alpha(h) dx + \int_{S_T} \varphi g(x, t, \alpha_m^{-1}(v_m)) d\sigma = 0$$

for all  $\varphi \in W_2^{1,1}(\Omega_T)$  which vanish when  $t = T$ .

Let

$$(4.7) \quad W_{m,l}(x, t) = \sum_{i=1}^l \beta_{m,i}^l(t) z_i(x)$$

denote the  $l$ th Galerkin approximate of  $v_m$ , where the  $z_i, i = 1, 2, \dots$  satisfy

$$-\Delta z_i = \lambda_i z_i \quad \text{in } G,$$

$$\frac{\partial z_i}{\partial \mathbf{n}} \Big|_{\partial G} = 0 \quad \text{on } \partial G.$$

It is clear that we can assume that the  $z_i$  form an orthonormal basis for  $L_2(G)$  and belong to  $W_2^1(G)$ . Here  $L_2(G)$  is the Hilbert space of square integrable functions over  $G$  with the usual inner product  $(\cdot, \cdot)_{L_2(G)}$  and  $W_2^1(G)$  is the Hilbert space with inner product

$$(\xi, \psi)_{W_2^1(G)} = (\xi, \psi)_{L_2(G)} + \sum_{i=1}^n \left( \frac{\partial \xi}{\partial x_i}, \frac{\partial \psi}{\partial x_i} \right)_{L_2(G)}.$$

For the construction of the  $W_{m,l}(x, t)$  [global in time] we refer to [2], [3]. The  $W_{m,l}(x, t)$  satisfy the problem

$$(4.8) \quad \frac{\partial}{\partial t} W_{m,l}(x, t) - \Delta k_m[\alpha_m^{-1}(W_{m,l}(x, t))] - \operatorname{div} \mathbf{b}(x, t, \alpha_m^{-1}(W_{m,l}(x, t))) + c(x, t, \alpha_m^{-1}(W_{m,l}(x, t))) = 0,$$

$$(4.9) \quad \{\nabla_x k_m[\alpha_m^{-1}(W_{m,l}(x, t))] + \mathbf{b}(x, t, \alpha_m^{-1}(W_{m,l}(x, t)))\} \cdot \mathbf{n} = g(x, t, \alpha_m^{-1}(W_{m,l}(x, t))), \quad (x, t) \in S_T.$$

$$(4.10) \quad W_{m,l}(x, 0) = V_l(x)$$

in the sense of the projection over the span of  $\{z_1, z_2, \dots, z_l\}$ , where

$$V_l(x) = \sum_{i=1}^l c_{m,i} z_i(x)$$

and

$$v(x, 0) = \alpha_m(h) = \sum_{i \geq 0} c_{m,i} z_i(x).$$

Namely, denoting with  $P_l$  the  $L_2(G)$  projection onto the linear span of  $\{z_1, z_2, \dots, z_l\}$ ,  $W_{m,l}$  is the unique element in  $W_2^{1,1}(\Omega_T)$  satisfying (4.10) and

$$\int_G \frac{\partial W_{m,l}}{\partial t} P_l w + \int_G \{\nabla k_m[\alpha_m^{-1}(W_{m,l})] + \mathbf{b}\} \cdot \nabla P_l w + \int_G c P_l w - \int_{\partial G} g(x, t, \alpha_m^{-1}(W_{m,l})) P_l w = 0$$

for all  $w \in W_2^1(G)$ .

Next we derive a priori bounds on the  $V_2$ -norm of  $W_{m,l}(x, t)$ , independent of  $m$  and  $l$ . To simplify the notation we will make the convention of indicating with  $C$  generic, nonnegative constants that depend upon quantities that will be specified as the constant appears. Set

$$(4.11) \quad U_{m,l} = \alpha_m^{-1}(W_{m,l})$$

then

$$(4.12) \quad W_{m,l} = \alpha_m(U_{m,l}).$$

LEMMA 1. *There is a constant  $C$  dependent only upon  $h, \Omega_T, K_0, K_1, \gamma_1, \gamma_0, \nu$  such*

that for all  $t \in [0, T]$  and all  $m, l$

$$(4.13) \quad \|W_{m,l}(\cdot, t)\|_{2,G} \leq C.$$

*Proof.* We take the inner product in  $L_2(\Omega_t)$  of (4.8) by  $W_{m,l}(x, t)$ . This gives

$$(4.14) \quad \int_0^t \frac{d}{dt} \|W_{m,l}(\cdot, \tau)\|_{2,G}^2 d\tau + \int_{\Omega_t} \nabla_x k_m(U_{m,l}) \cdot \nabla_x \alpha_m(U_{m,l}) dx d\tau \\ \leq \left| \int_{S_t} \alpha_m(U_{m,l}) g(x, t, U_{m,l}) d\sigma \right| + \left| \int_{\Omega_t} \mathbf{b}(x, \tau, U_{m,l}) \cdot \nabla_x \alpha_m(U_{m,l}) dx d\tau \right| \\ + \left| \int_{\Omega_t} W_{m,l} \cdot c(x, \tau, \alpha_m^{-1}(W_{m,l})) dx d\tau \right| = J_1 + J_2 + J_3.$$

We estimate the  $J_i, i = 1, 2, 3$ , as follows. Using (3.3) we obtain

$$(4.15) \quad J_1 \leq \int_{S_t} |\alpha_m(U_{m,l})| \{K_0 + K_1 |U_{m,l}|\} d\sigma \\ = \int_{\{(x,t) \mid |U_{m,l}| \leq 2/m\} \cap S_t} |\alpha_m(U_{m,l})| \{K_0 + K_1 |U_{m,l}|\} d\sigma \\ + \int_{\{(x,t) \mid |U_{m,l}| > 2/m\} \cap S_t} |\alpha_m(U_{m,l})| \{K_0 + K_1 |U_{m,l}|\} d\sigma = J'_1 + J''_1.$$

We have

$$|\alpha_m(U_{m,l}) - \alpha_m(0)| \leq \alpha'_m(\xi_{m,t}(x, t)) |U_{m,l}|,$$

where

$$|\xi_{m,t}(x, t)| \leq |U_{m,l}(x, t)|.$$

Therefore,

$$J'_1 \leq \int_{\{(x,t) \mid |U_{m,l}| \leq 2/m\} \cap S_t} \alpha'_m(\xi_{m,t}(x, t)) |U_{m,l}| \{K_0 + K_1 |U_{m,l}|\} d\sigma \\ + \int_{\{(x,t) \mid |U_{m,l}| \leq 2/m\} \cap S_t} |\alpha_m(0)| \{K_0 + K_1 |U_{m,l}|\} d\sigma.$$

Using now (4.3), we have

$$J'_1 \leq \int_{S_T} Cm \frac{2}{m} \{K_0 + 2K_1/m\} d\sigma + \nu \int_{S_T} \{K_0 + 2K_1/m\} d\sigma + C|S_T|,$$

where  $\nu$  is the constant appearing in (2.4) and  $|\Sigma|$  indicates the Lebesgue measure of the set  $\Sigma$ .

For  $J''_1$  we observe that from (2.4) and (4.2) it follows that

$$J''_1 \leq \gamma_1 \int_{S_t} (\nu + |U_{m,l}|) \{K_0 + K_1 |U_{m,l}|\} d\sigma \leq C_1 |S_T| + C_2 \int_{S_t} |U_{m,l}|^2 d\sigma \\ \leq C_1 |S_T| + C_2 \int_{S_t} |U_{m,l}|^2 d\sigma,$$

where  $C_i, i = 1, 2$  are constants which depend only upon  $\gamma_1, \nu, K_0, K_1$ , but not upon  $m$  and  $l$ . Substituting the estimates of  $J'_1$  and  $J''_1$  into (4.15) yields the existence of constants

$C_i, i = 1, 2$  such that

$$J_1 \leq C_1 + C_2 \int_{S_t} |U_{m,t}|^2 d\sigma.$$

By the trace theorem [7], for every  $\varepsilon > 0$  there is a constant  $C_2(\varepsilon)$  such that

$$(4.16) \quad J_1 \leq C_1 + C_2(\varepsilon) \int_{\Omega_t} |U_{m,t}|^2 dx d\tau + \varepsilon \int_{\Omega_t} |\nabla_x U_{m,t}|^2 dx d\tau.$$

We observe that by (2.4), (4.4) and (4.11), we have

$$(4.17) \quad |U_{m,t}| \leq \nu + \gamma_0^{-1} |W_{m,t}|.$$

Therefore from (4.16) we deduce that

$$(4.18) \quad J_1 \leq C_1(\varepsilon) + C_2(\varepsilon) \int_{\Omega_t} |W_{m,t}|^2 dx d\tau + \varepsilon \int_{\Omega_t} |\nabla_x U_{m,t}|^2 dx d\tau.$$

Next, we consider  $J_2$ . By (3.3)

$$(4.19) \quad \begin{aligned} J_2 &\leq \int_{\Omega_t} K_1 |U_{m,t}| \alpha'_m(U_{m,t}) |\nabla_x U_{m,t}| dx d\tau \\ &\leq \varepsilon \int_{\Omega_t} |\nabla_x U_{m,t}|^2 dx d\tau + \frac{K_1^2}{4\varepsilon} \int_{\Omega_t} \{\alpha'_m(U_{m,t}) |U_{m,t}|\}^2 dx d\tau. \end{aligned}$$

For the last integral by an analysis similar to that of  $J'_1$ , we obtain

$$\begin{aligned} \int_{\Omega_t} \{|U_{m,t}| \alpha'_m(U_{m,t})\}^2 &= \int_{\{(x,t) \mid |U_{m,t}| \leq 2/m\}} |U_{m,t}|^2 \alpha_m'^2(U_{m,t}) dx d\tau \\ &\quad + \int_{\{(x,t) \mid |U_{m,t}| > 2/m\}} |U_{m,t}|^2 \alpha_m'^2(U_{m,t}) dx d\tau \\ &\leq C|\Omega_t| + \gamma_1^2 \int_{\Omega_t} |U_{m,t}|^2 dx d\tau. \end{aligned}$$

By using (4.17), we see that there are constants  $C_1(\varepsilon), C_2(\varepsilon)$  such that

$$(4.20) \quad J_2 \leq C_1(\varepsilon) + C_2(\varepsilon) \int_{\Omega_t} |W_{m,t}|^2 dx d\tau + \varepsilon \int_{\Omega_t} |\nabla_x U_{m,t}|^2 dx d\tau.$$

Similar procedures yield

$$(4.21) \quad J_3 \leq C_1 + C_2 \int_{\Omega_t} |W_{m,t}|^2 dx d\tau,$$

where  $C_i, i = 1, 2$  depend upon  $\gamma_0, K_0, K_1, \nu, |\Omega_T|$  but not upon  $m$  and  $l$ .

We return now to (4.14). By using (4.3), (4.5) and the above estimates of the  $J_i$ 's,  $i = 1, 2, 3$ , we see that

$$(4.22) \quad \begin{aligned} &\|W_{m,t}(\cdot, t)\|_{2,G}^2 + (\gamma_0^2 - 2\varepsilon) \|\nabla_x U_{m,t}\|_{2,\Omega_t}^2 \\ &\leq \|V_t\|_{2,G}^2 + C_1(\varepsilon) + C_2(\varepsilon) \int_0^t \|W_{m,t}(\cdot, \tau)\|_{2,G}^2 d\tau, \end{aligned}$$

where  $C_i(\varepsilon)$ ,  $i = 1, 2$  are independent of  $m, l$ . We choose  $\varepsilon < \gamma_0^2/2$  and note that

$$\|V_i\|_{2,G}^2 \leq \|\alpha(h)\|_{2,G}^2.$$

The proof is concluded by an application of Gronwall’s inequality.

LEMMA 2. *There is a constant  $C$  which depends only upon  $h, \Omega_T, K_0, K_1, \gamma_1, \gamma_0, \nu$  such that*

$$(4.23) \quad \|U_{m,l}(\cdot, t)\|_{2,G}^2 + \int_0^t \|\nabla_x U_{m,l}(\cdot, \tau)\|_{2,G}^2 d\tau \leq C$$

holds for all  $m, l$  and all  $t \in [0, T]$ .

*Proof.* This is a direct consequence of (4.17) and (4.22).

From Lemma 1, it follows, by known analysis [2], [3], that for every  $i \in \mathbb{N}$ , there is a constant  $K(i)$ , which depends only upon  $h, \Omega_T, K_0, K_1, \gamma_1, \gamma_0, \nu$  and  $i$ , such that

$$(4.24) \quad |\beta_{m,i}^l(t) - \beta_{m,i}^l(s)| \leq K(i)(t-s)^{1/2}, \quad m, l = 1, 2, \dots$$

Therefore, for each  $i$  fixed,  $\{\beta_{m,i}^l(t)\}$  is a bounded equicontinuous family of functions.

Setting  $l = m$  and considering  $\{W_{m,m}\}$ , a subsequence can be selected via a diagonalization process and renamed  $\{W_m\}$ , so that each  $\beta_{m,i}^m(t)$  converges uniformly to a continuous limit  $\beta_i(t)$  as  $m$  tends to infinity. Setting

$$v = \sum_{i=1}^{\infty} \beta_i(t) z_i(x),$$

it follows from Lemma 1 that for each  $t$ ,  $\{W_m\}$  converges weakly to  $v$  in the space  $L_2(G)$ . Also, this convergence is uniform with respect to  $t$ .

From (4.23) it follows that the weak convergence of  $\{U_{m,i}\}$  in  $W_2^{1,0}(\Omega_T)$  can be incorporated into the diagonalization process so that

$$(4.25) \quad U_{m,m} = \alpha_m^{-1}(W_m) \rightharpoonup \xi(x, t) \in W_2^{1,0}(\Omega_T).$$

Next we will show that, for a suitable subsequence, again indexed with  $m$ ,

$$(4.26) \quad \alpha_m^{-1}(W_m) \rightarrow \alpha^{-1}(v) \quad \text{in } L_2(G)$$

for almost all  $t \in [0, T]$ . To show this the following preliminary results are needed.

LEMMA 3. *There is a subsequence of the  $\alpha_m^{-1}(W_m)$  (again indexed by  $m$ ) such that for almost all  $t \in [0, T]$ ,*

$$\int_G \alpha_m^{-1}(W_m)(W_m - v) dx \rightarrow 0$$

as  $m \rightarrow \infty$ .

*Remark.* The proof rests upon the following proposition.

PROPOSITION. *Let  $\{\varphi_n\}$  be a sequence of functions in  $W_2^{1,0}(\Omega_T)$  such that*

$$(4.27) \quad \|\varphi_n\|_{2,\Omega_T} + \|\nabla_x \varphi_n\|_{2,\Omega_T} \leq C, \quad n = 1, 2, \dots$$

Then, for each  $\varepsilon > 0$ , there exists an integer  $N(\varepsilon)$  which is independent of  $n$  such that

$$\sum_{i=N(\varepsilon)}^{\infty} \int_0^T \left| \int_G \varphi_n(x) z_i(x) dx \right|^2 dt < \varepsilon, \quad n = 1, 2, \dots$$

*Proof.* Recall that the  $\{z_i(x)\}$  satisfy

$$-\Delta z_i(x) = \lambda_i z_i(x),$$

$$\frac{\partial z_i}{\partial \mathbf{n}} \Big|_{\partial G} = 0.$$

It is well-known that the eigenvectors of the Laplacian are an orthonormal basis for  $L_2(G)$  and that the eigenvalues  $\lambda_i, i = 1, 2, 3, \dots$ , are positive and satisfy

$$\sum_{i=1}^{\infty} \frac{1}{\lambda_i} < \infty.$$

Let  $c_n^i(t) i = 1, 2, \dots$ , denote the Fourier coefficients of  $\varphi_n(x, t)$ . Then,

$$c_n^i(t) = \int_G \varphi_n(x, t) z_i(x) dx = -\frac{1}{\lambda_i} \int_G \varphi_n(x, t) \Delta z_i(x) dx$$

$$= \frac{1}{\lambda_i} \int_G \nabla_x \varphi_n(x, t) \cdot \nabla_x z_i(x) dx$$

$$\leq \frac{1}{\lambda_i} \|\nabla_x \varphi_n(\cdot, t)\|_{2,G} \|\nabla_x z_i(x)\|_{2,G} = \frac{1}{\sqrt{\lambda_i}} \|\nabla_x \varphi_n(\cdot, t)\|_{2,G}.$$

Therefore, by (4.27),

$$\int_0^T [c_n^i(t)]^2 dt \leq \frac{1}{\lambda_i} C.$$

For fixed  $\varepsilon > 0$ , there is  $N(\varepsilon)$  such that

$$C \sum_{i=N(\varepsilon)}^{\infty} \frac{1}{\lambda_i} \leq \varepsilon.$$

Hence,

$$\sum_{i=N(\varepsilon)}^{\infty} \int_0^T \left| \int_G \varphi_n(x, t) z_i(x) dx \right|^2 dt < \varepsilon, \quad n = 1, 2, \dots.$$

*Proof of Lemma 3.* It will be sufficient to show that

$$\int_G \alpha_m^{-1}(W_m)(W_m - v) dx \rightarrow 0$$

in  $L_2(0, T)$ . Let

$$\sum_{i=1}^{\infty} c_m^i(t) z_i(x)$$

denote the Fourier expansion of  $\alpha_m^{-1}(W_m(x, t))$  in  $L_2(G)$ . By (4.23) and the proposition, for each  $\varepsilon > 0$  there is  $N(\varepsilon)$  independent of  $m$  such that

$$\sum_{i=N(\varepsilon)}^{\infty} \int_0^T [c_m^i(t)]^2 \leq \varepsilon.$$

Hence,

$$\begin{aligned} & \int_0^T \left| \int_G \alpha_m^{-1}(W_m)(W_m - v) \, dx \right|^2 dt \\ &= \int_0^T \left| \int_G \sum_{i=1}^{N(\varepsilon)} c_m^i(t) z_i(x) \cdot (W_m - v) \, dx + \int_G \sum_{i>N(\varepsilon)} c_m^i(t) z_i(W_m - v) \right|^2 dt \\ &\leq \int_0^T \left[ \left[ \sum_{i=1}^{N(\varepsilon)} [c_m^i(t)]^2 \right]^{1/2} \left[ \sum_{i=1}^{N(\varepsilon)} [\beta_m^i(t) - \beta^i(t)]^2 \right]^{1/2} \right. \\ &\quad \left. + \left[ \sum_{i>N(\varepsilon)} [c_m^i(t)]^2 \right]^{1/2} \|(W_m - v)(\cdot, t)\|_{2,G} \right]^2 dt \\ &\leq 2\varepsilon \|W_m - v\|_{2,\infty,\Omega_T}^2 + 2\|\alpha_m^{-1}(W_m)\|_{2,\Omega_T}^2 \cdot \sup_{0 \leq t \leq T} \sum_{i=1}^{N(\varepsilon)} [\beta_m^i(t) - \beta^i(t)]^2. \end{aligned}$$

Since  $\beta_m^i(t) \rightarrow \beta^i(t)$  uniformly in  $t$ , and  $\|(W_m - v)(\cdot, t)\|_{2,G}^2, \|\alpha_m^{-1}(W_m)\|_{2,\Omega_T}$  are equibounded (see Lemmas 1 and 2), the result follows.

We select now a  $t \in [0, T]$  for which the result of Lemma 3 holds and fix it. Select any subsequence of the sequence for which the result of Lemma 3 holds. Now, the bound (4.23) implies the weak compactness of  $\{\alpha_m^{-1}(W_m(\cdot, t))\}$  in  $L_2(G)$ . Consequently, there exists a subsequence  $\{m_t\}$  of the original subsequence that was selected such that

$$\alpha_{m_t}^{-1}(W_{m_t}) \rightarrow \eta_t(x)$$

weakly in  $L_2(G)$ . We already know that  $W_m \rightarrow v$  weakly in  $L_2(G)$  and that the convergence is uniform in  $t$ .

LEMMA 4. *For this choice of  $t$ ,*

$$\int_G \alpha_{m_t}^{-1}(W_{m_t}) W_{m_t} \rightarrow \int_G \eta_t(x) v(x, t) \, dx$$

as  $m_t \rightarrow \infty$ .

*Proof.* We have

$$\begin{aligned} & \int_G \alpha_{m_t}^{-1}(W_{m_t}) W_{m_t} \, dx - \int_G \eta_t(x) v(x, t) \, dx \\ &= \int_G [\alpha_{m_t}^{-1}(W_{m_t}) - \eta_t] v(x, t) \, dx + \int_G \alpha_{m_t}^{-1}(W_{m_t})(W_{m_t} - v) \, dx \end{aligned}$$

The first integral converges to zero by the weak convergence of  $\alpha_{m_t}^{-1}(W_{m_t})$  to  $\eta_t(x)$ , and the second one converges to zero by virtue of Lemma 3.

LEMMA 5. *For this choice of  $t$ ,*

$$(4.28) \quad \eta_t(x) = \alpha^{-1}(v(x, t)).$$

*Proof.* For each  $m_t$  the function  $\alpha_{m_t}^{-1}(\cdot)$  is monotone increasing, and is a continuous map of  $L_2(G)$  into  $L_2(G)$ . If  $f(x) \in L_2(G)$  we have

$$(4.29) \quad \int_G [\alpha_{m_t}^{-1}(W_{m_t}) - \alpha_{m_t}^{-1}(f(x))] [W_{m_t} - f(x)] \, dx \geq 0.$$

Letting  $m_t \rightarrow \infty$  in (4.29) and using Lemma 4 and the uniform convergence of  $\alpha_m^{-1}$  to



$\alpha^{-1}$ , we obtain

$$(4.30) \quad \int_G [\eta_t(x) - \alpha^{-1}(f(x))][v(x, t) - f(x)] dx \geq 0.$$

Selecting  $f(x) = v(x, t) + \tau\psi(x)$  in (4.30) for arbitrary  $\psi(x) \in L_2(G)$ , we see that

$$(4.31) \quad \tau \int_G [\eta_t(x) - \alpha^{-1}(v + \tau\psi)]\psi dx \geq 0.$$

Dividing by  $\tau$  and letting  $\tau \rightarrow 0$ , we obtain

$$(4.32) \quad \int_G [\eta_t(x) - \alpha^{-1}(v(x, t))]\psi(x) dx \geq 0$$

for all  $\psi(x) \in L_2(G)$ . This proves the lemma.

Thus, we have shown that for each  $t$  for which the result of Lemma 3 holds, every subsequence selected from the sequence for which the result of Lemma 3 holds contains a subsequence  $\{m_i\}$  such that  $\alpha_{m_i}^{-1}(W_{m_i}) \rightarrow \alpha^{-1}(v(x, t))$  weakly in  $L_2(G)$ . Consequently, relabeling the subsequence of Lemma 2, we see that the entire sequence  $\alpha_m^{-1}(W_m(x, t)) \rightarrow \alpha^{-1}(v(x, t))$  weakly in  $L_2(G)$  for a.e.  $t \in [0, T]$ . From this result and the bound (4.23) we can obtain the strong  $L_2(\Omega_T)$  convergence of  $\alpha_m^{-1}(W_m)$  to  $\alpha^{-1}(v(x, t))$  as follows.

LEMMA 6. *The sequence  $\alpha_m^{-1}(W_m)$  converges strongly to  $\alpha^{-1}(v(x, t))$  in  $L_2(\Omega_T)$ .*

*Proof.* Set  $Z_m(x, t) = \alpha_m^{-1}(W_m(x, t)) - \alpha^{-1}(v(x, t))$ . By Lemma 2,  $Z_m \in L_2(G)$  for all  $t \in [0, T]$ . For each  $\varepsilon > 0$ , Friedrich's lemma [6, p. 72] implies the existence of a positive integer  $M(\varepsilon)$  and functions  $\varphi_1, \dots, \varphi_{M(\varepsilon)} \in L_2(G)$  with  $\|\varphi_i\|_{L_2(G)} = 1$  such that

$$\|Z_m(\cdot, t)\|_{L_2(G)}^2 \leq \varepsilon [\|Z_m(\cdot, t)\|_{L_2(G)}^2 + \|\nabla_x Z_m(\cdot, t)\|_{L_2(G)}^2] + \sum_{j=1}^{M(\varepsilon)} (Z_m(\cdot, t), \varphi_j)_{L_2(G)}^2$$

holds for all  $m$  and almost all  $t \in [0, T]$ . Note that the  $\varphi_j$  are independent of  $m$  and  $t$ .

Integrating the above inequality over  $[0, T]$  we obtain

$$(4.33) \quad \|Z_m\|_{L_2(\Omega_T)}^2 \leq \varepsilon [\|Z_m\|_{L_2(\Omega_T)}^2 + \|\nabla_x Z_m\|_{L_2(\Omega_T)}^2] + \sum_{j=1}^{M(\varepsilon)} \int_0^T (Z_m(\cdot, t), \varphi_j)_{L_2(G)}^2 dt.$$

From (4.23), it follows that  $\|Z_m\|_{W_2^1(\Omega_T)}$  are equibounded while from  $(Z_m(\cdot, t), \varphi_j)^2 \rightarrow 0$  for almost all  $t \in [0, T]$  it follows that we can use the Lebesgue dominated convergence theorem to pass to the limit under integral. Consequently, we see that there is a constant  $C$  such that

$$\lim_{m \rightarrow \infty} \|Z_m\|_{L_2(\Omega_T)} \leq \varepsilon C$$

for each  $\varepsilon > 0$ . Hence, the proof is finished.

We now conclude the proof of the theorem. From (4.23) and (4.5) we see that for all  $m$

$$\|\nabla_x k_m[\alpha_m^{-1}(W_m)]\|_{L_2(\Omega_T)} \leq C.$$

Therefore, we may incorporate into the diagonalization process

$$\lim_{m \rightarrow \infty} \iint_{\Omega_T} \varphi \frac{\partial}{\partial x_i} k_m[\alpha_m^{-1}(W_m)] dx dt = \iint_{\Omega_T} \varphi A_i$$

for all  $\varphi \in L_2(\Omega_T)$ . Since for all  $\varphi \in C_0^\infty(\Omega_T)$  we have

$$\iint_{\Omega_T} \varphi \frac{\partial}{\partial x_i} k_m[\alpha_m^{-1}(W_m)] dx dt = - \iint_{\Omega_T} k_m[\alpha_m^{-1}(W_m)] \frac{\partial \varphi}{\partial x_i} dx dt,$$

we see that the square integrable  $\Lambda_i$  equals the distributional derivative of  $k[\alpha^{-1}(v)]$  with respect to  $x_i$ . From a lemma of Stampacchia [8] we see that

$$\Lambda_i = \frac{\partial}{\partial x_i} k[\alpha^{-1}(v)] = k'[\alpha^{-1}(v)] \cdot \frac{\partial}{\partial x_i} [\alpha^{-1}(v)].$$

Consider the Galerkin approximate  $W_m$ . For

$$\varphi_N = \sum_{i=1}^N \mu_i(t) z_i(x)$$

where  $\mu_i \in C^1([0, T])$  and  $\mu_i(T) = 0, i = 1, \dots, N$ , we see that for all  $m > N, W_m$  satisfies

$$\begin{aligned} & \iint_{\Omega_T} \left\{ W_m \frac{\partial}{\partial t} \varphi_N - \nabla_x k_m[\alpha_m^{-1}(W_m)] \cdot \nabla_x \varphi_N - \mathbf{b}(x, t, \alpha_m^{-1}(W_m)) \cdot \nabla_x \varphi_N \right. \\ (4.34) \quad & \left. + c(x, t, \alpha_m^{-1}(W_m)) \varphi_N \right\} dx dt + \int_{G(0)} \alpha_m(h) \varphi_N dx + \int_{S_T} g(x, t, \alpha_m^{-1}(W_m)) \varphi_N d\sigma = 0. \end{aligned}$$

Consider the traces of  $\alpha_m^{-1}(W_m)$  and  $\alpha^{-1}(v)$  on  $S_T$ . For each  $\varepsilon > 0$ , there exists a constant  $K(\varepsilon)$  such that [see [7]],

$$\begin{aligned} & \|\alpha^{-1}(v) - \alpha_m^{-1}(W_m)\|_{2,S_T}^2 \leq \varepsilon \|\nabla_x(\alpha^{-1}(v) - \alpha_m^{-1}(W_m))\|_{2,\Omega_T}^2 \\ (4.35) \quad & + K(\varepsilon) \|\alpha^{-1}(v) - \alpha_m^{-1}(W_m)\|_{2,\Omega_T}^2. \end{aligned}$$

By (4.23),

$$(4.36) \quad \|\alpha^{-1}(v) - \alpha_m^{-1}(W_m)\|_{2,S_T}^2 \leq 4C\varepsilon + K(\varepsilon) \|\alpha^{-1}(v) - \alpha_m^{-1}(W_m)\|_{2,\Omega_T}^2.$$

By the strong convergence of  $\alpha_m^{-1}(W_m)$  to  $\alpha^{-1}(v)$  in  $L_2(\Omega_T)$ , we see that the second term on the right of (4.36) can be made arbitrarily small for  $m$  large enough. Hence, the trace of  $\alpha_m^{-1}(W_m)$  on  $S_T$  converges strongly to the trace of  $\alpha^{-1}(v)$  in  $L_2(S_T)$ . Consequently, from this discussion and (A1) we have

$$(4.37) \quad \lim_{m \rightarrow \infty} \int_{S_T} \varphi_N g(x, t, \alpha_m^{-1}(W_m)) d\sigma = \int_{S_T} \varphi_N g(x, t, \alpha^{-1}(v)) d\sigma.$$

Taking the limit as  $m \rightarrow \infty$  in (4.34) now yields

$$\begin{aligned} & \iint_{\Omega_T} \left\{ v \frac{\partial \varphi_N}{\partial t} - \nabla_x k(\alpha^{-1}(v)) \cdot \nabla_x \varphi_N - \mathbf{b}(x, t, \alpha^{-1}(v)) \cdot \nabla_x \varphi_N + c(x, t, \alpha^{-1}(v)) \varphi_N \right\} dx dt \\ (4.38) \quad & + \int_{G(0)} \varphi_N \alpha(h) dx + \int_{S_T} \varphi_N g(x, t, \alpha^{-1}(v)) d\sigma = 0 \end{aligned}$$

for all finite  $N$ . Since the  $\varphi_N$  are dense in  $W_2^{1,1}(\Omega_T)$  it follows that (4.38) holds true for all  $\varphi \in W_2^{1,1}(\Omega_T)$  such that  $\varphi(x, T) \equiv 0$ . From an argument of [6, pp. 156–157] we deduce that  $v \in V_2^{1,0}(\Omega_T)$ .

Setting now  $u = \alpha^{-1}(v)$ , the continuity of  $\alpha^{-1}(\cdot)$  implies that  $u \in V_2^{1,0}(\Omega_T)$ , while it is apparent that  $v \subset \alpha(u)$ . Hence a weak-solution of (1.1)–(1.5) exists.

**5. The Dirichlet problem.** If the condition (1.3) is replaced by

$$(1.3') \quad u_i = f_i \quad \text{on } S_i - \Gamma$$

then (1.1), (1.2), (1.3'), (1.4), (1.5) is a Dirichlet problem. Let  $\Psi$  denote a smooth function in  $\Omega_T$  which coincides with the Dirichlet boundary data induced by (1.3') when (1.1)–(1.5) is reduced to its weak formulation. Setting  $u = v + \Psi$ , creates a variational problem for  $v$  of the form (2.2) without the integral term over  $S_T$ . The test functions employed here are those  $\varphi \in W_2^{1,1}(\Omega_T)$  that vanish on  $S_T$  and  $G(T)$ . The definition of weak solution is analogous to the one given in § 2. The Galerkin method can be applied to this problem in a manner similar to the application in § 4. Thus, we conclude that the Dirichlet problem (1.1), (1.2), (1.3'), (1.4), (1.5) possesses a weak solution  $u \in V_2^{1,0}(\Omega_T)$ .

#### REFERENCES

- [1] J. R. CANNON AND C. DENSON HILL, *On the movement of a chemical reaction interface*, Indiana Univ. Math. J., 20 (1970), pp. 429–454.
- [2] J. R. CANNON, W. T. FORD AND A. V. LAIR, *Quasilinear parabolic systems*, J. Differential Equations, 20 (1976), pp. 441–472.
- [3] J. R. CANNON AND RICHARD EWING, *Quasilinear parabolic systems with nonlinear boundary conditions*, to appear.
- [4] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 132 (1968), pp. 51–87.
- [5] ———, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [6] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Mono., 23, American Mathematical Society, Providence, RI, 1968.
- [7] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary value Problems and applications*, I, Springer-Verlag, New York, 1972.
- [8] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble) (1), 15 (1965), pp. 189–258.

## CONCERNING THE EIGENVALUES OF $DA$ UNDER VARIATIONS OF THE ENTRIES OF $D$ AND $A^*$

D. J. HARTFIEL†

**Abstract.** Let  $A$  be a positive definite matrix and  $D$  a nonsingular, nonnegative diagonal matrix. This paper is a study of the eigenvalues of  $DA$  under variations between zero and infinity of specified entries of  $D$  and  $A$ . Intervals containing these eigenvalues are determined and these intervals are seen to be the best possible.

**Introduction.** The differential equation

$$D\ddot{x} + Ax = 0,$$

where  $D$  is a diagonal matrix with positive main diagonal and  $A$  is positive definite Hermitian, describes an elastic system such as the spring-mass system. The behavior of the system is, in part, mathematically determined by the generalized eigenvalues of  $D\lambda + A$ , which are the eigenvalues of  $-D^{-1}A$ . Variational studies on these eigenvalues have been made by Rayleigh, Courant, Weyl (see [1]) as well as many others [4].

Recently, the author [2] considered the differential equation of price stability of multiple markets

$$\dot{x} = DAx,$$

where  $D = \text{diag}(d_1, \dots, d_n)$ ; each  $d_i > 0$  representing the speed of adjustment to which the  $i$ th market responds to a discrepancy between demand and supply, and where  $A$  is an  $m$ -matrix due to the assumption of all goods being gross substitutes. The study considered how  $D$  affects the eigenvalues of  $DA$  and, hence, how speeds of adjustment affect return speeds of prices, given that there is a discrepancy between demand and supply.

The present paper considers the analogous problem for the elastic equation above. The point of interest in this study is how the entries in  $D$  affect the eigenvalues of  $DA$ . Also of interest is how changes in  $A$  affect the eigenvalues of  $DA$ . Thus this work adds to the variational study of differential equations describing elastic systems.

Throughout the paper, let  $n$  be a positive integer, and  $A = (a_{ij})$  an  $n \times n$  positive definite Hermitian matrix. Let  $D = \text{diag}(d_1, \dots, d_n)$  be an  $n \times n$  nonnegative, nonsingular matrix. The study of the eigenvalue behavior of  $DA$ , as selected entries in  $D$  are allowed to vary between zero and infinity, will be given in § 1. As the eigenvalues of  $DA$  are as those of  $EAE$ , where  $E = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})$ , it follows that  $DA$  always has positive eigenvalues. Intervals containing these eigenvalues will be determined. Section 2 concerns the study of the eigenvalue behavior of  $A$  under variations of its entries. Here we first consider varying the main diagonal entries of  $A$ , that is we study  $A + \mathcal{D}$ , where  $\mathcal{D} = \text{diag}(\partial_1, \dots, \partial_n)$  is a nonnegative diagonal matrix. Intervals are found which contain the eigenvalues of  $A + \mathcal{D}$ , as selected entries of  $\mathcal{D}$  are allowed to vary from zero to infinity. The paper is then concluded by showing how some of the results presented in § 1 and § 2 are applied to physical situations.

**1. The study for variations in  $D$ .** We begin this study by considering the monotone behavior of the eigenvalues of  $DA$ . This work utilizes the following notation: Let  $B$  be

\* Received by the editors September 5, 1978 and in revised form September 24, 1979.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

an  $n \times n$  matrix with real eigenvalues. We denote the list of these eigenvalues as

$$\alpha_n(B) \leq \alpha_{n-1}(B) \leq \dots \leq \alpha_1(B).$$

Using this notation, we have the following

**THEOREM 1.1.** *If  $\min_i d_i \geq 1$ , then  $\alpha_k(A) \leq \alpha_k(DA)$  for  $k = 1, \dots, n$ .*

*Proof.* For  $k = 1$ , Rayleigh's principle [1, p. 141] yields that  $\alpha_1(DA) = \max_{|x|=1} x^*EAE x$ , where  $E = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})$ . From this,

$$\alpha_1(DA) = \max_{|E^{-1}y|=1} y^*Ay.$$

Factoring  $A = U^*FU$ , where  $U$  is a unitary matrix and  $F$  a diagonal matrix yields that

$$\alpha_1(DA) = \max_{|E^{-1}y|=1} y^*U^*FUy = \max_{|E^{-1}U^*w|=1} w^*Fw.$$

Now, since  $|E^{-1}(U^*w)| = 1$  implies that  $|U^*w| \geq 1$ , and hence,  $|w| \geq 1$ , it follows that

$$\alpha_1(DA) \geq \max_{|w|=1} w^*Fw = \alpha_1(A).$$

For  $k > 1$ , the minimax theorem [1, p. 146] yields that

$$\begin{aligned} \alpha_k(DA) &= \min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |x|=1}} x^*EAE x \right) \\ &= \min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(w, Ey_1^{-1})=0 \\ \vdots \\ (w, Ey_{k-1}^{-1})=0}} w^*Aw \right) \\ &= \max_{z_1, \dots, z_{k-1}} \left( \max_{\substack{(w, z_1)=0 \\ \vdots \\ (w, z_{k-1})=0 \\ |E^{-1}w|=1}} w^*Aw \right) \\ &= \min_{z_1, \dots, z_{k-1}} \left( \max_{\substack{(U^*x, z_1)=0 \\ \vdots \\ (U^*x, z_{k-1})=0 \\ |E^{-1}U^*x|=1}} x^*Fx \right) \\ &= \min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |E^{-1}U^*x|=1}} x^*Fx \right). \end{aligned}$$

Now as  $|E^{-1}U^*x| = 1$ , it follows that  $|U^*x| \geq 1$ , and hence,  $|x| \geq 1$ . Thus

$$\begin{aligned} \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |E^{-1}U^*x|=1}} x^*Fx &\geq \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |x|=1}} x^*Fx. \end{aligned}$$

Hence  $\alpha_k(DA) \geq \alpha_k(A)$ .  $\square$

Using this theorem we can now establish the range of the eigenvalues of  $DA$  when certain entries of  $D$  are allowed to vary from zero to infinity.

**THEOREM 1.2.** *Let  $d_1, \dots, d_r$  vary between zero and infinity and fix  $d_{r+1}, \dots, d_n$ .*

*Let  $L = \text{diag}(d_{r+1}, \dots, d_n)$ ,  $L^{1/2} = \text{diag}(\sqrt{d_{r+1}}, \dots, \sqrt{d_n})$  and partition  $A = \begin{pmatrix} P & Q \\ Q^* & S \end{pmatrix}$  where  $P$  is  $r \times r$ . Then*

- (i) *for  $i = n - r + 1, \dots, n$ ,  $0 < \alpha_i(DA)$ ;*
- (ii) *for  $i = 1, \dots, n - r$ ,  $\alpha_i(LS) = a_i \leq \alpha_i(DA)$ ;*
- (iii) *for  $i = 1, \dots, r$ ,  $\alpha_i(DA) < \infty$ ; and*
- (iv) *for  $i = r + 1, \dots, n$ ,  $\alpha_i(DA) \leq b_i = \alpha_{i-r}[L(S - Q^*P^{-1}Q)]$ ,*

*where all numbers in these intervals are achieved with the possible exception of  $a_i$  and  $b_i$ . The number  $a_i$  (similarly  $b_i$ ) is achieved if and only if  $a_i$  (similarly  $b_i$ ) is an eigenvalue of  $L^{1/2}SL^{1/2}$  and if  $a_i = a_{i+1} = \dots = a_{i+s-1} > a_{i+s}$  (similarly  $b_i = b_{i-1} = \dots = b_{i-s+1} < b_{i-s}$ ) then  $a_i$  (similarly  $b_i$ ) is of multiplicity at least  $s$  having at least  $s$  linearly independent eigenvectors each of which is in  $\ker Q$ .*

*Proof.* Let  $E$  be the nonnegative diagonal matrix such that  $E^2 = D$ . Then the eigenvalues of  $DA$  are those of  $EAE$ . Now if  $d_1, \dots, d_r$  approach zero,  $EAE$  approaches

$$\begin{pmatrix} 0 & 0 \\ 0 & L^{1/2}SL^{1/2} \end{pmatrix}.$$

Hence

- (i)  $0 < \alpha_i(DA)$  for  $i = n - r + 1, \dots, n$ , while
  - (ii)  $\alpha_i(LS) = a_i \leq \alpha_i(DA)$  for  $i = 1, \dots, n - r$ .
- Further, as  $d_1, \dots, d_r$  approach infinity,  $E^{-1}A^{-1}E^{-1}$  approaches

$$\begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}, \quad \text{where } C = L^{-1/2}(S - Q^*P^{-1}Q)^{-1}L^{-1/2}.$$

Thus

- (iii)  $\alpha_i(DA) < \infty$  for  $i = 1, \dots, r$ , while
- (iv)  $\alpha_i(DA) \leq b_i = \alpha_{i-r}[L(S - Q^*P^{-1}Q)]$  for  $i = r + 1, \dots, n$ .

Further, the above arguments show that, with the possible exception of the end points  $a_i$  for  $i = 1, \dots, n - r$  and  $b_i$  for  $i = r + 1, \dots, n$ , all numbers in the  $i$ th interval are achieved by  $\alpha_i(DA)$  for  $i = 1, \dots, n$ .

We now consider when the boundary points  $a_i$  for  $i = 1, \dots, n - r$  and  $b_i$  for  $i = r + 1, \dots, n$  are achieved. For this, suppose first that  $\alpha_k(DA)$  achieves its upper bound  $b_k$  for some  $k \geq r + 1$ . Then, by the minimax theorem,

$$\min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(x, y_1) = 0 \\ \vdots \\ (x, y_{k-1}) = 0 \\ |x| = 1}} x^*EAE x \right) = b_k$$

for all  $d_1, \dots, d_r$  sufficiently large. Hence, by replacing  $Ex$  by  $w$ , we have

$$\max_{z_1, \dots, z_{k-1}} \left( \max_{\substack{(w, z_1) = 0 \\ \vdots \\ (w, z_{k-1}) = 0 \\ |E^{-1}w| = 1}} w^*Aw \right) = b_k$$

for all  $d_1, \dots, d_r$  sufficiently large.

Make two selections of values for  $d_1, \dots, d_r$  yielding diagonal matrices  $E_1(e_{ij}^{(1)})$  and  $E_2 = (e_{ij}^{(2)})$  so that  $e_{ii}^{(1)} e_{ii}^{(2)}$  for all  $i$ , and so that the above equation holds with  $E$  replaced by  $E_1$  and by  $E_2$ . Since

$$\begin{array}{ccc} \max_{\substack{(w, z_1)=0 \\ \vdots \\ (w, z_{k-1})=0 \\ |E_1^{-1}w|=1}} w^*Aw & \cong & \max_{\substack{(w, z_1)=0 \\ \vdots \\ (w, z_{k-1})=0 \\ |E_2^{-1}w|=1}} w^*Aw \end{array}$$

for all choices  $z_1, \dots, z_{n-1}$  with equality holding for some choices  $z_1, \dots, z_{n-1}$  so that  $b_k$  is achieved, it follows that, for such choices, the max is achieved at some  $w$  so that  $|E_1^{-1}w| = |E_2^{-1}w| = 1$ , and hence,  $w_1 = \dots = w_r = 0$ . From the proof of the minimax theorem it follows that  $y = E_1^{-1}w$  is an eigenvector for  $E_1AE_1$  and  $E_2AE_2$  for the eigenvalue  $b_k$ . Hence, by direct calculation,  $y$  is an eigenvector for  $EAE$  for all choices of  $d_1, \dots, d_r$  with corresponding eigenvalue  $b_k$ . Thus  $b_k$  is an eigenvalue of  $L^{1/2}SL^{1/2}$  with a corresponding eigenvector  $v$  so that  $Qv = 0$ . Further, if  $b_k = b_{k-1} = \dots = b_{k-s+1} < b_{k-s}$ , then  $b_k$  must have multiplicity at least  $s$  and having at least  $s$  linearly independent eigenvectors in  $\ker Q$ .

Conversely, suppose  $b_k = b_{k-1} = \dots = b_{k-s+1} < b_{k-s}$  and  $b_k$  is an eigenvalue of  $L^{1/2}SL^{1/2}$  having at least  $s$  linearly independent eigenvectors in  $\ker Q$ . Then  $b_k$  is an eigenvalue for  $EAE$  for all choices of  $d_1, \dots, d_r$  so that for each  $D$ , there is some  $i \leq k$  so that  $b_k = \alpha_i(DA) = \alpha_{i-1}(DA) = \dots = \alpha_{i-s+1}(DA) < \alpha_{i-s}(DA)$ . As  $\alpha_j(DA)$  is non-decreasing in  $D$  for  $j = 1, \dots, n$ , it follows that  $b_k$  is an upper bound for some  $\alpha_i(DA), \alpha_{i-1}(DA), \dots, \alpha_{i-s+1}(DA)$  and this bound is achieved.

The proof concerning the bound  $a_i$  is argued similarly and hence omitted.  $\square$

Although intervals containing the eigenvalues of  $DA$  can be specified, it is not possible, in general, to specify the list of these eigenvalues. To see this, let  $d = (d_1, \dots, d_n)$  and consider the quotient

$$q(d) = \frac{\alpha_n(DA)}{\alpha_1(DA)}.$$

Since  $q(d) = q(\alpha d)$  for any positive number  $\alpha$ , we may assume  $d_1 = 1$ . With this assumption, and the inclusion principle [1, p. 149],

$$\lim_{\mu \rightarrow 0} q(d) = 0 \quad \text{and} \quad \lim_{l \rightarrow \infty} q(d) = 0,$$

where  $\mu = \max_{i>1} d_i$  and  $l = \min_{i>1} d_i$ . Hence,  $q(d)$  achieves a maximal value for some  $\hat{d} > 0$ . Further,  $q(\hat{d}) \leq 1$ .

Now, if  $A$  is real and tridiagonal, with  $a_{ii+1} \neq 0$  for  $i = 1, \dots, n$  then  $DA$  has distinct eigenvalues for all nonsingular nonnegative diagonal matrices  $D$  [3, p. 166]. Thus,  $q(\hat{d}) < 1$ . From this it follows that if we pick two positive numbers  $\epsilon_1$  and  $\epsilon_2$  so that  $q(\hat{d}) < \epsilon_1/\epsilon_2 < 1$ , then there is no  $D$  so that  $\alpha_n(DA) = \epsilon_1$  and  $\alpha_1(DA) = \epsilon_2$ .

This concludes our work on the eigenvalues of  $DA$  under variations of certain entries of  $D$ .

**2. The study for variations in  $A$ .** In this section we study how varying the entries of  $A$  affect the eigenvalues of  $A$ . The initial work concerns the varying of the main diagonal entries of  $A$ . Our result in this regard follows:

**THEOREM 2.1.** *Let  $\partial_1, \dots, \partial_r$  vary between zero and infinity and fix  $\partial_{r+1}, \dots, \partial_n$  at zero. Partition  $A = \begin{pmatrix} P & Q \\ Q^* & S \end{pmatrix}$ , where  $P$  is  $r \times r$ . Then*

(i) for  $i = r + 1, \dots, n$ ;  $\alpha_i(A) = a_i \leq \alpha_i(A + \mathcal{D}) \leq b_i = \alpha_{i-r}(S)$ , where all numbers in this interval are achieved with the possible exception of  $b_i$  and  $a_i$ . The number  $b_i$  is achieved if and only if  $b_i$  is an eigenvalue of  $A$  and if  $b_i = b_{i+1} = \dots = b_{i-s+1} < b_{i-s}$  then  $b_i$  has multiplicity at least  $s$  with at least  $s$  linearly independent eigenvectors having their first  $r$  components zero;

(ii) for  $i = 1, \dots, r$ ;  $a_i = \alpha_i(A) \leq \alpha_i(A + \mathcal{D}) < \infty$ , where all numbers in this interval are achieved with the possible exception of  $a_i$ ;

(iii) the bound  $a_i$  in (i) and (ii) is achieved if and only if  $a_i$  is an eigenvalue of  $A$  and if  $a_i = a_{i-1} = \dots = a_{i-s+1} < a_{i-s}$ . Then  $A$  has at least  $s$  linearly independent eigenvectors having their first  $r$  components zero.

*Proof.* First write

$$A + \mathcal{D} = \begin{pmatrix} P + \mathcal{D}_1 & Q \\ Q^* & S \end{pmatrix},$$

where  $P + \mathcal{D}_1$  is  $r \times r$ . Then, using the partitioned form of the inverse, we obtain

$$(A + \mathcal{D})^{-1} = \begin{pmatrix} X & -(P + \mathcal{D}_1)^{-1}QW \\ -WQ^*(P + \mathcal{D}_1)^{-1} & W \end{pmatrix},$$

where  $X = (P + D_1 - QS^{-1}Q^*)^{-1}$  and  $W = (S - Q^*(P + \mathcal{D}_1)^{-1}Q)^{-1}$ . Thus, by applying the adjoint formula for the inverse, it follows that as  $\partial_1, \dots, \partial_r$  approach infinity,  $X$  approaches 0 and  $(P + \mathcal{D}_1)^{-1}$  approaches 0. Hence,  $(A + \mathcal{D})^{-1}$  approaches

$$\begin{pmatrix} 0 & 0 \\ 0 & S^{-1} \end{pmatrix}.$$

Now by Weyl's inequality [1, p. 157], the eigenvalues of  $A + \mathcal{D}$  are nondecreasing in  $\mathcal{D}$  and so (ii) and the first part of (i) follow.

Consider now the boundary point  $b_i$  in (i). For this, suppose this boundary point is achieved. Then, there are numbers  $d_1, \dots, d_r$  so that if  $\partial_1 > d_1, \dots, \partial_r > d_r$  then

$$\alpha_i(A + D) = \alpha_i(A + \mathcal{D})$$

where  $D = \text{diag}(d_1, \dots, d_r, 0, \dots, 0)$ . Now write  $\mathcal{D} = D + R$ . Then applying the minimax theorem we have

$$\min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |x|=1}} x^*(A + D)x \right) = \min_{y_1, \dots, y_{k-1}} \left( \max_{\substack{(x, y_1)=0 \\ \vdots \\ (x, y_{k-1})=0 \\ |x|=1}} [x^*(A + D)x + x^*Rx] \right).$$

This minimum is achieved at an eigenvector, belonging to the eigenvalue  $b_i$ , which must, by the above equality, be of the form  $w$  where  $w_1 = \dots = w_r = 0$ . Hence, by direct calculation,  $w$  is an eigenvector, belonging  $b_i$ , of  $A$ . Further, if  $b_i = b_{i-1} = \dots = b_{i-s+1} < b_{i-s}$  then  $b_i$  must be of multiplicity at least  $s$  with at least  $s$  linearly independent eigenvectors having their first  $s$  components zero.

Conversely, if  $b_i = b_{i-1} = \dots = b_{i-s+1} < b_{i-s}$  and  $b_i$  is an eigenvalue of  $A$  having multiplicity at least  $s$  with at least  $s$  linearly independent eigenvectors having their first  $r$  components zero, then for each  $\mathcal{D}$  there is a  $k$  so that  $b_i = \alpha_k(A + \mathcal{D})$ . Now as the eigenvalues of  $A + \mathcal{D}$  are nondecreasing in  $d_1, \dots, d_r$ , it follows that  $b_i$  is an upper bound for

$$\alpha_k(A + \mathcal{D}), \alpha_{k-1}(A + \mathcal{D}), \dots, \alpha_{k-s+1}(A + \mathcal{D}),$$

for some  $k$ , and this bound is achieved.

The bound  $a_i$  is argued in a similar way and hence is omitted.  $\square$



From the above results, it is seen that intervals containing the eigenvalues of  $A + \mathcal{D}$  can be determined. However, it is not, in general, possible to specify these eigenvalues.

For example,  $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} + \begin{pmatrix} d & 0 \\ 0 & 0 \end{pmatrix}$  has eigenvalues  $((4+d) \pm \sqrt{(4+d)^2 - 4(4+2d-1)})/2$  and so only one eigenvalue can be specified.

Specified eigenvalues can, of course, be achieved by allowing all of the entries of  $A$  to vary. For our next result we show that this can, in fact, be done when  $A$  is real and tridiagonal. The theorem will also provide a converse to the known theorem if  $A$  is a real tridiagonal matrix with  $a_{ii+1}a_{i+1,i} \neq 0$  for  $i = 1, \dots, n-1$ , then  $A$  has distinct eigenvalues [3, p. 166].

**THEOREM 2.2.** *Let  $\lambda_1 > \dots > \lambda_n > 0$  be a list of numbers. Then there is a real symmetric tridiagonal matrix  $A$ , with arbitrarily small  $a_{ii-1} > 0$  for  $i = 2, \dots, n$  and  $a_{11} > \dots > a_{nn}$ , having its list of eigenvalues  $\lambda_1, \dots, \lambda_n$ .*

*Proof.* First note that if  $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ ,  $Q = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$  and  $B = Q^t A Q$  then

$$\begin{aligned} b_{11} &= a_{11} \cos^2 \theta + 2a_{12} \sin \theta \cos \theta + a_{22} \sin^2 \theta, \\ b_{12} &= a_{12} \cos 2\theta + (a_{22} - a_{11})(\sin 2\theta)/2, \\ b_{22} &= a_{22} \cos^2 \theta - 2a_{12} \sin \theta \cos \theta + a_{11} \sin^2 \theta. \end{aligned}$$

The proof now proceeds by an induction on  $n$ .

If  $n = 1$ , there is nothing to prove. Thus, suppose the theorem holds for every list of  $n$  distinct positive numbers where  $n < n_1$ . Now let  $\lambda_1 > \dots > \lambda_n$  be a list of  $n = n_1$  distinct numbers. By the induction hypothesis there is a symmetric tridiagonal matrix  $B$ , having arbitrarily small  $b_{ii-1} > 0$  with  $b_{11} > b_{22} > \dots > b_{n-1n-1}$ , and having eigenvalues  $\lambda_1, \dots, \lambda_{n-1}$ . Set  $T = B \oplus (\lambda_n)$ . Note that  $B$  can be chosen so that  $b_{11} > \dots > b_{n-1n-1} > \lambda_n$ . Define an  $n \times n$  plane rotation as  $P = P(i, j, \theta)$ , where  $p_r$ , the  $r$ th row of  $P$ , is defined as

$$P_r = \begin{cases} e_r, \text{ the } r\text{th unit vector} & \text{for } r \notin \{i, j\}, \\ (\cos \theta)e_i + (-\sin \theta)e_j & \text{for } r = i, \\ (\sin \theta)e_i + (\cos \theta)e_j & \text{for } r = j. \end{cases}$$

Pick  $P_1 = P(n-1, n, \theta_1)$ , with  $-\pi/4 < \theta_1 < 0$  so that  $P_1^t A P_1 = B_1 = (b_{ij}^{(1)})$ , where  $b_{nn-1}^{(1)} = (\lambda_n - b_{n-1n-1})(\sin 2\theta_1)/2 > 0$ . Then  $b_{nn-2}^{(1)} = (\sin \theta_1)a_{n-1n-2} < 0$ . Note that  $\lim_{\theta \rightarrow 0} B_1 = T$ , so  $\theta_1$  can be chosen sufficiently small so that  $b_{11}^{(1)} > \dots > b_{nn}^{(1)}$ .

Now let  $P_2 = P_2(n-2, n-1, \theta_2)$  and set  $P_2^t B_1 P_2 = B_2 = (b_{ij}^{(2)})$ . Take  $\theta_2$  so that  $b_{nn-2}^{(2)} = (\cos \theta_2)b_{nn-2}^{(1)} - (\sin \theta_2)b_{nn-1}^{(1)} = 0$ , i.e.,

$$\tan \theta_2 = \frac{b_{nn-2}^{(1)}}{b_{nn-1}^{(1)}} < 0.$$

Thus  $\pi/4 < \theta_2 < 0$  and so  $b_{n-2n-1}^{(2)} = b_{n-2n}^{(1)} \sin \theta_2 + b_{n-1n}^{(1)} \cos \theta_2 > 0$ . Consider then

$$b_{n-1n-2}^{(2)} = b_{n-1n-2}^{(1)} \cos 2\theta_2 + (b_{n-1n-1}^{(1)} - b_{n-2n-2}^{(1)})(\sin 2\theta_2)/2.$$

As  $\tan 2\theta_2 < 0$  and  $(-2b_{n-1n-2}^{(1)})/(b_{n-1n-1}^{(1)} - b_{n-2n-2}^{(1)}) > 0$  it follows that  $b_{n-1n-2}^{(2)} \neq 0$ . Note again that  $\lim_{\theta \rightarrow 0} B_2 = T$ , hence  $\theta_1$  can be chosen sufficiently small so that  $b_{n-1n-2}^{(2)} > 0$  and  $b_{11}^{(2)} > \dots > b_{nn}^{(2)} > 0$ .

This argument can now be repeated to obtain the matrix of the theorem.  $\square$

By applying real diagonal orthogonal transformations to  $A$ , we have the following:

**COROLLARY 2.1.** *Let  $\lambda_1 > \dots > \lambda_n > 0$  be a list of numbers. Then there is a real symmetric tridiagonal matrix  $A$ , with arbitrarily small  $a_{ii-1} < 0$  for  $i = 2, \dots, n$  and  $a_{11} > \dots > a_{nn}$ , having its list of eigenvalues  $\lambda_1, \dots, \lambda_n$ .*

This then concludes the work of this section. In the final section, some application of these results is given.

**3. Some application.** Consider an elastic system, such as a spring-mass system, as described by the matrix differential equation

$$D\ddot{x} + Ax = 0,$$

where

$$A = \begin{pmatrix} a_{11} + a_{12} & -a_{12} & 0 & \cdots & 0 & 0 \\ -a_{12} & a_{12} + a_{23} & -a_{23} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -a_{nn-1} & a_{nn-1} \end{pmatrix},$$

with  $a_{ii} > 0$  and  $a_{ii+1} > 0$  for all  $i$ . Thus, if  $e = (1, 1, \dots, 1)^t$  then  $Ae = (\rho_1, 0, \dots, 0)^t$ , where  $\rho_1$  is positive [1, p. 104].

First we note that given numbers  $\lambda_1 > \dots > \lambda_n > 0$ , there need not be a matrix, having the above form, with these numbers as eigenvalues. To see this, consider

$$A = \begin{pmatrix} a + b & -a \\ -a & a \end{pmatrix},$$

where  $a$  and  $b$  are positive. If we let  $\lambda_1 = 100$  and  $\lambda_2 = 101$ , then  $2a + b = 201$ , while  $ab = 10100$ . Further, by the inclusion principle,

$$\lambda_1 \leq a < a + b \leq \lambda_2.$$

This implies that  $0 < b \leq 1$ . But now  $ab \leq 100$ , a contradiction. Thus there is no matrix, with the above form, having  $\lambda_1 = 100$  and  $\lambda_2 = 101$  as eigenvalues.

As a consequence we see that the spring-mass system

$$\ddot{x} + Ax = 0$$

can not have arbitrary generalized eigenvalues. Although this is the case, we will show that the spring-mass system

$$D\ddot{x} + Ax = 0$$

can have arbitrary generalized eigenvalues.

**THEOREM 3.1.** *Let  $\lambda_1 > \dots > \lambda_n$  be a list of numbers. Then there is a real tridiagonal matrix  $A$ , with  $a_{ii-1} < 0$  for  $i = 1, \dots, n$ ,  $Ae = (\rho, 0, \dots, 0)^t = s$ , where  $\rho > 0$ , and a nonsingular nonnegative diagonal matrix  $D$  so that  $D^{-1}A$  has eigenvalues  $\lambda_1, \dots, \lambda_n$ .*

*Proof.* Let  $B$  be a tridiagonal matrix, as assured by Corollary 2.1, so that  $b_{ii-1} < 0$ ,  $b_{11} > |b_{12}|$ ,  $b_{nn} > |b_{n-1n}|$ , and  $b_{ii} > |b_{ii-1}| + |b_{ii+1}|$  for  $i = 2, \dots, n - 1$  and so that  $B$  has eigenvalues  $\lambda_1, \dots, \lambda_n$ .

Consider the equation

$$Bx = s,$$

where  $s = (\rho_1, 0, \dots, 0)$  with  $\rho_1 > 0$ . As  $B$  is an  $M$  matrix,  $B^{-1} > 0$ . Hence  $x > 0$ . Set  $F = \text{diag}(x_1, \dots, x_n)$  and  $A = FBF$ . Then  $Ae = Fs = (\rho_2, 0, \dots, 0)$  with  $\rho_2 > 0$ .

Now set  $D = \text{diag}(1/x_1^2, \dots, 1/x_n^2)$ . Then  $D^{-1}A$  has the same eigenvalues as  $B$  and hence the result follows.  $\square$

Other such results can also be obtained from the work of the previous sections. Theorem 1.1 shows the effect on the natural frequencies of the spring-mass system under changes of various mass elements. Theorem 1.2 gives necessary and sufficient conditions for a natural frequency to remain stationary under variations of mass elements. Finally, Theorem 1.2 can be used to show the effect on the natural frequencies of replacing springs by stronger springs in the spring-mass system. Other such results are no doubt also possible.

## REFERENCES

- [1] JOEL N. FRANKLIN, *Matrix Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.
- [2] D. J. HARTFIEL, *The effect of  $D$  on the maximum eigenvalue of  $DA$  where  $A$  is an  $m$ -matrix*, SIAM. J. Appl. Math., 35 (1978), pp. 119–122.
- [3] MARVIN MARCUS AND HENRYK MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA, 1964.
- [4] ALEXANDER WEINSTEIN AND WILLIAM STENGER, *Methods of Intermediate Problems for Eigenvalues*, Academic Press, New York, 1972.

## ON THE ASYMPTOTIC BEHAVIOR OF RESOLVENTS OF VOLTERRA EQUATIONS\*

GUSTAF GRIPENBERG†

**Abstract.** This paper considers the asymptotic behavior of the resolvent  $r$  of the Volterra equation

$$(1) \quad x(t) + \int_0^t a(t-s)x(s) ds = f(t), \quad t \geq 0$$

given by

$$(2) \quad r(t) + \int_0^t a(t-s)r(s) ds = a(t), \quad t \geq 0.$$

Both integrability and pointwise estimates for  $r$  are established and some of these are shown to be uniform with respect to  $a$ . These results in turn give information about the asymptotic properties of the solution  $x$  of (1).

**1. Introduction.** The purpose of this paper is to study some aspects of the asymptotic behavior of the resolvent of the linear Volterra equation

$$(1.1) \quad x(t) + \int_0^t a(t-s)x(s) ds = f(t), \quad t \in R_+ = [0, \infty),$$

i.e., the unique (provided  $a$  is locally integrable) solution of the equation

$$(1.2) \quad r(t) + \int_0^t a(t-s)r(s) ds = a(t), \quad t \in R_+.$$

The importance of the resolvent derives from the fact that the solution of (1.1) is given by

$$(1.3) \quad x(t) = f(t) - \int_0^t r(t-s)f(s) ds, \quad t \in R_+,$$

but the properties of the resolvent are also important when one studies nonlinear Volterra integral equations. As seen from (1.3) it is interesting to determine when  $r \in L^1(R_+)$  and this question has been studied in [3], [4], [6], [8]–[10]. Other properties of the resolvents have been studied in [2], [7], [8]. In this work the use of the Laplace transform plays a crucial role, and we define for functions  $a$  such that  $e^{-\sigma t}a(t) \in L^1(R_+)$ ,  $\sigma > 0$ ,

$$\hat{a}(s) = \int_0^\infty e^{st}a(t) dt, \quad s = \sigma + i\tau, \quad \sigma > 0, \quad \tau \in R,$$

$$\hat{a}(i\tau) = \lim_{\sigma \rightarrow 0^+} \hat{a}(\sigma + i\tau).$$

**2. Statement of results.** First we recall the following result proved (with a slightly greater constant in (2.4) below) in [4].

---

\* Received by the editors March 7, 1979, and in revised form September 20, 1979.

† Institute of Mathematics, Helsinki University of Technology, Espoo 15, Finland.

THEOREM 1. Assume that

(2.1)  $a \in L^1_{loc}(R_+),$

(2.2)  $a$  is nonnegative, nonincreasing and convex on  $(0, \infty),$

(2.3)  $r$  is the solution of (1.2).

Then

(2.4)  $\|r\|_{L^1(R_+)} \leq 320.$

The idea of the proof is that one combines the estimates in [10, pp. 319, 320] with the method in [10, pp. 322, 323] to get the estimate

$$\|r\|_{L^1(R_+)} \leq 40 \int_0^\delta a(t) dt + 640 \left( \int_0^\delta a(t) dt \right)^{-1}$$

or, by the same reasoning the estimate

$$\|r\|_{L^1(R_+)} \leq 40 \int_0^\infty a(t) dt,$$

and then one chooses  $\delta$  appropriately. (For details, see [4]). As shown in [8, Thm. 2] the constant in (2.4) can be replaced by 1 if one also assumes that  $a$  is positive and  $\log a$  is convex.

If in addition to (2.2) it is assumed that  $-a'$  (" $'$ " =  $d/dt$ ) is convex, then we get a stronger result.

THEOREM 2. Assume that (2.1)–(2.3) hold and that

(2.5)  $-a'$  is convex on  $(0, \infty).$

Then

(2.6)  $\|\text{var}(r; [t, \infty))\|_{L^1(R_+)} \leq 122000.$

It follows from [2, Thm. 2] that if  $\log |a'|$  is convex too, then the constant in (2.6) can be replaced by 1. The interesting point in (2.4) and (2.6) is, of course, that the constants are independent of  $a$ .

We also have

COROLLARY 1. Assume that (2.3) holds and that  $a = a_1 + a_2$ , where  $a_1$  satisfies (2.1), (2.2), (2.5) and

(2.7)  $\text{var}(a_2; [t, \infty)) \in L^1(R_+),$

(2.8)  $\hat{a}(s) \neq -1, \quad \text{Re } s \geq 0.$

Then

(2.9)  $\text{var}(r; [t, \infty)) \in L^1(R_+).$

This result can be used in the study of nonlinear Volterra equations, cf. [5, Thm. 3].

In the next theorem we consider the asymptotic behavior of the function  $\int_t^\infty r(u) du, t \in R_+,$  (it follows from the assumptions below that  $r \in L^1(R_+)$ ).

THEOREM 3. Assume that (2.3) and (2.8) hold, and that

(2.10)  $a = a_1 + a_2,$  where  $a_1$  satisfies (2.1) and (2.2), and  $a_2$  satisfies (2.7),

(2.11)  $\liminf_{t \rightarrow \infty} t^{-\alpha} \int_0^t a(u) du > 0$  for some  $\alpha \in (0, 1).$

Then

$$(2.12) \quad \int_t^\infty r(u) \, du = O(t^{-\alpha}) \quad \text{as } t \rightarrow \infty.$$

If  $a$  satisfies (2.1), (2.2) and  $\log a$  is convex, then this result is quite easy to establish, see [2, Thm. 1]. If one considers the case  $a(t) = t^{-\beta}$ ,  $\beta \in (0, 1)$ , then one sees that the exponent in (2.12) is the best possible. The asymptotic behavior of  $\int_t^\infty r(u) \, du$  is uniform, under certain assumptions, as seen from the following result.

**COROLLARY 2.** *Assume that (2.1)–(2.3) hold, and that*

$$(2.13) \quad \text{there exist positive constants } \alpha, \gamma \text{ and } T, \alpha \in (0, 1) \\ \text{such that } t^{-\alpha} \int_0^t a(u) \, du \geq \gamma \text{ when } t \geq T.$$

Then

$$(2.14) \quad 0 \leq \int_t^\infty r(s) \, ds \leq ((\gamma\alpha)^{-1} + 110(\gamma(1-\alpha))^{-1} + \gamma^{-1} + 40T^\alpha)t^{-\alpha}, \quad t \geq T.$$

Observe that the inequality in (2.14) is only of interest when  $t$  is large since by [7, Thm. 1.4] we have  $0 \leq \int_t^\infty r(s) \, ds \leq 1$ ,  $t \in \mathbb{R}_+$ .

In the next theorem we consider the question concerning bounds on  $r(t)$ . The result is very similar to Theorem 3.

**THEOREM 4.** *Assume that (2.3), (2.8) and (2.11) hold and that*

$$(2.15) \quad a = a_1 + a_2, \text{ where } a_1 \text{ satisfies (2.1), (2.2) and (2.5),}$$

$$(2.16) \quad a_2 \in AC_{\text{loc}}((0, \infty)), \quad \lim_{t \rightarrow \infty} a_2'(t) = 0,$$

$$(2.17) \quad \int_t^\infty \text{var}(a_2'; [u, \infty)) \, du \in L^1(\mathbb{R}_+).$$

Then

$$(2.18) \quad r(t) = O(t^{-\alpha-1}) \quad \text{as } t \rightarrow \infty.$$

In analogy with Corollary 2 we have the following result for the uniform behavior of the resolvents.

**COROLLARY 3.** *Assume that (2.1)–(2.3), (2.5) and (2.13) hold. Then*

$$(2.19) \quad |r(t)| \leq (110(\alpha\gamma)^{-1} + 27000(\gamma(1-\alpha))^{-1} + 110\gamma^{-1} + 14000T^\alpha)t^{-\alpha-1}, \quad t \geq T.$$

The proofs of Theorems 3 and 4 and Corollaries 2 and 3 rely on ideas used in [1]. From [4, Thms. 1–3] and Corollary 3 we get the following result.

**COROLLARY 4.** *Assume that (2.3) and (2.8) hold and that*

$$(2.20) \quad a = a_1 + a_2 + a_3, \text{ where } a_1 \text{ satisfies the assumptions of Corollary 3,}$$

$$(2.21) \quad a_2 \in BV(\mathbb{R}_+), \quad \int_1^\infty t^{-\alpha} \text{var}(a_2; [t, \infty)) < \infty,$$

$$(2.22) \quad a_3 \in L^1(\mathbb{R}_+).$$

Then  $r \in L^1(\mathbb{R}_+)$ .

It is straightforward, using (1.3) and the fact that  $\int_0^\infty r(s) \, ds = 1$  under the assumptions in the corollaries below, to see that we have the following applications of Theorems 3 and 4.

COROLLARY 5. Assume that (1.1) and the assumptions of Theorem 3 hold and that

$$(2.23) \quad f \in BV(R_+) \text{ and } \text{var}(f; [t, \infty)) = O(t^{-\alpha}) \text{ as } t \rightarrow \infty.$$

Then

$$(2.24) \quad x(t) = O(t^{-\alpha}) \text{ as } t \rightarrow \infty.$$

COROLLARY 6. Assume that (1.1) and the assumptions of Theorem 4 hold, and that

$$(2.25) \quad f \in L^\infty(R_+) \text{ and } f(t) = O(t^{-\alpha}) \text{ as } t \rightarrow \infty.$$

Then (2.24) holds.

**3. Proofs of Theorem 2 and Corollary 1.** To prove Theorem 2 we first assume that

$$(3.1) \quad a(0) < \infty.$$

Now it is easy to conclude from (1.2), (2.2), (3.1) and the fact that  $r \in L^1(R_+)$ , see [10], that  $r$  is absolutely continuous on  $R_+$ . Consequently it is sufficient to show that

$$(3.2) \quad \int_0^\infty t|r'(t)| dt \leq 121048.$$

Let  $b(t) = tr'(t)$ . Obviously  $\hat{b}(s) = (d/d\tau)i(s\hat{r}(s))$ ,  $s = \sigma + i\tau$ ,  $\sigma > 0$  and so we have by (1.2),

$$(3.3) \quad \hat{b}(s) = -\hat{r}(s) - s\hat{a}'(s)(1 + \hat{a}(s))^{-2}, \quad \text{Re } s > 0.$$

Here “'” denotes differentiation with respect to  $\tau$ . In view of Theorem 1 we have only to consider the second term in (3.3).

Recall that, as a consequence of (2.2) and (2.5),  $\hat{a}(i\tau)$  exists and is twice continuously differentiable when  $\tau \neq 0$ , and we have the estimates, see [1, Lemma 4.1, Lemma 5.1],

$$(3.4) \quad |\hat{a}(i\tau)| \geq 2^{-3/2} \int_0^{|\tau|^{-1}} a(u) du, \quad \tau \neq 0,$$

$$(3.5) \quad |\hat{a}'(i\tau)| \leq 40 \int_0^{|\tau|^{-1}} ua(u) du, \quad \tau \neq 0,$$

$$(3.6) \quad |\hat{a}''(i\tau)| \leq 6000 \int_0^{|\tau|^{-1}} u^2 a(u) du, \quad \tau \neq 0.$$

It follows from (2.2) that  $\text{Re } \hat{a}(i\tau) \geq 0$ , see [10, p. 320], and so

$$(3.7) \quad |1 + \hat{a}(i\tau)| \geq \max\{1, |\hat{a}(i\tau)|\}, \quad \tau \neq 0.$$

Let

$$(3.8) \quad h(\tau) = \tau\hat{a}'(i\tau)(1 + \hat{a}(i\tau))^{-2}, \quad \tau \neq 0.$$

We can differentiate  $h$  when  $\tau \neq 0$ , and we obtain

$$(3.9) \quad \begin{aligned} h'(\tau) &= \hat{a}'(i\tau)(1 + \hat{a}(i\tau))^{-2} + i\tau\hat{a}''(i\tau)(1 + \hat{a}(i\tau))^{-2} \\ &\quad - 2i\tau(\hat{a}'(i\tau))^2(1 + \hat{a}(i\tau))^{-3}, \quad \tau \neq 0. \end{aligned}$$

By (2.2) and (3.4)–(3.7) we have  $|\tau\hat{a}''(i\tau)| \leq 6000 \int_0^{|\tau|^{-1}} ua(u) du$  and  $|\tau\hat{a}'(i\tau)(1 + \hat{a}(i\tau))^{-1}| \leq 80 \cdot 2^{1/2}$ , and so we conclude in the same way as in the proof of [4, Thm. 4]

(taking  $\delta = +\infty$  in [4, line (5.6)] when  $\int_0^\infty a(s) ds < 4$ ) that

$$(3.10) \quad \int_{-\infty}^\infty |h'(\tau)| d\tau \leq 24156.$$

Since (3.5) and (3.7) yield  $\lim_{|\tau| \rightarrow \infty} h(\tau) = 0$ , and since  $(d/d\tau)i(s\hat{s}(s)) + r(\hat{s})$  is bounded and analytic in  $\text{Re } s > 0$ , we deduce, in the same way as in the proof of [4, Thm. 4], from Theorem 1, (3.3), (3.8) and (3.10) that (3.2) holds (cf. also [10]).

If  $\lim_{t \rightarrow 0^+} a(t) = +\infty$ , then we define the functions  $a_n$  by  $a_n(t) = a(t + 1/n)$  and let  $r_n$  be the resolvent associated with  $a_n$ . By the previous result, (2.6) holds with  $r$  replaced by  $r_n$  and it is easy to see from (1.2), (2.2) and the definition of  $a_n$  that

$$(3.11) \quad r_n \rightarrow r \quad \text{in } L^1(0, T) \text{ as } n \rightarrow \infty \text{ for all } T > 0.$$

By (1.2) and (2.1) we conclude that  $r$  is continuous on  $(0, \infty)$  (note that we get  $\text{ess sup}_{t>0} |tr(t)| < \infty$  by (2.6) (with  $r_n$ ) and (3.11), and so it follows from (3.11) that  $\text{var}(r; [t, \infty)) \leq \liminf \text{var}(r_n; [t, \infty))$  for all  $t > 0$ , and now (2.6) follows by Fatou's lemma and the corresponding result for  $r_n$ ).

To prove Corollary 1, note that by (2.8) we may, without loss of generality, assume that  $\lim_{t \rightarrow \infty} a_2(t) = 0$  (and so by (2.7)  $a_2 \in L^1(\mathbb{R}_+)$ ), and that  $a_1(0) < \infty$ . Choose a sequence of functions  $\{b_n\}$  such that

$$(3.12) \quad \lim_{t \rightarrow \infty} b_n(t) = 0, \quad \lim_{n \rightarrow \infty} \|\text{var}(b_n; [t, \infty))\|_{L^1(\mathbb{R}_+)} = 0,$$

and such that  $c_n \stackrel{\text{def}}{=} a_2 - b_n \in BV(\mathbb{R}_+)$ . Let  $r_n$  be the resolvent associated with  $a_1 + c_n$ . We are going to show that

$$(3.13) \quad \sup_{n \geq 1} \|\text{var}(r_n; [t, \infty))\|_{L^1(\mathbb{R}_+)} < \infty.$$

This statement follows if we can show that  $(d/d\tau)(sr_n(s))$ ,  $\text{Re } s > 0$ , is the Laplace-Stieltjes transform of a measure with total variation bounded independently of  $n$ . From (1.2) and the definition of  $r_n$  we obtain

$$(3.14) \quad \begin{aligned} \frac{d}{d\tau}(sr_n(s)) &= i\hat{r}_n(s) + is\hat{a}'_1(s)(1 + \hat{a}_1(s))^{-2}(1 + \hat{c}_n(s)(1 + \hat{a}_1(s))^{-1})^{-2} \\ &\quad + \frac{d}{d\tau}(s\hat{c}_n(s))(1 + \hat{a}_1(s) + \hat{c}_n(s))^{-2} \\ &\quad - i\hat{c}_n(s)(1 + \hat{a}_1(s) + \hat{c}_n(s))^{-2}, \quad \text{Re } s > 0. \end{aligned}$$

We have by (1.2), (2.3) and the definitions of  $c_n, r_n$ ,

$$(3.15) \quad \hat{r}_n(s) = (\hat{r}(s) - \hat{b}_n(s)(1 - \hat{r}(s)))(1 - \hat{b}_n(s)(1 - \hat{r}(s)))^{-1}, \quad \text{Re } s > 0,$$

$$(3.16) \quad \begin{aligned} &(1 + \hat{c}_n(s)(1 + \hat{a}_1(s))^{-1})^{-2} \\ &= (1 - \hat{a}_2(s)(1 - \hat{r}(s))^2(1 - \hat{b}_n(s)(1 - \hat{r}(s)))^{-2}, \quad \text{Re } s > 0, \end{aligned}$$

and

$$(3.17) \quad (1 + \hat{a}_1(s) + \hat{c}_n(s))^{-2} = (1 - \hat{r}(s))^2(1 - \hat{b}_n(s)(1 - \hat{r}(s)))^{-2}, \quad \text{Re } s > 0.$$

Now it is easy to see, from Theorem 1, (2.7), (2.8), (3.12) and (3.14)–(3.17), that (3.13) holds when we observe that  $r \in L^1(\mathbb{R}_+)$ , see [6].



From (3.12) we note that (3.11) holds in this case too, and since it readily follows that  $r(t) - a_2(t)$  is continuous on  $(0, \infty)$  we can complete the proof of Corollary 1 in the same way as above.

**4. Proofs of Theorem 3 and Corollary 2.** We may again assume that  $\lim_{t \rightarrow \infty} a_2(t) = 0$ . Let  $x(t) = \int_t^\infty r(u) du$ . Since  $\int_0^\infty a(u) du = +\infty$  by (2.11) we have  $\int_0^\infty r(u) du = 1$ , (that  $r \in L^1(\mathbb{R}_+)$  follows from results in [6]), and so

$$(4.1) \quad \hat{x}(s) = s^{-1}(1 + \hat{a}(s))^{-1}, \quad \text{Re } s > 0.$$

We define the function  $H$  by

$$(4.2) \quad H(t) = e^{-t^2}, \quad t \in \mathbb{R}.$$

By standard Fourier transform results, we know that (recall the notation  $s = \sigma + i\tau$ )

$$(4.3) \quad x(t) = \lim_{\sigma \rightarrow 0^+} \lim_{n \rightarrow \infty} (2\pi)^{-1} \int_{-\infty}^\infty e^{i\tau t} H(n^{-1}\tau) \hat{x}(s) d\tau.$$

From (2.1), (2.2), (2.7) and (2.8) it follows that  $\inf_{\text{Re } s \geq 0} |1 + \hat{a}_2(s)(1 + \hat{a}_1(s))^{-1}|^{-1} > 0$ , and so there exists a positive constant  $c_1$  such that (cf. (3.7)),

$$(4.4) \quad |1 + \hat{a}(s)| \geq c_1 \max \{1, |\hat{a}_1(s)|\}, \quad \text{Re } s \geq 0.$$

Choose constants  $T$  and  $c_2$  so that

$$(4.5) \quad \int_0^t a_1(u) du \geq c_2 t^\alpha, \quad t \geq T.$$

This is possible by (2.7), (2.10) and (2.11). Observe that  $\hat{a}_1(i\tau)$  is still continuously differentiable when  $\tau \neq 0$ , and (3.4) and (3.5) hold with  $a$  replaced by  $a_1$  (and also when  $i\tau$  is replaced by  $\sigma + i\tau$  on the left side and  $a$  by  $e^{-\sigma u} a_1(u)$  on the right side).

Let  $t > T$  be arbitrary. By the dominated convergence theorem, (3.4), (4.1), (4.4) and (4.5) we have for some constant  $c_3$ ,

$$(4.6) \quad \lim_{\sigma \rightarrow 0^+} \int_{\sigma \leq |\tau| \leq t^{-1}} |\hat{x}(s)| d\tau \leq c_3 \int_0^{t^{-1}} \tau^{-1+\alpha} d\tau = c_3 \alpha^{-1} t^{-\alpha}.$$

We also get by (3.4), (4.1), (4.4) and (4.5),

$$(4.7) \quad \lim_{\sigma \rightarrow 0^+} \int_{|\tau| \leq \sigma} |\hat{x}(s)| d\tau \leq c_3 c_2 \lim_{\sigma \rightarrow 0^+} \int_0^\sigma \sigma^{-1} \left( \int_0^{\sigma^{-1}} e^{-\sigma u} a_1(u) du \right)^{-1} d\tau = 0.$$

An integration by parts gives

$$(4.8) \quad \begin{aligned} \int_{|\tau| > t^{-1}} e^{i\tau t} H(n^{-1}\tau) \hat{x}(s) d\tau &= (it)^{-1} H(n^{-1}t^{-1}) (e^{-i\hat{x}(\sigma - it^{-1})} - e^{i\hat{x}(\sigma + it^{-1})}) \\ &\quad - (it)^{-1} \int_{|\tau| > t^{-1}} e^{i\tau t} n^{-1} H'(n^{-1}\tau) \hat{x}(s) d\tau \\ &\quad - (it)^{-1} \int_{|\tau| > t^{-1}} e^{i\tau t} H(n^{-1}\tau) \hat{x}'(s) d\tau, \quad \text{Re } s > 0. \end{aligned}$$

Here we used  $\lim_{|\tau| \rightarrow \infty} \hat{x}(s) = 0$ , and this fact combined with (4.2) also yields

$$(4.9) \quad \lim_{n \rightarrow \infty} \int_{|\tau| > t^{-1}} e^{i\tau t} n^{-1} H'(n^{-1}\tau) \hat{x}(s) d\tau = 0.$$

By (3.4), (4.1), (4.4) and (4.5) we obtain

$$(4.10) \quad |\hat{x}(\sigma \pm it^{-1})| \leq 2^{-1} c_3 e^{\sigma t} t^{1-\alpha}.$$

From (4.1) we have

$$(4.11) \quad \hat{x}'(s) = -(i + i\hat{a}_1(s) + is\hat{a}'_1(s) + \frac{d}{d\tau}(s\hat{a}_2(s)))(s(1 + \hat{a}(s)))^{-2},$$

and since (2.8) implies that  $\sup_{\text{Re } s > 0} |(d/d\tau)(s\hat{a}_2(s))| < \infty$ , it follows from (2.2), (2.10), (3.4), (3.5), (4.2), (4.4) and (4.11) that for some constants  $c_4$  and  $c_5$ ,

$$(4.12) \quad \lim_{\sigma \rightarrow 0^+} \lim_{n \rightarrow \infty} \left| \int_{t^{-1} < |\tau| < T^{-1}} e^{i\tau} H(n^{-1}\tau) \hat{x}'(s) d\tau \right| \leq c_4 \int_{t^{-1}}^{T^{-1}} \tau^{-2} \left( \int_0^{\tau^{-1}} a_1(u) du \right)^{-1} d\tau \leq c_5 \int_T^t \tau^{-\alpha} d\tau \leq c_5(1-\alpha)^{-1} t^{1-\alpha}.$$

In the same way, it follows that there exists a constant  $c_6$  such that

$$(4.13) \quad \lim_{\sigma \rightarrow 0^+} \lim_{n \rightarrow \infty} \left| \int_{|\tau| \geq T^{-1}} e^{i\tau} H(n^{-1}\tau) \hat{x}'(s) d\tau \right| \leq c_6 \int_{T^{-1}}^{\infty} \tau^{-2} d\tau = c_6 T.$$

Combining (4.2), (4.3), (4.6)–(4.10), (4.12) and (4.13) we get the desired conclusion. This completes the proof of Theorem 3.

To prove Corollary 2 we have only to note that in this case  $c_1 = 1$ ,  $c_2 = \gamma$ ,  $c_3 = \gamma^{-1} 2^{5/2}$ ,  $c_4 = 2^{5/2} + 640$ ,  $c_5 = \gamma^{-1} c_4$  and  $c_6 = 2 + 160 \cdot 2^{1/2}$ , and that  $T/t \leq T^\alpha t^{-\alpha}$ . By [7, Thm. 1.4] we have  $\int_0^t r(s) ds \leq 1$ , and since  $\int_0^\infty r(s) ds = 1$ , the conclusion of Corollary 2 follows.

**5. Proofs of Theorem 4 and Corollary 3.** By (2.8), (2.16) and (2.17) we may, without loss of generality, assume that  $\lim_{t \rightarrow \infty} a_2(t) = 0$ , and we note that (2.7) holds. We also observe that (4.4) holds and we choose  $T$  and  $c_2$  so that (4.5) holds. From (2.16) and (2.17) we deduce that  $\hat{a}_2(i\tau)$  is twice continuously differentiable when  $\tau \neq 0$  and we have the estimates

$$(5.1) \quad |\hat{a}'_2(i\tau)| \leq c_3 |\tau|^{-1}, \quad |\hat{a}''_2(i\tau)| \leq c_4 |\tau|^{-2}, \quad \tau \neq 0$$

for some constants  $c_3$  and  $c_4$ , cf. [10, p. 320], (not the same constants as in the proof of Theorem 3). We also have (3.4)–(3.6) with  $a$  replaced by  $a_1$ .

It is easy to conclude from (2.3), (2.8) and (2.15)–(2.17) that  $r \in L^1(\mathbb{R}_+)$ , see [6], and that  $r$  is continuous on  $(0, \infty)$ . We clearly have

$$(5.2) \quad \hat{r}(i\tau) = \hat{a}(i\tau)(1 + \hat{a}(i\tau))^{-1}.$$

Using the definition (4.2) we deduce from standard results concerning Fourier transforms that

$$(5.3) \quad r(t) = \lim_{n \rightarrow \infty} (2\pi)^{-1} \int_{-\infty}^{\infty} e^{i\tau t} H(n^{-1}\tau) \hat{r}(i\tau) d\tau, \quad t > 0.$$

As  $\hat{a}(i\tau)$  is continuously differentiable when  $\tau \neq 0$ , the same holds for  $r(i\tau)$  and an

integration by parts in (5.3) gives

$$(5.4) \quad r(t) = (2\pi it)^{-1} \left( \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} e^{it\tau} n^{-1} H'(n^{-1}\tau) \hat{r}(i\tau) d\tau \right. \\ \left. + \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} e^{it\tau} H(n^{-1}\tau) \hat{r}'(i\tau) d\tau \right), \quad t > 0.$$

Here we used  $\lim_{|\tau| \rightarrow \infty} \hat{r}(i\tau) = 0$ , and combining this fact with (4.2) we obtain

$$(5.5) \quad \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} e^{it\tau} n^{-1} H'(n^{-1}\tau) \hat{r}(i\tau) d\tau = 0.$$

Next let  $t > T$  be arbitrary and fixed. By (2.2), (2.15), (3.4), (3.5), (4.4), (4.5), (5.1) and (5.2), there exist constants  $c_5$  and  $c_6$  such that

$$(5.6) \quad \int_{|\tau| < t^{-1}} |\hat{r}'(i\tau)| d\tau \leq c_5 \int_0^{t^{-1}} \left( \left( \int_0^{\tau^{-1}} u a_1(u) du \right) \left( \int_0^{\tau^{-1}} a_1(u) du \right)^{-2} \right. \\ \left. + \tau^{-1} \left( \int_0^{\tau^{-1}} a_1(u) du \right)^{-1} \right) d\tau \leq c_6 t^{-\alpha}.$$

By an additional integration by parts we obtain

$$(5.7) \quad \int_{|\tau| > t^{-1}} e^{it\tau} H(n^{-1}\tau) \hat{r}'(i\tau) d\tau = (it)^{-1} H(n^{-1}t^{-1}) (e^{-i\hat{r}'(-it^{-1})} - e^{i\hat{r}'(it^{-1})}) \\ - (it)^{-1} \int_{\tau > t^{-1}} e^{it\tau} n^{-1} H'(n^{-1}\tau) \hat{r}'(i\tau) d\tau \\ - (it)^{-1} \int_{\tau > t^{-1}} e^{it\tau} H(n^{-1}\tau) \hat{r}''(i\tau) d\tau.$$

Here we used the fact that  $\lim_{|\tau| \rightarrow \infty} \hat{r}'(i\tau) = 0$  (a consequence of (3.5), (4.4) and (5.1)) and we also note that

$$(5.8) \quad \lim_{n \rightarrow \infty} \int_{|\tau| > t^{-1}} e^{it\tau} n^{-1} H'(n^{-1}\tau) \hat{r}'(i\tau) d\tau = 0.$$

From (2.2), (2.15), (3.4), (3.5), (4.4), (4.5), (5.1) and (5.2) we get

$$(5.9) \quad |\hat{r}'(\pm it^{-1})| \leq c_7 \left( \left( \int_0^t u a_1(u) du \right) \left( \int_0^t a_1(u) du \right)^{-2} + t \left( \int_0^t a_1(u) du \right)^{-1} \right) \leq c_8 t^{-\alpha+1}$$

for some constants  $c_7$  and  $c_8$ .

By (5.2) we obtain

$$(5.10) \quad \hat{r}''(i\tau) = \hat{a}''(i\tau)(1 + \hat{a}(i\tau))^{-2} - 2(\hat{a}'(i\tau))^2(1 + \hat{a}(i\tau))^{-3}, \quad \tau \neq 0.$$

Combine (2.2), (2.15), (3.4)–(3.6), (4.4), (4.5) and (5.1) with (5.9). Then we see that

there exist constants  $c_9$  and  $c_{10}$  so that

$$\begin{aligned}
 \int_{t^{-1} < |\tau| < T^{-1}} |\hat{r}''(i\tau)| \, d\tau &\leq c_9 \int_{t^{-1}}^{T^{-1}} \left( \left( \int_0^{\tau^{-1}} u^2 a_1(u) \, du \right) \left( \int_0^{\tau^{-1}} a_1(u) \, du \right)^{-2} \right. \\
 (5.11) \qquad \qquad \qquad &+ \left. \left( \int_0^{\tau^{-1}} u a_1(u) \, du \right)^2 \left( \int_0^{\tau^{-1}} a_1(u) \, du \right)^{-3} \right. \\
 &+ \left. \tau^{-2} \left( \int_0^{\tau^{-1}} a_1(u) \, du \right)^{-1} \right) \, d\tau \\
 &\leq c_{10} t^{1-\alpha}.
 \end{aligned}$$

In the same way we also deduce that

$$\begin{aligned}
 \int_{|\tau| \geq T^{-1}} |\hat{r}''(\tau)| \, d\tau &\leq c_{11} \int_{T^{-1}}^{\infty} \left( \left( \int_0^{\tau^{-1}} u^2 a_1(u) \, du \right) \left( \int_0^{\tau^{-1}} a_1(u) \, du \right)^{-1} \right. \\
 (5.12) \qquad \qquad \qquad &+ \left. \left( \int_0^{\tau^{-1}} u a_1(u) \, du \right)^2 \left( \int_0^{\tau^{-1}} a_1(u) \, du \right)^{-2} + \tau^{-2} \right) \, d\tau \leq 3c_{11} T.
 \end{aligned}$$

The conclusion of Theorem 4 now follows from (4.2), (5.4)–(5.9), (5.11) and (5.12).

To prove Corollary 3 we have only to note that in this case  $c_1 = 1$ ,  $c_2 = \gamma$ ,  $c_5 = 640$ ,  $c_6 = (\alpha\gamma)^{-1}c_5$ ,  $c_7 = 320$ ,  $c_8 = \gamma^{-1}c_7$ ,  $c_9 = (96000 + 51200 \cdot 2^{1/2})$ ,  $c_{10} = (\gamma(1 - \alpha))^{-1}c_9$ ,  $c_{11} = (24000 \cdot 2^{1/2} + 51200)$  and that  $Tt^{-1} \leq T^\alpha t^{-\alpha}$ .

REFERENCES

- [1] R. W. CARR AND K. B. HANNSGEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961–984.
- [2] G. GRIPENBERG, *On positive nonincreasing resolvents of Volterra equations*, J. Differential Equations, 30 (1978), pp. 380–390.
- [3] ———, *A Volterra equation with nonintegrable resolvent*, Proc. Amer. Math. Soc., 73 (1979), pp. 57–60.
- [4] ———, *Integrability properties of resolvents of Volterra equations*, Rep. HTKK-MAT-A117, Helsinki Univ. of Technology, Helsinki, Finland, 1978.
- [5] ———, *On nonlinear Volterra equations with nonintegrable kernels*, this Journal, to appear.
- [6] 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.
- [7] J. J. LEVIN, *Resolvents and bounds for linear and nonlinear Volterra equations*, Trans. Amer. Math. Soc., 228 (1977), pp. 207–222.
- [8] R. K. MILLER, *On Volterra integral equations with nonnegative integrable resolvents*, J. Math. Anal. Appl., 22 (1968), pp. 319–340.
- [9] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, RI, 1934.
- [10] D. F. SHEA AND S. WAINGER, *Variants of the Wiener–Lévy theorem, with applications to stability problems for some integral equations*, Amer. J. Math., 97 (1975), pp. 312–342.

## SOME NEW INEQUALITIES RELATED TO CERTAIN ULTRASPHERICAL POLYNOMIALS\*

GIOVANNI MONEGATO†

**Abstract.** A remarkable class of polynomials  $\{E_{n+1}^{(\mu)}(x)\}$ ,  $n \geq 1$  was first considered by Stieltjes in 1894 and later studied by Szegő. This class can be uniquely defined, for example, by the orthogonality relation

$$\int_{-1}^1 (1-x^2)^{\mu-1/2} P_n^{(\mu)}(x) E_{n+1}^{(\mu)}(x) x^k dx = 0, \quad k = 0, 1, \dots, n, \quad \mu > -\frac{1}{2},$$

where  $P_n^{(\mu)}(x)$  is the classical ultraspherical polynomial.

In this paper we consider an inequality given by Szegő, which involves these new polynomials, and derive some new bounds for  $P_n^{(\mu)}(x)$ ,  $E_{n+1}^{(\mu)}(x)$  and their first derivatives.

**1. Introduction.** In this paper, we are concerned with classical ultraspherical polynomials and certain new polynomials which may be associated to them and give rise to a class of quadrature formulas which have recently been studied.

We first review some known results which are needed to derive the inequalities we will present in the next two sections.

Let  $P_n^{(\mu)}(x)$ ,  $\mu > -\frac{1}{2}$ , be the (ultraspherical) polynomial of degree  $n$  orthogonal in  $(-1, 1)$  with respect to the weight function  $w(x) = (1-x^2)^{\mu-1/2}$ ; and  $Q_n^{(\mu)}(x)$ , the associated function of the second kind:

$$(1) \quad (1-x^2)^{\mu-1/2} Q_n^{(\mu)}(x) = \frac{1}{2} \frac{\Gamma(2\mu)}{\Gamma(\mu+1/2)} \int_{-1}^1 (1-t^2)^{\mu-1/2} \frac{P_n^{(\mu)}(t)}{x-t} dt = q_n^{(\mu)}(x).$$

For the properties of this function we refer to [5, §4.61 and §4.62].  $P_n^{(\mu)}(x)$  and  $Q_n^{(\mu)}(x)$  are two linearly independent solutions of the following linear homogeneous differential equation of the second-order

$$(2) \quad (1-x^2)y'' - (2\mu+1)xy' + n(n+2\mu)y = 0.$$

Furthermore, we define  $Q_n^*(\mu; x)$  by

$$(3) \quad \lim_{\varepsilon \rightarrow +0} [q_n^{(\mu)}(x+i\varepsilon) + q_n^{(\mu)}(x-i\varepsilon)] = 2(1-x^2)^{\mu-1/2} Q_n^*(\mu; x),$$

and note that this new function is analytic on  $(-1, 1)$  and satisfies (2). The following expansion is due to Szegő [4]

$$(4) \quad (1-\cos^2 \varphi)^{\mu-1/2} \left[ Q_n^*(\mu; \cos \varphi) + i \frac{\pi}{2} \frac{\Gamma(2\mu)}{\Gamma(\mu+1/2)} P_n^{(\mu)}(\cos \varphi) \right] \\ = \sqrt{\pi} \frac{\Gamma(n+2\mu)}{\Gamma(n+\mu+1)} \sum_{\nu=0}^{\infty} f_{\nu}^{(\mu)} e^{i(n+2\nu+1)\varphi}, \quad 0 < \varphi < \pi,$$

$$f_0^{(\mu)} = 1, \quad f_{\nu}^{(\mu)} = \frac{\nu-\mu}{\nu} \frac{n+\nu}{n+\mu+\nu} f_{\nu-1}^{(\mu)}, \quad \nu = 1, 2, \dots$$

Let now  $E_{n+1}^{(\mu)}(x)$  be the polynomial of degree  $n+1$ ,  $n = 1, 2, \dots$ , uniquely defined (up

\* Received by the editors December 27, 1978.

† Istituto di Calcoli Numerici, Università di Torino, I-10123 Torino, Italy. This work was performed under the auspices of the Italian Research Council.

to a constant factor) by the relation

$$(5) \quad \int_{-1}^1 (1-x^2)^{\mu-1/2} P_n^{(\mu)}(x) E_{n+1}^{(\mu)}(x) x^k dx = 0, \quad k = 0, 1, \dots, n.$$

These polynomials were first considered for  $\mu = \frac{1}{2}$  by Stieltjes [6] who introduced them via the relation

$$(6) \quad [(1-x^2)^{\mu-1/2} Q_n^{(\mu)}(x)]^{-1} = E_{n+1}^{(\mu)}(x) + \frac{a_1^{(\mu)}}{x} + \frac{a_2^{(\mu)}}{x^2} + \dots.$$

Szegő [4] studied them further and proved that when  $0 < \mu \leq 2$  their zeros are all in  $(-1, 1)$  and interlace with those of  $P_n^{(\mu)}(x)$ . In the case  $\mu > 2$  little is known about the zeros of  $E_{n+1}^{(\mu)}(x)$ , but numerical results for some  $\mu > 2$  indicate the presence of complex zeros [2].

The polynomials  $E_{n+1}^{(\mu)}(x)$  give rise to the remarkable class of quadrature formulas

$$(7) \quad \int_{-1}^1 (1-x^2)^{\mu-1/2} f(x) dx = \sum_{i=1}^n A_{i,n}^{(\mu)} f(\xi_{i,n}^{(\mu)}) + \sum_{j=1}^{n+1} B_{j,n}^{(\mu)} f(x_{j,n}^{(\mu)}) + R_n(f),$$

where  $\xi_{i,n}^{(\mu)}, i = 1, \dots, n$ , are the zeros of  $P_n^{(\mu)}(x)$ , and  $x_{j,n}^{(\mu)}, j = 1, 2, \dots, n+1$  are the zeros of  $E_{n+1}^{(\mu)}(x)$ . They have been studied in [1], [2], [3]. We recall, in particular, that (7) exists for any  $n \geq 1$  when  $0 \leq \mu \leq 2$  and has polynomial degree  $3n+1$  ( $3n+2$  when  $n$  is odd), i.e.,  $R_n(f) = 0$  whenever  $f(x)$  is a polynomial of degree  $3n+1$  ( $3n+2$ ).

Furthermore, letting

$$E_{n+1}^{(\mu)}(\cos \varphi) = \lambda_0^{(\mu)} \cos(n+1)\varphi + \lambda_1^{(\mu)} \cos(n-1)\varphi + \dots + \begin{cases} \lambda_{n/2}^{(\mu)} \cos \varphi, & n \text{ even,} \\ \frac{1}{2} \lambda_{(n+1)/2}^{(\mu)}, & n \text{ odd,} \end{cases}$$

where

$$\lambda_0^{(\mu)} = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n+\mu+1)}{\Gamma(n+2\mu)},$$

Szegő shows that for  $0 < \mu < 1$  one has

$$(8) \quad \lambda_i^{(\mu)} < 0, \quad i \geq 1 \quad \text{and} \quad \lambda_0^{(\mu)} > - \sum_{i \geq 1} \lambda_i^{(\mu)}.$$

Finally, introducing

$$e_{n+1}^{(\mu)}(\varphi) = \lambda_0^{(\mu)} \sin(n+1)\varphi + \lambda_1^{(\mu)} \sin(n-1)\varphi + \dots + \begin{cases} \lambda_{n/2}^{(\mu)} \sin \varphi, & n \text{ even,} \\ 0, & n \text{ odd} \end{cases}$$

and considering the product of (4) and

$$[E_{n+1}^{(\mu)}(\cos \varphi) - i e_{n+1}^{(\mu)}(\varphi)],$$

Szegő proves the following inequality, valid for  $0 < \mu < 1$ ,

$$(9) \quad (1 - \cos^2 \varphi)^{\mu-1/2} [Q_n^*(\mu; \cos \varphi) E_{n+1}^{(\mu)}(\cos \varphi) + \frac{\pi}{2} \frac{\Gamma(2\mu)}{\Gamma(\mu+\frac{1}{2})} P_n^{(\mu)}(\cos \varphi) e_{n+1}^{(\mu)}(\varphi)] > 1, \\ 0 < \varphi < \pi.$$

In the next sections, we use (9) as point of departure from which we derive new bounds for  $P_n^{(\mu)}(x)$ ,  $E_{n+1}^{(\mu)}(x)$  and their first derivatives, which do not seem to be present in the literature. In particular, we use some of these inequalities to derive lower bounds for  $H_{i,n}^{(\mu)}$ , the Christoffel constants associated to  $P_n^{(\mu)}(x)$ .

**2. Bounds for  $P_n^{(\mu)}(x)$  and its first derivative.** In this section, we examine some of the consequences of inequality (9), which plays, in all that follows, an essential role. More precisely, by evaluating (9) at the zeros  $\xi_{i,n}^{(\mu)} = \cos \varphi_{i,n}^{(\mu)}$  of  $P_n^{(\mu)}(x)$ , we derive the following

LEMMA. When  $0 < \mu < 1$ , we have

$$(10) \quad |\sin \varphi_{i,n}^{(\mu)}|^{2\mu-1} |Q_n^*(\mu; \cos \varphi_{i,n}^{(\mu)})| > \frac{\sqrt{\pi}}{2} \frac{\Gamma(n+2\mu)}{\Gamma(n+\mu+1)}.$$

*Proof.* From (9), we obtain

$$(11) \quad |\sin \varphi_{i,n}^{(\mu)}|^{2\mu-1} Q_n^*(\mu; \cos \varphi_{i,n}^{(\mu)}) E_{n+1}^{(\mu)}(\cos \varphi_{i,n}^{(\mu)}) > 1, \quad 0 < \mu < 1;$$

because of (8), when  $0 < \mu < 1$ , we also have

$$(12) \quad |E_{n+1}^{(\mu)}(\cos \varphi)| < 2\lambda_0^{(\mu)}, \quad |e_{n+1}^{(\mu)}(\varphi)| < 2\lambda_0^{(\mu)},$$

and hence, using the explicit value of  $\lambda_0^{(\mu)}$ , from (11) we finally derive (10).

Using the above result it is then possible to prove

THEOREM 1. Let  $H_{i,n}^{(\mu)}$ ,  $i = 1, \dots, n$  be the Christoffel constants associated to  $P_n^{(\mu)}(x)$  and  $0 < \mu < 1$ , then

$$(13) \quad H_{i,n}^{(\mu)} > \pi \frac{2^{1-2\mu}}{\Gamma(\mu)} \frac{\Gamma(n+2\mu)}{\Gamma(n+\mu+1)} \frac{1}{|P_n^{(\mu)'}(\cos \varphi_{i,n}^{(\mu)})|},$$

where  $P_n^{(\mu)'}(x)$  denotes the first derivative of  $P_n^{(\mu)}(x)$ .

*Proof.* Consider the Gaussian quadrature formula

$$\int_{-1}^1 (1-x^2)^{\mu-1/2} f(x) dx = \sum_{i=1}^n H_{i,n}^{(\mu)} f(\xi_{i,n}^{(\mu)}) + R_{g,n}(f),$$

where

$$H_{i,n}^{(\mu)} = \frac{1}{P_n^{(\mu)'}(\xi_{i,n}^{(\mu)})} \int_{-1}^1 (1-x^2)^{\mu-1/2} \frac{P_n^{(\mu)}(x)}{x - \xi_{i,n}^{(\mu)}} dx.$$

Recalling (1) and (3), and using Lebesgue's convergence theorem, the Christoffel constants may be expressed as follows:

$$(14) \quad H_{i,n}^{(\mu)} = 2 \frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(2\mu)} |\sin \varphi_{i,n}^{(\mu)}|^{2\mu-1} |Q_n^*(\mu; \cos \varphi_{i,n}^{(\mu)})| \frac{1}{|P_n^{(\mu)'}(\cos \varphi_{i,n}^{(\mu)})|}.$$

Using (10) and recalling the duplication formula of the gamma function, we finally have (13).

Since [5, (4. 7.14)]

$$(15) \quad \frac{dP_n^{(\mu)}(x)}{dx} = 2\mu P_{n-1}^{(\mu+1)}(x)$$

and [5, (7.33.1)]

$$(16) \quad |P_{n-1}^{(\mu+1)}(\cos \varphi)| < \frac{\Gamma(n+2\mu+1)}{\Gamma(n)\Gamma(2\mu+2)},$$

a first consequence of Theorem 1 is the following

COROLLARY 1. For  $0 < \mu < 1$ ,

$$(17) \quad H_{i,n}^{(\mu)} > \sqrt{\pi} (2\mu+1) \Gamma(\mu + \frac{1}{2}) \frac{\Gamma(n)}{(n+2\mu)\Gamma(n+\mu+1)}.$$

Using the alternative representation [5, (15.3.2)] for the Christoffel constants

$$(18) \quad H_{i,n}^{(\mu)} = 2^{2-2\mu} \frac{\pi}{[\Gamma(\mu)]^2} \frac{\Gamma(n+2\mu)}{\Gamma(n+1)} \frac{1}{1-[\xi_{i,n}^{(\mu)}]^2} \frac{1}{[P_n^{(\mu)'(\xi_{i,n}^{(\mu)})]^2},$$

together with (13), we find

**COROLLARY 2.** *When  $0 < \mu < 1$ , the following bound holds*

$$(19) \quad |P_n^{(\mu)'(\xi_{i,n}^{(\mu)})}| < \frac{2}{\Gamma(\mu)} \frac{\Gamma(n+\mu+1)}{\Gamma(n+1)} \frac{1}{1-[\xi_{i,n}^{(\mu)}]^2}.$$

We remark that, when  $0 < \delta < \varphi_{i,n}^{(\mu)} < \pi - \delta$ ,  $\delta$  being a constant, the upper bound given by (19) is of order  $O(n^\mu)$ , while the one obtained by means of (15) and (16) is of order  $O(n^{2\mu+1})$ , with  $0 < \mu < 1$ .

A very crude upper bound for  $H_{i,n}^{(\mu)}$  is

$$H_{i,n}^{(\mu)} < \frac{1}{2} \int_{-1}^1 (1-x^2)^{\mu-1/2} dx = \frac{\sqrt{\pi}}{4} \frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(\mu + 1)},$$

so that from (13) we may immediately derive

**COROLLARY 3.** *When  $0 < \mu < 1$ ,*

$$(20) \quad |P_n^{(\mu)'(\xi_{i,n}^{(\mu)})}| > 4 \frac{\Gamma(\mu+1)}{\Gamma(2\mu)} \frac{\Gamma(n+2\mu)}{\Gamma(n+\mu+1)}.$$

Finally, if we evaluate (9) at the zeros  $x_{j,n}^{(\mu)} = \cos \theta_{j,n}^{(\mu)}$ ,  $j = 1, 2, \dots, n+1$ , of  $E_{n+1}^{(\mu)}(x)$ , we find

$$|\sin \theta_{j,n}^{(\mu)}|^{2\mu-1} \frac{\pi}{2} \frac{\Gamma(2\mu)}{\Gamma(\mu + \frac{1}{2})} |P_n^{(\mu)}(\cos \theta_{j,n}^{(\mu)})| |e_{n+1}^{(\mu)}(\theta_{j,n}^{(\mu)})| > 1,$$

and, taking account of the second inequality in (12),

$$(21) \quad |\sin \theta_{j,n}^{(\mu)}|^{2\mu-1} |P_n^{(\mu)}(\cos \theta_{j,n}^{(\mu)})| > \frac{1}{\sqrt{\pi}} \frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(2\mu)} \frac{\Gamma(n+2\mu)}{\Gamma(n+\mu+1)}.$$

The last inequality implies, in particular, that for the smallest relative maximum of  $|\sin \theta|^{2\mu-1} |P_n^{(\mu)}(\cos \theta)|$ ,  $0 < \mu < 1$ , we have the lower bound given by (21). This follows from the known fact that the  $x_{j,n}^{(\mu)}$ 's interlace with the  $\xi_{i,n}^{(\mu)}$ 's. For the Legendre case ( $\mu = 1/2$ ), we find

$$(22) \quad |P_n(\cos \theta_{j,n}^{(\mu)})| > \frac{1}{\sqrt{\pi}} \frac{\Gamma(n+1)}{\Gamma(n+3/2)} \sim \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{n}},$$

which is reasonably sharp, considering, when  $n$  is even, the smallest relative maximum of  $|P_n(x)|$  is known to be  $|P_n(0)| \sim \sqrt{2/\pi} \cdot 1/\sqrt{\pi}$ .

**3. Bounds for  $E_{n+1}^{(\mu)'(x)}$ .** In this section, we consider the polynomial  $E_{n+1}^{(\mu)}(x)$  and derive lower and upper bounds for its derivative, when  $0 < \mu < 1$ .

From the cosine expansion of  $E_{n+1}^{(\mu)}(\cos \varphi)$ , and inequalities (8), noting that  $|U_n(\cos \varphi)| \leq n+1$ , where  $U_n(\cos \varphi) = \sin(n+1)\varphi/\sin \varphi$ , one easily obtains

$$(23) \quad |E_{n+1}^{(\mu)'(\cos \varphi)}| < \frac{2}{\sqrt{\pi}} \frac{\Gamma(n+\mu+1)}{\Gamma(n+2\mu)} (n+1)^2.$$



To obtain a lower bound for  $|E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})|$ , we make use of the integration rule (7) and recall that for its weights  $B_{j,n}^{(\mu)}$ , the following expression is known [2]

$$B_{j,n}^{(\mu)} = \frac{h_n^{(\mu)}}{k_n^{(\mu)}} \frac{2^n \lambda_0^{(\mu)}}{P_n^{(\mu)}(\cos \theta_{j,n}^{(\mu)}) E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})} > 0,$$

$$B_{j,n}^{(\mu)} = B_{n+2-j,n}^{(\mu)}, \quad j = 1, 2, \dots, n + 1,$$

where

$$h_n^{(\mu)} = \int_{-1}^1 (1-x^2)^{\mu-1/2} [P_n^{(\mu)}(x)]^2 dx$$

and

$$k_n^{(\mu)} = \frac{2^n \Gamma(n + \mu)}{n! \Gamma(\mu)}.$$

Putting  $f(x) = [P_n^{(\mu)}(x)]^2$  in (7), we get

$$h_n^{(\mu)} = 2^n \lambda_0^{(\mu)} \frac{h_n^{(\mu)}}{k_n^{(\mu)}} \sum_{j=1}^{n+1} \frac{P_n^{(\mu)}(\cos \theta_{j,n}^{(\mu)})}{E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})} > 2^{n+1} \lambda_0^{(\mu)} \frac{h_n^{(\mu)}}{k_n^{(\mu)}} \frac{P_n^{(\mu)}(\cos \theta_{j,n}^{(\mu)})}{E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})}.$$

Recalling (21), we thus have proved the following

**THEOREM 2.** For  $0 < \mu < 1$  we have

$$|\sin \theta_{j,n}^{(\mu)}|^{2\mu-1} |E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})| > \frac{4^{1-\mu} \Gamma(n+1)}{\sqrt{\pi} \Gamma(n+\mu)},$$

and, in particular, when  $\frac{1}{2} \leq \mu < 1$ ,

$$|E_{n+1}^{(\mu)'}(\cos \theta_{j,n}^{(\mu)})| > \frac{4^{1-\mu} \Gamma(n+1)}{\sqrt{\pi} \Gamma(n+\mu)}.$$

The inequalities presented in this section could be used, for example, together with (16) and (21), to derive bounds for the numbers  $B_{j,n}^{(\mu)}$ .

**4. Conclusion.** In this paper we have tried to show how one can obtain some inequalities relating to the orthogonal polynomials  $P_n^{(\mu)}(x)$  in a way slightly different from those usually presented. A fundamental role is played by the polynomials  $E_{n+1}^{(\mu)}(x)$  and their connection to the solutions of the differential equation (2), which is expressed by the relations (5) and (6). Particularly important has been inequality (9). Up to the present, very little is known about the polynomials  $E_{n+1}^{(\mu)}(x)$ . We think that new information about them will also lead to new or more accurate results on the polynomials  $P_n^{(\mu)}(x)$ .

REFERENCES

[1] A. S. KRONROD, *Nodes and Weights for Quadrature Formulae. Sixteen-Places tables*, Nauka, Moscow, 1964; English transl., Consultants Bureau, New York, 1965.  
 [2] G. MONEGATO, *A note on extended Gaussian quadrature rules*, Math. Comp., 30 (1976), pp. 812–817.  
 [3] ———, *Positivity of the weights of extended Gauss–Legendre quadrature rules*, *Ibid.*, 32 (1978), pp. 243–245.  
 [4] G. SZEGÖ, *Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören*, Math. Ann., 110 (1934), pp. 501–513.  
 [5] ———, *Orthogonal Polynomials*, vol. 23, 4th ed., American Mathematical Society, Providence, RI, 1975.  
 [6] T. J. Stieltjes, *Correspondance d’Hermite et de Stieltjes*, vol. II, Gauthier-Villars, Paris, 1905, pp. 439–441.

# ON NONLINEAR VOLTERRA EQUATIONS WITH NONINTEGRABLE KERNELS\*

GUSTAF GRIPENBERG†

**Abstract.** In this paper, the asymptotic behavior of solutions of the nonlinear real Volterra equation

$$x(t) + \int_0^t g(x(t-s))a(s) ds = f(t), \quad t \geq 0$$

is studied. Here  $a$  and  $f$  are given and  $x$  is the unknown function. The assumptions on the function  $f$  are rather weak, and in most cases it is assumed that  $\int_0^\infty a(s) ds = +\infty$ .

**1. Introduction.** The purpose of this paper is to study the asymptotic behavior of the solutions of the real Volterra equation

$$(1.1) \quad x(t) + \int_0^t g(x(t-s))a(s) ds = f(t), \quad t \in \mathbb{R}_+ = [0, \infty),$$

where  $a$ ,  $g$  and  $f$  are prescribed real functions and  $x$  is the unknown. We will always assume that a solution  $x$  of (1.1) exists and we note that the a priori bounds on the solution, which we establish below, can be used to prove this existence.

The objective here is not to treat very general kernels  $a$  but to consider cases where  $\int_0^\infty a(s) ds = +\infty$  and where we have to assume very little of the function  $f$ . For earlier investigations of (1.1) and related equations under different assumptions, see [2], [5], [8]–[19]. All our results can immediately be formulated as statements for the integrodifferential equation

$$(1.2) \quad x'(t) + \int_{[0,t]} g(x(t-s)) d\mu(s) = f_0(t), \quad x(0) = x_0, \quad t \in \mathbb{R}_+$$

which corresponds to (1.1) if we take  $a(t) = \mu([0, t])$  and  $f(t) = x_0 + \int_0^t f_0(s) ds$ .

**2. Statement of results.** Recall that the resolvent kernel associated with the (locally integrable) function  $a$  is defined to be the unique solution of the equation

$$(2.1) \quad r(t) + \int_0^t a(t-s)r(s) ds = a(t), \quad t \in \mathbb{R}_+.$$

**THEOREM 1.** *Assume that*

$$(2.2) \quad a: \mathbb{R}_+ \rightarrow \mathbb{R}_+ \text{ is nonincreasing,}$$

$$(2.3) \quad \text{the resolvent } r \text{ associated with } a \text{ belongs to } L^1(\mathbb{R}_+),$$

$$(2.4) \quad g \in C(\mathbb{R}),$$

$$(2.5) \quad \text{there exists a constant } k_1 \geq 0 \text{ such that } g(z) + k_1 \geq g(y) \text{ if } z > y,$$

$$(2.6) \quad \text{there exist constants } k_2 \text{ and } k_3 \text{ such that } |g(z) - g(y)| \leq k_2|z - y| + k_3, \quad z, y \in \mathbb{R},$$

$$(2.7) \quad \liminf_{|y| \rightarrow \infty} y^{-1}g(y) > 0,$$

\* Received by the editors March 7, 1979, and in revised form October 1, 1979.

† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

(2.8)  $f \in L^\infty(\mathbb{R}_+)$ ,

(2.9)  $x \in L^\infty_{loc}(\mathbb{R}_+)$  satisfies (1.1).

Then  $x \in L^\infty(\mathbb{R}_+)$ .

This theorem is closely related to [5, Thm. 1], where (2.3) and (2.7) are replaced by the assumption that  $a \in L^1(\mathbb{R}_+)$ , (then (2.3) is automatically satisfied but this fact is not needed there). In [5, Thm. 2], the assumptions (2.5)–(2.7) are replaced by  $(y - y_0)g(y) \geq 0$  for some  $y_0 \in \mathbb{R}$  and it is moreover assumed that  $\log a$  is convex. In the case when  $f \in BV(\mathbb{R}_+)$ , there are several boundedness results which do not depend on (2.3) and where the assumptions on the function  $g$  are weaker than the ones used here; see, e.g., [2], [11]. For cases where  $f$  satisfies some integrability conditions, see [5], [16], [17]. Note finally that (2.3) is not a consequence of (2.2) (see [3]), and that this assumption is clearly necessary. For some sufficient conditions for (2.3) to hold, see [4], [6], [7], [13].

As noted above, Theorem 1 is an advance over earlier results only in the case when  $\int_0^\infty a(s) ds = +\infty$ . If moreover  $\lim_{t \rightarrow \infty} a(t) > 0$ , then one can prove a stronger assertion (in this case (2.3) again holds automatically).

**THEOREM 2.** *Assume that (2.4), (2.6), (2.8) and (2.9) hold and that*

(2.10)  $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is such that  $\text{var}(a; \mathbb{R}_+) < a(0)$ ,

(2.11)  $\limsup_{y \rightarrow \infty} g(y) > 0, \quad \liminf_{y \rightarrow \infty} g(y) > -\infty,$

$\liminf_{y \rightarrow -\infty} g(y) < 0, \quad \limsup_{y \rightarrow -\infty} g(y) < \infty,$

either  $\limsup_{y \rightarrow \infty} g(y) = \infty$  or

(2.12)  $\limsup_{y \rightarrow \infty} |g(y+z) - g(y)| = 0$  for all  $d > 0$  and either

$\liminf_{y \rightarrow -\infty} g(y) = -\infty$  or  $\limsup_{y \rightarrow -\infty} |g(y+z) - g(y)| = 0$

for all  $d > 0$ .

Then  $x \in L^\infty(\mathbb{R}_+)$ .

It is quite clear that some assumptions like (2.11) are necessary to guarantee the boundedness of the solution. To see that something more is needed, take  $f \in C^1(\mathbb{R}_+) \cap L^\infty(\mathbb{R}_+)$ ,  $f(0) = 0$ , and define  $g$  by  $g(y) = f'(y) - 1$ ,  $y \geq 0$  and  $g(y) = y + f'(0) - 1$ ,  $y < 0$ . If  $a(t) \equiv 1$ , then  $x(t) = t$  satisfies (1.1).

Next we proceed to study the asymptotic behavior of bounded solutions of (1.1). First we have

**THEOREM 3.** *Assume that (2.3), (2.4) hold and that*

(2.13)  $a \in BV(\mathbb{R}_+)$ ,

(2.14)  $\int_0^\infty a(s) ds = +\infty,$

(2.15)  $\text{Re} \int_{[0, \infty)} e^{i\omega t} d\mu(t) \geq 0, \omega \in \mathbb{R}$  where  $\mu([0, t]) = a(t), t \in \mathbb{R}_+$ ,

(2.16)  $\text{if } \operatorname{Re} \int_{[0, \infty)} e^{i\omega t} d\mu(t) = 0 \text{ for some } \omega \in \mathbb{R}, \text{ then } \int_{[0, \infty)} e^{i\omega t} d\mu(t) = 0,$

(2.17)  $\operatorname{var}(r; [t, \infty)) \in L^1(\mathbb{R}_+),$

(2.18)  $f$  is a measurable function on  $\mathbb{R}_+$  such that

$$\limsup_{t \rightarrow \infty} \sup_{|s| \leq d} |f(t+s) - f(t)| = 0 \text{ for all } d > 0,$$

(2.19)  $x \in L^\infty(\mathbb{R}_+)$  satisfies (1.1),

and either (2.8) holds or

(2.20)  $\lim_{t \rightarrow \infty} a(t) \neq 0.$

Then

(2.21)  $\limsup_{t \rightarrow \infty} \sup_{|s| \leq d} |x(t+s) - x(t)| = 0 \text{ for all } d > 0$

and

(2.22)  $\lim_{t \rightarrow \infty} g(x(t)) = 0.$

Theorem 3 is an extension of results in [12], [19] (with the exception that  $a \in L^1(\mathbb{R}_+)$  is considered in [19]) and the proof relies heavily on these results. The main difference compared with Theorem 3 is that in [12], [19] it is assumed that  $\operatorname{var}(a; [t, \infty)) \in L^1(\mathbb{R}_+)$ , whereas we invoke this condition on the resolvent kernel  $r$ . Note also that if  $\lim_{t \rightarrow \infty} a(t) = 0$  and  $\operatorname{var}(a; [t, \infty)) \in L^1(\mathbb{R}_+)$ , then (2.14) cannot hold. As an example of a function  $a$ , for which (2.3) and (2.17) are satisfied, take  $a = a_1 + a_2$  such that  $a_1$  and  $-a_1'$  are nonnegative, nonincreasing and convex on  $\mathbb{R}_+$ ,  $a_2 \in BV(\mathbb{R}_+)$ ,  $\operatorname{var}(a_2; [t, \infty)) \in L^1(\mathbb{R}_+)$  and such that  $\int_0^\infty e^{-st} a(t) dt \neq -1$ ,  $\operatorname{Re} s \geq 0$ , (this is the case if (2.15) and (2.16) hold); for details see [6].

In the next theorem, we replace (2.13), (2.15) and (2.16) by the stronger assumption (2.2) and instead of (2.17) we use a technical condition saying that  $a$  is not a constant on arbitrarily long intervals.

THEOREM 4. Assume that (2.2)–(2.4), (2.14), (2.18) and (2.19) hold and that either (2.8) and

(2.23)  $\text{there exists } t_0 > 0 \text{ such that}$   
 $(a(t) - a(t + t_0))^{-1} \in L^\infty_{\text{loc}}(\mathbb{R}_+)$

or (2.20) hold. Then (2.21) and (2.22) hold.

If  $\lim_{t \rightarrow \infty} a(t) \neq 0$ , then Theorem 4 is very close to some results in [14], [15] where more general kernels are treated but where, on the other hand, somewhat more is assumed of the function  $g$  (i.e., that it should vanish or be a constant on no interval). Consequently the main interest is in the case when  $\lim_{t \rightarrow \infty} a(t) = 0$ , a case studied only under stronger assumptions on  $f$  in [14], [15]. In contrast to Theorem 3, the proof of Theorem 4 does not (once (2.3) is known to hold) depend on transform techniques.

If we drop the assumption that  $f \in L^\infty(\mathbb{R}_+)$ , then it is no longer true (at least in the linear case  $g(x) = x$ ) that (2.22) follows from the assumptions of Theorems 3 or 4 when  $\lim_{t \rightarrow \infty} a(t) = 0$ . But if we assume more of  $a$  and  $g$  then (2.21) still holds. This will be a

consequence of the following result for the corresponding limit equation (cf. [9]),

$$(2.24) \quad x'(t) + \int_{[0,\infty)} g(x(t-s)) \, d\mu(s) = 0 \quad \text{a.e. } t \in \mathbb{R}.$$

**THEOREM 5.** *Assume that (2.2) and (2.4) hold and that*

$$(2.25) \quad a \text{ is convex,}$$

$$(2.26) \quad x \in L^\infty(\mathbb{R}) \cap AC_{loc}(\mathbb{R}) \text{ satisfies (2.24)}$$

$$\text{where } \mu([0, t]) = a(t), t \in \mathbb{R}_+,$$

$$(2.27) \quad g \text{ is nondecreasing.}$$

*Then  $x$  is a constant. If instead of (2.27)*

$$(2.28) \quad g \text{ is nonincreasing}$$

*then  $x$  is monotone nonincreasing or nondecreasing and if  $x$  is not a constant, then*

$$(2.29) \quad \left( \int_0^\infty a(s) \, ds \right)^{-1} \in \left[ \inf_{L \in G_1} L, \sup_{L \in G_1} L \right],$$

$$G_1 = \left\{ |g(z) - g(y)|/|z - y| \mid \inf_{t \in \mathbb{R}} x(t) \leq z < y \leq \sup_{t \in \mathbb{R}} x(t) \right\}.$$

For other investigations of the limit equation (2.24) and its relation to (1.1) and (1.2), see [8], [9], [16], [19]. For the existence of nonconstant solutions of (2.24) under the assumption (2.28), see [1]. Concerning (1.1), we have

**COROLLARY 1.** *Assume that (2.2), (2.4), (2.18), (2.19), (2.25) and either (2.27) or (2.28) and*

$$(2.30) \quad \left( \int_0^\infty a(s) \, ds \right)^{-1} \notin \left[ \inf_{L \in G_2} L, \sup_{L \in G_2} L \right],$$

$$G_2 = \left\{ |g(z) - g(y)|/|z - y| \mid \liminf_{t \rightarrow \infty} x(t) \leq z < y \leq \limsup_{t \rightarrow \infty} x(t) \right\}$$

*hold. Then (2.21) holds.*

From the proof of Theorem 5 we can also deduce the following result for the equation

$$(2.31) \quad x''(t) + \int_{[0,\infty)} g(x(t-s)) \, d\mu(s) = 0 \quad \text{a.e. } t \in \mathbb{R}.$$

**COROLLARY 2.** *Assume that (2.2), (2.4), (2.25) and (2.28) hold and that*

$$(2.32) \quad x \in L^\infty(\mathbb{R}) \cap C^1(\mathbb{R}), x' \in AC_{loc}(\mathbb{R}) \text{ satisfies the equation (2.31)}$$

$$\text{where } \mu([0, t]) = a(t), t \in \mathbb{R}_+.$$

*Then  $x$  is a constant.*

Using this corollary we obtain

**COROLLARY 3.** *Assume that (2.4), (2.18), (2.19) and (2.28) hold and that*

$$(2.33) \quad a \in C^1(\mathbb{R}_+), a(0) = 0, a \neq 0, a' \text{ is nonnegative, nonincreasing and convex.}$$

*Then (2.21) and (2.22) hold.*

Observe that it is a very strong assumption to assume that  $x$  is bounded in Corollaries 2 and 3.

**3. Proof of Theorem 1.** Define the function  $z$  by

$$(3.1) \quad z(t) = x(t) - f(t), \quad t \in R_+.$$

By (1.1), this function satisfies the equation

$$(3.2) \quad z(t) + \int_0^t a(t-s)g_1(z(s)) ds = F(t), \quad t \in R_+,$$

where

$$(3.3) \quad g_1(y) = \sup_{0 \leq v \leq y} g(v), \quad y \geq 0, \quad g_1(y) = \inf_{y \leq v < 0} g(v), \quad y < 0$$

and

$$F(t) = \int_0^t a(t-s)(g_1(z(s)) - g(x(s))) ds, \quad t \in R_+.$$

Since  $|g_1(y) - g(y)| \leq k_1, y \in R$  by (2.5) and (3.3), it follows from (2.2), (2.6), (2.8) and (3.1) that

$$(3.4) \quad F \text{ is Lipschitz continuous on } R_+.$$

Choose real constants  $c_1, c_2 > 0$  such that

$$(3.5) \quad y^{-1}g(y) \geq c_1 \quad \text{when } |y| \geq c_2.$$

By (2.7), this is possible. Let  $c_3, c_4 \in (0, 1)$  be some constants which satisfy

$$(3.6) \quad (c_1^{-1}c_3 + c_4) \int_0^\infty |r(s)| ds \leq 2^{-1},$$

where  $r$  is the resolvent associated with  $a$ .

Suppose that  $x \notin L^\infty(R_+)$ . Then, by (2.8) and (3.1), there exists a sequence  $\{t_n\}$  such that  $|z(t_n)| \rightarrow \infty$  as  $t_n \rightarrow \infty$ . We may assume (see (2.4) and (2.7)) that the following conditions are satisfied (if necessary replace  $x$  and  $f$  by  $-x$  and  $-f$ , and  $g$  by the function  $g_2(y) = -g(-y)$ )

$$(3.7) \quad z(t_n) \rightarrow +\infty \quad \text{as } n \rightarrow \infty,$$

$$(3.8) \quad z(t_n) = \sup_{0 \leq s \leq t_n} |z(s)|, \quad n \geq 1,$$

$$(3.9) \quad g_1(z(t_n)) = \sup_{0 \leq s \leq t_n} |g_1(z(s))|, \quad n \geq 1.$$

Define the numbers  $T_n$  by

$$(3.10) \quad T_n = \inf \{t \geq 0 | z(s) \geq (1 - c_3)z(t_n), g_1(z(s)) \geq (1 - c_4)g_1(z(t_n)) \text{ on } [t, t_n]\}.$$

Proceeding in the same way as in the proof of [5, Thm. 1], we deduce from (2.2), (2.4)–(2.6), (3.2)–(3.4) and (3.7)–(3.10) that

$$(3.11) \quad t_n - T_n \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

(Observe that the assumption  $a \in L^1(R_+)$  was not essential at this point in the proof of [5, Thm. 1].)

Next we conclude from (2.4)–(2.6), (3.1), (3.3), (3.9) and (3.10) that there exist constants  $c_5$  and  $c_6$  such that for all  $n$ ,

$$(3.12) \quad |g(x(t)) - g_1(z(t_n))| \leq c_4 g_1(z(t_n)) + c_5 \quad \text{on } [T_n, t_n],$$

$$(3.13) \quad |g(x(t))| \leq g_1(z(t_n)) + c_6 \quad \text{on } [0, T_n].$$

By (1.1) and (3.1), we have

$$g(x(t)) + \int_0^t a(t-s)g(x(s)) ds = g(x(t)) - z(t), \quad t \in \mathbb{R}_+,$$

and solving  $g(x(t))$  from the left side of this equation with the aid of the resolvent associated with  $a$  (see (2.1)), we get

$$z(t) - \int_0^t r(t-s)z(s) ds + \int_0^t r(t-s)g(x(s)) ds = 0, \quad t \in \mathbb{R}_+.$$

This equation yields, for every  $n$ ,

$$(3.14) \quad \begin{aligned} &g_1(z(t_n)) \int_0^{t_n - T_n} r(s) ds + z(t_n) \left( 1 - \int_0^{t_n - T_n} r(s) ds \right) \\ &\leq \sup_{T_n \leq t \leq t_n} (|g(x(t)) - g_1(z(t_n))| + |z(t) - z(t_n)|) \int_0^\infty |r(s)| ds \\ &\quad + \sup_{0 \leq t \leq T_n} (|g(x(t))| + |z(t)|) \int_{t_n - T_n}^\infty |r(s)| ds. \end{aligned}$$

We may safely assume that  $\int_0^\infty a(s) ds = +\infty$  since the other case is covered in [5, Thm. 1]. But then it is easy to see from (2.1) that  $\int_0^\infty r(s) ds = 1$  and so we get a contradiction from (2.3), (3.3) and (3.5)–(3.14) (note that  $|g(x(t)) - g_1(z(t_n))| + |z(t) - z(t_n)| \leq (c_3 c_1^{-1} + c_4) |g_1(z(t_n))| + c_5$ ). This completes the proof of Theorem 1.

**4. Proof of Theorem 2.** Suppose that  $g(x(t)) \notin L^\infty(\mathbb{R}_+)$ . We again introduce the definition (3.1) and note that  $z$  satisfies the equation

$$(4.1) \quad z(t) + \int_0^t a(t-s)g(z(s)) ds = F(t), \quad t \in \mathbb{R}_+,$$

where we now have  $F(t) = \int_0^t a(t-s)(g(z(s)) - g(x(s))) ds$ ,  $t \in \mathbb{R}_+$ . From (2.6), (2.8), (2.10) and (3.1) we conclude that (3.4) still holds and that  $g(z(t)) \notin L^\infty(\mathbb{R}_+)$ . Consequently, we can choose a sequence  $\{t_n\}$  such that  $|g(z(t_n))| \rightarrow \infty$  as  $t_n \rightarrow \infty$ . We may again assume this sequence to be such that

$$(4.2) \quad g(z(t_n)) = \sup_{0 \leq t \leq t_n} |g(z(t))|, \quad n \geq 1.$$

Since  $x \in L_{loc}^\infty(\mathbb{R}_+)$  and since (2.4), (2.8), (2.10), (3.4) and (4.1) hold, we see that  $z$  is locally Lipschitz continuous and differentiating (4.1) we get (writing  $a(t) = a(0) + \int_0^t da(s)$ )

$$(4.3) \quad z'(t) + a(0)g(z(t)) + \int_0^t g(z(t-s)) da(s) = F'(t) \quad \text{a.e. } t \in \mathbb{R}_+.$$

It follows from (2.10) and (4.2) that

$$a(0)g(z(t_n)) + \int_0^{t_n} g(z(t_n-s)) da(s) \geq (a(0) - \text{var}(a; \mathbb{R}_+))g(z(t_n)), \quad n \geq 1,$$

and using this inequality together with (3.4) in (4.3) we conclude that  $z'(t) < 0$  for a.e.  $t$  in a neighborhood of  $t_n$  when  $n$  is sufficiently large. This is a contradiction since by (2.11) we may assume that  $z(t) \leq z(t_n)$  when  $t \leq t_n$  for all  $n$ .

Suppose next that  $g(x(t)) \in L^\infty(\mathbb{R}_+)$  but  $x \notin L^\infty(\mathbb{R}_+)$ . By (2.8) and (3.1) we also have  $z \notin L^\infty(\mathbb{R}_+)$  and consequently there exists a sequence  $\{t_n\}$  such that  $|z(t_n)| \rightarrow \infty$  as  $t_n \rightarrow \infty$ . We may without loss of generality assume that

$$(4.4) \quad z(t_n) = \sup_{0 \leq t \leq t_n} z(t), \quad n \geq 1$$

and by (2.11) also

$$(4.5) \quad g(z(t_n)) \geq c_1 > 0, \quad n \geq 1$$

for some constant  $c_1$ .

It is easy to see from (1.1), (2.10), (3.1) and the assumption that  $g(x(t)) \in L^\infty(\mathbb{R}_+)$  that

$$(4.6) \quad z \text{ is Lipschitz continuous on } \mathbb{R}_+.$$

Since we clearly cannot have  $\limsup_{y \rightarrow \infty} g(y) = \infty$ , we must by (2.13) have  $\lim_{y \rightarrow \infty} \sup_{|z| \leq d} |g(y+z) - g(y)| = 0$  for all  $d > 0$ . This fact combined with (2.8), (3.1), (4.5) and (4.6) implies that there exists a sequence  $\{T_n\}$  such that (3.11) holds and for every  $n$ ,

$$(4.7) \quad \begin{aligned} g(x(t)) &\geq 2^{-1}c_1 \quad \text{and} \\ \text{var}(a; \mathbb{R}_+) \text{ess sup}_{T_n \leq s \leq t} |g(x(t)) - g(x(s))| &\leq 2^{-2}c_1(a(0) - \text{var}(a; \mathbb{R}_+)) \end{aligned}$$

for a.e.  $t \in (t_n - 1, t_n)$ .

We have by routine estimates

$$\begin{aligned} a(0)g(x(t)) + \int_0^t g(x(t-s)) da(s) \\ \geq a(t)g(x(t)) - \text{var}(a; \mathbb{R}_+) \text{ess sup}_{T_n \leq s \leq t} |g(x(t)) - g(x(s))| \\ - 2 \text{var}(a; [t - T_n, \infty)) \|g(x(t))\|_{L^\infty(\mathbb{R}_+)} \quad \text{for a.e. } t \in (t_n + 1, t_n), n \geq 1. \end{aligned}$$

It follows from this inequality and (1.1), (3.1), (3.11) and (4.7) that when  $n$  is large enough, then  $z'(t) < 0$  for a.e.  $t \in (t_n - 1, t_n)$ . But this statement combined with (4.4) yields a contradiction and the proof of Theorem 2 is completed.

**5. Proof of Theorem 3.** Using the resolvent equation (2.1), one easily sees that (1.1) implies

$$(5.1) \quad x(t) + \int_0^t r(t-s)g_1(x(s)) ds = f(t) - \int_0^t r(t-s)f(s) ds, \quad t \in \mathbb{R}_+,$$

where now  $g_1(y) = g(y) - y$ . Since (2.3) was assumed to hold, it follows directly from (2.1) and (2.14) by Laplace transform arguments that

$$(5.2) \quad \int_0^\infty r(s) ds = 1.$$



Suppose that  $\lim_{t \rightarrow \infty} a(t) = 0$  and (2.8) holds. By (2.3), (2.8), (2.18) and (5.2), we have (observe that (1.1), (2.4) (2.13) and (2.19) yield  $f \in L^\infty_{loc}(\mathbb{R}_+)$ )

$$(5.3) \quad f(t) - \int_0^t r(t-s)f(s) ds \in L^\infty(\mathbb{R}_+), \quad \lim_{t \rightarrow \infty} \left( f(t) - \int_0^t r(t-s)f(s) ds \right) = 0.$$

Now we want to apply [19, Corollary 3b] to (5.1) and in view of (2.4), (2.17), (5.2), (5.3) and the definition of  $g_1$ , we have only to check that the assumptions (2.13), (2.15) and (2.16) hold when  $a$  is replaced by  $r$ . In [19, Corollary 3b] it is assumed that  $x$  is a bounded solution of (1.1), (2.4), (2.8), (2.13), (2.15) and (2.16) hold (with the unnecessary restriction that  $\text{Re } \hat{\mu}(\omega) = 0$  for at most denumerably many  $\omega$ , see below) and that  $\text{var} (a : [t, \infty)) \in L^1(\mathbb{R}_+)$ ,  $\hat{\mu}(0) = 0$  and  $\lim_{t \rightarrow \infty} f(t) = f(\infty)$  exists. Then it is asserted that (2.21) and  $\lim_{t \rightarrow \infty} (x(t) + \int_0^\infty a(s) ds g(x(t))) = f(\infty)$  hold.

We immediately obtain from (2.1), (2.3) and (2.13) that  $r \in BV(\mathbb{R}_+)$ . Let  $\nu$  be the finite Borel measure defined by  $\nu([0, t]) = r(t)$ ,  $t \in \mathbb{R}_+$ . From (2.1) it follows that

$$(5.4) \quad \hat{\nu}(\omega) = (\hat{\mu}(\omega) - i\omega |\hat{\mu}(\omega)/\omega|^2) |1 + i\hat{\mu}(\omega)/\omega|^{-2}, \quad \omega \in \mathbb{R},$$

where we have used the definition  $\hat{\mu}(\omega) = \int_{[0, \infty)} e^{i\omega t} d\mu(t)$ . Now it is clear from (5.4) that we can replace  $a$  by  $r$  in (2.15) and (2.16) as we note that it follows from (5.2) that  $\hat{\nu}(0) = 0$ . An application of [19, Corollary 3b] shows that (2.21) and (2.22) hold in this case, but note that we must use [12, Lemma 4] in the proof of [19, Lemma 2.1].

Assume next that  $\lim_{t \rightarrow \infty} a(t) \neq 0$ . Write  $f$  in the form  $f = f_1 + f_2$ , where

$$(5.5) \quad f_1 \in L^\infty(\mathbb{R}_+), \quad f_2 \in C^1(\mathbb{R}_+) \quad \text{and} \quad \lim_{t \rightarrow \infty} f_1(t) = 0 = \lim_{t \rightarrow \infty} f_2'(t).$$

Put  $R(t) = 1 - \int_0^t r(s) ds$ ,  $t \in \mathbb{R}_+$ . Since (2.20) holds, it follows from (2.13), (2.15) and (2.17) that (see [13])

$$(5.6) \quad R \in L^1(\mathbb{R}_+).$$

An integration by parts yields

$$f_2(t) - \int_0^t r(t-s)f_2(s) ds = R(t)f_2(0) + \int_0^t R(t-s)f_2'(s) ds,$$

and we see from (5.5) and (5.6) that (5.3) still holds. Now we can complete the proof for the case when  $\lim_{t \rightarrow \infty} a(t) \neq 0$  in the same way as above.

**6. Proof of Theorem 4.** Define the function  $h$  by  $h(t) = f(t) - \int_t^{t+1} f(s) ds$ ,  $t \in \mathbb{R}_+$ . We observe that

$$(6.1) \quad h \in L^\infty(\mathbb{R}_+), \quad \lim_{t \rightarrow \infty} h(t) = 0,$$

and if we define the function  $F$  by

$$(6.2) \quad F(t) = f(t) - h(t) + \int_0^t a(t-s)(g(x(s)) - h(s) - g(x(s))) ds, \quad t \in \mathbb{R}_+,$$

then

$$(6.3) \quad F \in AC_{loc}(\mathbb{R}_+), \quad F' \in L^\infty(\mathbb{R}_+) \quad \text{and} \quad \lim_{t \rightarrow \infty} F'(t) = 0.$$

This follows by (2.2), (2.4), (2.18) and (2.19). Putting

$$(6.4) \quad z(t) = x(t) - h(t), \quad t \in \mathbb{R}_+,$$

we now have

$$(6.5) \quad z(t) + \int_0^t a(t-s)g(z(s)) \, ds = F(t), \quad t \in \mathbb{R}_+.$$

The next step is to establish the crucial

LEMMA 6.1. *Let the hypothesis of Theorem 4 hold and let  $\delta, T$  be arbitrary positive constants. Then there exist a positive constant  $T_0$  and an integer  $N_0$  such that for all  $n \geq N_0$  there exist closed intervals  $I_n \subset [nT_0, (n+1)T_0]$  such that  $m(I_n) \geq T$  and  $|g(z(t))| \leq \delta, t \in I_n$ .*

*Proof.* Arguing in the same way as in [10, p. 853], we conclude from (2.20), (6.4) and (6.5) that there exist constants  $c_1$  and  $c_2$  such that for any  $t_1, t_2, 0 < t_1 < t_2 < \infty$ ,

$$(6.6) \quad \begin{aligned} & - \int_{t_1}^{t_2} \int_0^u (g(z(u)) - g(z(u-s)))^2 \, da(s) \, du + a(\infty) \int_{t_1}^{t_2} (g(z(s)))^2 \, ds \\ & \leq c_1 + c_2 \int_{t_1}^{t_2} |F'(s)| \, ds + 2^{-1} \int_0^{t_1} (g(z(t_1-s)))^2 (a(s) - a(\infty)) \, ds \\ & \quad - 2^{-1} \int_0^{t_2} (g(z(t_2-s)))^2 (a(s) - a(\infty)) \, ds, \end{aligned}$$

where  $a(t) = a(0) + \int_0^t da(s)$  and  $a(\infty) = \lim_{t \rightarrow \infty} a(t)$ . The following estimate is a consequence of (2.2) ( $c_2 = \|g(z(t))\|_{L^\infty(\mathbb{R}_+)}$ ).

$$(6.7) \quad \begin{aligned} & \int_0^{t_1} (g(z(t_1-s)))^2 (a(s) - a(\infty)) \, ds - \int_0^{t_2} (g(z(t_2-s)))^2 (a(s) - a(\infty)) \, ds \\ & \leq c_2^2 \int_0^{t_2-t_1} (a(s) - a(\infty)) \, ds. \end{aligned}$$

This inequality and (6.3) imply that the right side of the inequality in (6.6) divided by  $t_2 - t_1$  can be made arbitrarily small, provided one first chooses  $t_2 - t_1$  to be large and then  $t_1$  to be large. This shows, since we may assume that  $a(t) \neq a(\infty)$  (otherwise we can invoke Theorem 3), that the hypotheses used in the proof of [10, Lemma 4] are satisfied. Hence we conclude that if  $\delta_1$  and  $T_1$  are positive constants, then there exists a positive constant  $T_2$  and an integer  $N_2$  such that for all  $n \geq N_2$  there exist closed intervals  $J_n$  such that

$$(6.8) \quad J_n \subset [nT_2, (n+1)T_2], \quad m(J_n) \geq T_1, \quad n \geq N_2$$

and

$$(6.9) \quad \sup \{|g(z(s)) - g(z(t))| \mid s, t \in J_n\} \leq \delta_1, \quad n \geq N_2.$$

If  $a(\infty) > 0$ , then the conclusion of the lemma follows from (6.6)–(6.9) by a simple argument. Just note that  $g(z(t))$  cannot behave like a nonzero constant on large intervals.

Suppose that  $a(\infty) = 0$ . By (2.2) and (6.3) we may differentiate (6.5) and we obtain

$$(6.10) \quad z'(t) + a(0)g(z(t)) + \int_0^t g(z(t-s)) \, da(s) = F'(t) \quad \text{a.e. } t \in \mathbb{R}_+.$$

Let  $0 < t_1 < t_2 < \infty$  be arbitrary. From (2.2) we have

$$(6.11) \quad \left| a(0)g(z(t_2)) + \int_0^{t_2} g(z(t_2-s)) da(s) \right| \leq a(0) \sup_{t_1 \leq t \leq t_2} |g(z(t_2)) - g(z(t))| + 2a(t_2 - t_1) \sup_{t \geq 0} |g(z(t))|.$$

Since  $\lim_{t \rightarrow \infty} a(t) = 0$  and (2.4), (2.19), (6.1), (6.3), (6.4) and (6.8)–(6.11) hold, we see that if  $\delta_3$  and  $T_3$  are given constants then there exist a positive constant  $T_4$  and an integer  $N_4$  such that for all  $n \geq N_4$  there exist closed intervals  $K_n$  such that

$$(6.12) \quad K_n \subset [nT_4, (n+1)T_4], \quad m(K_n) \geq T_3, \quad n \geq N_4,$$

$$(6.13) \quad \sup \{|x(s) - x(t)| \mid s, t \in K_n\} \leq \delta_3, \quad n \geq N_4$$

and

$$(6.14) \quad \sup \{|g(x(s)) - g(x(t))| \mid s, t \in K_n\} \leq \delta_3, \quad n \geq N_4.$$

Again we conclude that (5.1) holds. Let  $0 < t_1 < t_2 < \infty$  be arbitrary. We have

$$(6.15) \quad \begin{aligned} |g(x(t_2))| \left| \int_0^{t_2-t_1} r(s) ds \right| &\leq \left| x(t_2) + \int_0^{t_2} r(t_2-s)(g(x(s)) - x(s)) ds \right| \\ &\quad + |x(t_2)| \left| 1 - \int_0^{t_2-t_1} r(s) ds \right| \\ &\quad + \sup_{t_1 \leq t \leq t_2} (|g(x(t)) - g(x(t_2))| + |x(t) - x(t_2)|) \int_0^\infty |r(s)| ds \\ &\quad + \text{ess sup}_{t \geq 0} (|x(t)| + |g(x(t))|) \int_{t_2-t_1}^\infty |r(s)| ds. \end{aligned}$$

Let  $\delta$  be arbitrary. By (2.3), (2.4), (2.19), (5.1)–(5.3) and (6.12)–(6.15) we observe that if we choose  $\delta_3 \leq 2^{-1}$  sufficiently small and  $T_3 \geq T$  large enough and let  $t_1$  and  $t_2$  be the endpoints of the interval  $K_n$  for some  $n$  sufficiently large, then it follows that  $|g(x(t_2))| \leq 2^{-1}\delta$ . Consequently the assertion of Lemma 6.1 holds with  $g(z(t))$  replaced by  $g(x(t))$ . But the desired conclusion now follows by (2.4), (6.1) and (6.4) and the proof of Lemma 6.1 is completed.

Suppose that  $\limsup_{t \rightarrow \infty} |g(x(t))| > 0$ . We may without loss of generality assume that  $g_+ \stackrel{\text{def}}{=} \limsup_{t \rightarrow \infty} g(x(t)) > 0$ . Let  $\{t_n\}$  be a sequence such that

$$(6.16) \quad g(z(t_n)) \rightarrow g_+ \quad \text{as } t_n \rightarrow \infty.$$

Moreover, we choose this sequence so that

$$(6.17) \quad \text{for any integer } m, \text{ the set } \{t \mid z(t) \leq z(t_n)\} \cap (t_n - m^{-1}, t_n) \text{ is nonempty.}$$

If this is impossible to achieve, then  $\lim_{t \rightarrow \infty} z(t)$  exists. But then it follows from (2.4) and Lemma 6.1 that  $\lim_{t \rightarrow \infty} g(z(t)) = 0$  and so (2.22) follows by (2.4), (6.1) and (6.4).

Choose  $\delta = 2^{-1}g_+$  and  $T \geq t_0$ . For every  $n$  sufficiently large, let  $[q_n, s_n]$  be one of the intervals given by Lemma 6.1 so that

$$(6.18) \quad 0 < t_n - s_n \leq 2T_0.$$

By (2.2), Lemma 6.1 and our choice of  $\delta$  we obtain for all  $n$  sufficiently large,

$$\begin{aligned}
 (6.19) \quad & a(t_n)g(z(t_n)) + \int_0^{t_n} (g(z(t_n - s)) - g(z(t_n))) da(s) \\
 & \geq a(\infty)g(z(t_n)) - a(0) \sup_{2^{-1}t_n \leq s \leq t_n} (g(z(s)) - g(z(t_n))) \\
 & \quad - 2(a(2^{-1}t_n) - a(\infty)) \sup_{t \geq 0} |g(z(t))| \\
 & \quad + (a(t_n - s_n) - a(t_n - q_n))(g(z(t_n)) - 2^{-1}g_+).
 \end{aligned}$$

Now it follows from (2.2) and either (2.20) or (2.23) combined with (6.3), (6.10), (6.16), (6.18), (6.19), the continuity of  $g(z(t))$ , the choice of  $T$  and the definition of  $g_+$  that if we choose  $n$  large enough, then  $z'(t) < 0$  for a.e.  $t \in (t_n - m^{-1}, t_n)$  for some positive integer  $m$ . But this statement contradicts (6.17) and hence (2.22) holds. It is an easy consequence of (2.4), (2.22), (6.1), (6.3), (6.4) and (6.10) that (2.21) also holds. This completes the proof of Theorem 4.

**7. Proof of Theorem 5.** By (2.2), (2.4) and (2.24)–(2.26) we may clearly assume that  $x$  is continuously differentiable and (2.24) holds for all  $t \in \mathbb{R}$ .

Write (2.24) in the form

$$(7.1) \quad x'(t) = -a(\infty)g(x(t)) - \int_{-\infty}^t a'(t-s)(g(x(s)) - g(x(t))) dt, \quad t \in \mathbb{R}.$$

It is not difficult to see from this equation combined with (2.2) and (2.27) or (2.28) that  $x$  cannot be bounded unless  $g(x(t)) \equiv 0$  or  $\lim_{t \rightarrow \infty} a(t) = a(\infty) = 0$ . Hence we may assume that  $a(\infty) = 0$ .

We claim that if (2.27) holds but  $x$  is not a constant or if (2.28) holds but  $x$  is not monotone (nondecreasing or nonincreasing) then there exist points  $s_0, t_0, -\infty < s_0 < t_0 < \infty$  such that

$$(7.2) \quad g(x(t)) \neq g(x(s_0)) = g(x(t_0)) \quad \text{when } t \in (s_0, t_0).$$

To see this we observe by (2.2) and (7.1) that if (2.27) holds then  $g(x(t))$  is monotone if and only if  $x$  is a constant and if (2.28) holds then  $g(x(t))$  is monotone if and only if  $x(t)$  is monotone.

We may without loss of generality assume that  $g(x(t)) < g(x(t_0)), t \in (s_0, t_0)$ . By (2.27) or (2.28) we can choose  $t_1 \in (s_0, t_0)$  to be the smallest number such that

$$(7.3) \quad g(x(t_1)) = \min_{s_0 \leq t \leq t_0} g(x(t)) \quad \text{and} \quad x'(t_1) = 0.$$

Let  $(c)_+ = \max\{0, c\}$  and  $(c)_- = -(-c)_+$ . We claim that

$$(7.4) \quad p_1 \stackrel{\text{def}}{=} - \int_{-\infty}^{t_1} a'(t_1 - s)(g(x(s)) - g(x(t_1))) ds < 0.$$

If this is not the case then it follows from (7.1) since  $x'(t_1) = 0$  that  $\int_{-\infty}^{t_1} a'(t_1 - s)|g(x(s)) - g(x(t_1))| ds = 0$  and we deduce from (2.25), (2.27) or (2.28), (7.1) and our choice of  $t_1$  that  $x'(t) = 0, t > t_1$  which is a contradiction by (7.2). Now it is possible to define the real number  $s_1$  by

$$(7.5) \quad s_1 = \sup \{t < t_1 \mid g(x(t)) < g(x(t_1))\}.$$

Next we are going to define three sequences  $\{t_i\}$ ,  $\{p_i\}$  and  $\{s_i\}$  with the following properties for all  $i \geq 1$ :

$$(7.6) \quad s_i < t_{i+1} \leq s_{i-1} < t_i,$$

$$(7.7) \quad g(x(t_i)) \geq g(x(t_0)), \quad i \text{ even} \quad \text{and} \quad g(x(t_i)) \leq g(x(t_1)), \quad i \text{ odd},$$

$$(7.8) \quad (-1)^i g(x(t)) \leq (-1)^i g(x(t_i)) = (-1)^i g(x(s_i)), \quad s_i \leq t \leq t_i,$$

$$(7.9) \quad g(x(t_{i+1})) = (-1)^{i+1} \max_{s_i \leq t \leq t_i} \{(-1)^{i+1} g(x(t))\}, \quad x'(t_{i+1}) = 0,$$

$$(7.10) \quad p_i = \int_{-\infty}^{t_i} a'(t_i - s)((-1)^i (g(x(s)) - g(x(t_i))))_+ ds,$$

$$(7.11) \quad p_{i+1} \leq p_i$$

and

$$(7.12) \quad s_i = \sup \{t < t_i \mid (-1)^i g(x(t)) > (-1)^i g(x(t_i))\}.$$

The numbers  $s_0, t_1, p_1$  and  $s_1$  have already been defined above. It is a consequence of the monotonicity of  $g$  that we may choose the numbers  $t_i$  to satisfy  $x'(t_i) = 0$ . Since (7.4) holds, we have only to check that (7.11) holds as then all the other requirements are easily seen to be satisfied (especially the crucial fact  $s_i > -\infty$  for all  $i$ ). At this point we may assume that (7.6), (7.8) and (7.9) hold. We have by (2.2), (2.25), (7.1), (7.6), (7.8) and (7.9),

$$\begin{aligned} & \int_{-\infty}^{t_{i+1}} a'(t_{i+1} - s)((-1)^{i+1} (g(x(s)) - g(x(t_{i+1}))))_+ ds \\ &= \int_{-\infty}^{t_{i+1}} a'(t_{i+1} - s)((-1)^i (g(x(s)) - g(x(t_{i+1}))))_+ ds \\ &\leq \int_{-\infty}^{t_{i+1}} a'(t_{i+1} - s)((-1)^i (g(x(s)) - g(x(t_i))))_+ ds \\ &\cong \int_{-\infty}^{t_{i+1}} a'(t_i - s)((-1)^i (g(x(s)) - g(x(t_i))))_+ ds = p_i, \end{aligned}$$

and (7.11) follows.

From (2.2), (2.4) and (2.26) we conclude that  $x$  is Lipschitz continuous, hence  $g(x(t))$  is uniformly continuous. Combining this fact with (7.2), (7.6) and (7.7) we see that  $\lim_{i \rightarrow \infty} t_i = -\infty$  and so by (7.6), (7.8) and (7.9) we have  $\lim_{i \rightarrow \infty} g(x(t_{2i})) = \limsup_{t \rightarrow -\infty} g(x(t))$ . But this statement gives a contradiction in view of (2.2), (7.4), (7.10) and (7.11) and the desired conclusion follows.

Assume next that (2.28) holds,  $x$  is not a constant and (2.29) does not hold. The first case to consider is

$$(7.13) \quad L_0 \int_0^\infty a(s) ds > 1,$$

where  $L_0 = \inf_{L \in G_1} L > 0$ . We may, without loss of generality, assume that  $x$  is nondecreasing. It is not difficult to see from (2.2) and (7.1) that  $x'(t)$  is uniformly continuous on  $\mathbf{R}$ . Consequently there exists  $t_0$  so that  $x'(t_0) = \max_{t \in \mathbf{R}} x'(t)$ . Since  $L_0 > 0$  we see from (2.25) and (7.1) that  $x'(t) > 0$  for all  $t \in \mathbf{R}$  and we can choose a sequence  $\{t_n\}$ ,  $t_n \rightarrow \infty$  so

that  $x'(t_n) = \min_{t_0 \leq t \leq t_n} x'(t)$ . Then we have by (2.2), (2.28), (7.1) and the definition of  $L_0$ ,

$$\begin{aligned} x'(t_n) &\geq -L_0 x'(t_n) \int_{t_0}^{t_n} a'(t_n - s)(t_n - s) \, ds \\ &= L_0 x'(t_n) \int_0^{t_n - t_0} (a(s) - a(t_n - t_0)) \, ds. \end{aligned}$$

But since  $x'(t_n) > 0$  for all  $n$ , we get a contradiction by (7.13) if we choose  $t_n$  large enough.

The second case that we have to consider is

$$(7.14) \quad L_1 \int_0^\infty a(s) \, ds < 1,$$

where  $L_1 = \sup_{L \in G_1} L$  (note that if  $L_1 = 0$  then it follows immediately from (7.1) that  $x$  is a constant). Again we assume that  $x$  is nondecreasing and choose  $t_0$  so that  $x'(t_0) = \max_{t \in \mathbb{R}} x'(t)$ . From (2.2), (2.28), (7.1) and the definition of  $L_1$  we deduce that

$$x'(t_0) \leq -L_1 x'(t_0) \int_{-\infty}^{t_0} a'(t_0 - s)(t_0 - s) \, ds = L_1 x(t_0) \int_0^\infty a(s) \, ds.$$

Since  $x'(t_0) > 0$  we have a contradiction by (7.14). This completes the proof of Theorem 5.

**8. Proof of Corollaries 1, 2 and 3.** To prove Corollary 1, it is sufficient to show that (2.22) holds with  $x$  replaced by  $z$  as defined in (6.4) since (6.1) holds. Applying (6.3), Theorem 5 and [9, Thm. 1a] to the equation (6.10), we obtain the conclusion of Corollary 1.

To establish Corollary 2 we have to make the following changes in the proof of Theorem 5. We replace equation (7.1) by

$$(8.1) \quad x''(t) + a(\infty)g(x(t)) + \int_{-\infty}^t a'(t-s)(g(x(s)) - g(x(t))) \, ds = 0, \quad t \in \mathbb{R}$$

and we may assume that  $x \in C^2(\mathbb{R})$ . From (2.2) and (2.28) we again conclude that  $x$  cannot be bounded unless  $g(x(t)) \equiv 0$  or  $a(\infty) = 0$ . Let us assume below that  $a(\infty) = 0$ . By (2.2) and (2.28) we also see that  $x$  cannot be monotone (nondecreasing or nonincreasing) if it is not a constant. In the rest of the proof of Theorem 5 we have only to choose the numbers  $t_i, i \geq 1$  so that  $(-1)^i x''(t_i) \geq 0$  instead of demanding  $x'(t_i) = 0$ . This is possible by (2.28). The fact that  $x'(t_i) = 0, i \geq 0$  was only used in deriving (7.4) and (7.11) with the aid of (7.1), and we can just as well use  $(-1)^i x''(t_i) \geq 0$  and (8.1) to achieve the same results. Finally we note that since  $x'$  is now Lipschitz continuous by (8.1), it follows that  $x$  is Lipschitz continuous too, since it is bounded. This completes the proof of Corollary 2.

For the proof of Corollary 3, we define the function  $h$  by  $h(t) = f(t) - 2 \int_t^{t+1} \int_t^s f(u) \, du \, ds, t \in \mathbb{R}_+$ . From (2.18) we deduce that (6.1) holds and if the function  $F$  is defined by (6.2), then it follows from (2.4), (2.18), (2.19) and (2.33) that

$$(8.2) \quad F \in C^1(\mathbb{R}_+), \quad F' \in AC_{loc}(\mathbb{R}_+), \quad F'' \in L^\infty(\mathbb{R}_+) \quad \text{and} \quad \lim_{t \rightarrow \infty} F''(t) = 0.$$

Defining the function  $z$  by (6.4), we see from (2.33) and (8.2) that (6.5) holds,

$z \in C^1(\mathbb{R}_+)$ ,  $z' \in AC_{loc}(\mathbb{R}_+)$  and

$$(8.3) \quad z''(t) + a'(0)g(z(t)) + \int_0^t a''(t-s)g(z(s)) ds = F''(t) \quad \text{a.e. } t \in \mathbb{R}_+.$$

Proceeding in the same way as in the proof of [9, Thm. 1a], we conclude that if  $\{t_n\}$  is a sequence such that  $t_n \rightarrow \infty$ , then there exists a subsequence (also denoted by  $\{t_n\}$ ), and a function  $y \in C^2(\mathbb{R})$  such that

$$(8.4) \quad z(t - t_n) \rightarrow y(t) \text{ uniformly on compact subsets of } \mathbb{R} \text{ as } n \rightarrow \infty,$$

where  $y$  satisfies the equation

$$y''(t) + a'(0)g(y(t)) + \int_{-\infty}^t a''(t-s)g(y(s)) ds = 0, \quad t \in \mathbb{R}.$$

But by (2.28) and (2.33) we know that  $y$  must be a constant, and hence, (2.21) follows from (6.1), (6.4) and (8.4).

We have still to show that (2.22) holds. Define the function  $v$  by

$$(8.5) \quad v(t) = x(t) - f(t), \quad t \in \mathbb{R}_+.$$

By (1.1) and (2.33) we see that  $v \in C^2(\mathbb{R}_+)$  and

$$(8.6) \quad v'(t) + \int_0^t a'(t-s)g(x(s)) ds = 0, \quad t \in \mathbb{R}_+.$$

From this equation we conclude by (2.4), (2.19) and (2.33) that  $v'$  is Lipschitz continuous. Hence (2.18), (2.21) and (8.5) yield

$$(8.7) \quad \lim_{t \rightarrow \infty} v'(t) = 0.$$

Let  $r$  be the resolvent kernel associated with  $a'$ . Using the resolvent equation (2.1) we obtain from (8.6)

$$(8.8) \quad \int_0^t r(t-s)g(x(s)) ds = \int_0^t r(t-s)v'(s) ds - v'(t), \quad t \in \mathbb{R}_+.$$

It is a consequence of (2.33) that  $r \in L^1(\mathbb{R}_+)$  (see [13]), and hence, it follows from (2.4), (2.19), (2.21), (8.7) and (8.8) that  $\lim_{t \rightarrow \infty} g(x(t)) = 0$  (recall that since  $a' \neq 0$  we have  $\int_0^\infty r(s) ds > 0$ ). This completes the proof of Corollary 3.

REFERENCES

[1] O. DIEKMANN, *Limiting behaviour in an epidemic model*, Nonlinear Anal., Theory, Methods Appl., 1 (1977), pp. 459-470.  
 [2] G. GRIPENBERG, *Bounded solutions of a Volterra equation*, J. Differential Equations, 28 (1978), pp. 18-22.  
 [3] ———, *A Volterra equation with nonintegrable resolvent*, Proc. Amer. Math. Soc., 73 (1979), pp. 57-60.  
 [4] ———, *Integrability properties of resolvents of Volterra equations*, Rep. HTKK-MAT-A117, Helsinki Univ. of Technology, Helsinki, Finland, 1978.  
 [5] ———, *On the boundedness of solutions of Volterra equations*, Indiana Math. J., 28 (1979), pp. 279-290.  
 [6] ———, *On the asymptotic behavior of resolvents of Volterra equations*, this Journal, this issue, pp. 654-662.  
 [7] G. S. JORDAN AND R. L. WHEELER, *A generalization of the Wiener-Levy theorem applicable to some Volterra equations*, Proc. Amer. Math. Soc., 57 (1976), pp. 109-114.  
 [8] J. J. LEVIN, *On some geometric structures for integrodifferential equations*, Advances in Math., 22 (1976), pp. 146-186.

- [9] J. J. LEVIN AND D. F. SHEA, *On the asymptotic behaviour of the bounded solutions of some integral equations, I–III*, J. Math. Anal. Appl., 37 (1972), pp. 42–82, pp. 288–326, pp. 537–575.
- [10] S-O. LONDEN, *On the asymptotic behavior of the bounded solutions of a nonlinear Volterra equation*, this Journal, 5 (1974), pp. 849–875.
- [11] ———, *On the variation of the solutions of a nonlinear integral equation*, J. Math. Anal. Appl., 52 (1975), pp. 430–449.
- [12] ———, *On a Volterra integrodifferential equation with  $L^\infty$ -perturbation and non-countable zero-set of the transforming kernel*, J. Integral Equations, to appear.
- [13] D. F. SHEA AND S. WAINGER, *Variants of the Wiener–Levy theorem, with some applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.
- [14] C. C. SHILEPSKY, *A note on the asymptotic behavior of an integral equation*, Proc. Amer. Math. Soc., 33 (1972), pp. 111–113.
- [15] ———, *The asymptotic behavior of an integral equation with an application to Volterra’s population equation*, J. Math. Anal. Appl., 48 (1974), pp. 764–779.
- [16] O. J. STAFFANS, *On the asymptotic spectra of the bounded solutions of a nonlinear Volterra equation*, J. Differential Equations, 24 (1977), pp. 365–382.
- [17] ———, *Boundedness and asymptotic behavior of solutions of a Volterra equation*, Michigan Math. J., 24 (1977), pp. 77–95.
- [18] ———, *Some energy estimates for a nondifferentiated Volterra equation*, J. Differential Equations, 32 (1979), pp. 285–293.
- [19] ———, *On a nonlinear integral equation with a nonintegrable perturbation*, Rep. HTKK-MAT-A148, Helsinki Univ. of Technology, Helsinki, Finland, 1979.



## STABILITY OF AN AGE-DEPENDENT POPULATION\*

FRANK J. S. WANG†

**Abstract.** This paper considers a nonlinear deterministic population model in which the death rate rises as the population grows. It is an age-dependent version of a logistic population whose growth is controlled by limited resources and is in the form of a partial differential equation with respect to time and age. We prove that the solution of our equation behaves asymptotically just like the solution of the logistic equation  $dN/dt = N(a - bN)$  and is, therefore, globally asymptotically stable. This implies that there are no steady oscillations, and that in the long run the population size and age-structure become fixed, independent of the initial conditions. Possible applications in fish and animal population dynamics are studied.

**1. Introduction.** The logistic equation  $dN/dt = N(a - bN)$  has proved to be a very useful model for population growth. It applies when environmental pressures force the death rate up or the birth rate down as the population grows. Its solution has been applied, with remarkable success, to fit the growth curves of various types of populations (see [9], [10], [11]). Among deterministic models, the chief disadvantage of the logistic models is that they yield no information concerning the age distribution of the population and, in fact, are based on the assumption that the birth and death processes are age-independent. The age-dependent analogue of the logistic model was first proposed by Von Foerster [13], in connection with cell populations. It has then been generalized by several authors—Gurtin [8], Griffel [6] and Rorres [12]; and the existence and stability of their solutions has been extensively studied. It is the purpose of this paper to propose a different age-dependent version of the logistic model and investigate the stability of its solutions.

We define the age-density function  $p(x, t)$  such that for any  $a, b$  the number of members at time  $t$  with age between  $a$  and  $b$  is

$$\int_a^b p(x, t) dx.$$

Here,  $p$  is a continuous function, representing some kind of smoothing or statistical average of the true integer-valued population size. We shall assume that the mean number  $\beta(x)$  of offsprings produced per unit time by an individual of age  $x$  is independent of both  $t$  and the size of the population. Hence, the birth process is described by the integral equation

$$p(0, t) = \int_0^\infty \beta(x)p(x, t) dx, \quad t \geq 0.$$

Here,  $\beta(t)$  is assumed to be nonnegative and continuous.

Suppose the struggle for survival is dominated by competition with other individuals in the population. Then individuals of different ages should be weighted differently to account for their possible different ecological impacts on the surrounding environment, e.g., the average food consumed in a day by an individual may be different in different age-groups. Let  $c(x)$ , a nonnegative function on  $[0, \infty)$ , represent the weighing function. We therefore assume that the death probability of an individual at time  $t$  is a

---

\* Received by the editors October 4, 1978, and in revised form April 13, 1979.

† Mathematics Department, University of Montana, Missoula, Montana 59812.

function of the “size”

$$s(t) = \int_0^\infty c(x)p(x, t) dx$$

after the adjustment according to the age-distribution of the population (see [12]) More precisely, we assume that there exist a nondecreasing function  $\lambda$  such that  $\lambda(s(t))$  is the probability density function at time  $t$  of an individual’s dying at age  $x$ ; i.e.,  $\lambda(s(t)) \Delta t + o(\Delta t)$  is the probability of an individual of age  $x$  dying in the interval  $(t, t + \Delta t]$ ; we assume that this is independent of  $x$ .

Since it is usually true that the more producing power (producing offsprings, of course) the individual has, the better physical condition it has and is thus less susceptible to competition from other individuals; it is reasonable to assume that the weight  $c(x)$  we put on an individual of age  $x$  is proportional to its birth rate  $\beta(x)$ . Without loss of generality, we shall take the constant of proportionality to be one and thus assume  $\beta(x) = c(x)$  and  $s(t) = p(0, t)$ . Applications in fish and animal population dynamics will be given in the last section. This will illustrate the validity of the underlying biological assumptions for some species.

In § 2, we set up the basic equation of the model and reduce the problem to the solution of an integral equation. We then transform this integral equation into a linear renewal equation and relate the asymptotic behavior of  $s(t)$  to the asymptotic behavior of the solution of the renewal equation. In § 3, we use some well-known properties of the solutions of the renewal equations to show that, under appropriate conditions on  $\beta(x)$  and  $\lambda(s)$ ,  $s(t)$  is globally asymptotically stable; that is,  $s(t)$  tends to a limit  $s^* \geq 0$  which is independent of the initial conditions, and  $s^* > 0$  if and only if

$$\mu = \int_0^\infty \beta(x) dx,$$

the mean number of offspring born to a member surviving to a great age, is greater than 1.

**2. Basic equations.** Consider the group of individuals who are of age  $x$  at time  $t$ . If  $t$  is increased by  $h$  units, these individuals age by  $h$  units; thus, assuming that  $p$  has partial derivatives,

$$\lim_{h \rightarrow 0} \frac{p(x+h, t+h) - p(x, t)}{h} = p_x + p_t$$

is the rate at which the population of this group is changing in time. Since this rate plus the number  $\lambda(s(t))p(x, t)$  of individuals (per unit age and time) of age  $x$  who died at  $t$  must equal to zero, we obtain

$$(1) \quad p_x + p_t + \lambda(s(t))p(x, t) = 0,$$

where

$$(2) \quad s(t) = p(0, t) = \int_0^\infty \beta(x)p(x, t) dx.$$

Let  $\varphi$  be the age-density function for the initial population, i.e.,

$$(3) \quad p(x, 0) = \varphi(x) \quad \text{for } x \geq 0.$$

The p.d.e. (1) together with the integral equation (2) and the initial condition (3) constitutes our model.

Equation (1), with the initial condition (3), can easily be solved by integrating along characteristics. One finds that

$$\begin{aligned}
 (4) \quad p(x, t) &= s(t-x) \exp\left(-\int_{t-x}^t \lambda(s(u)) du\right) & 0 \leq x \leq t, \\
 &= \varphi(x-t) \exp\left(-\int_0^t \lambda(s(u)) du\right) & x \geq t \geq 0.
 \end{aligned}$$

Substituting (4) into (2), we obtain

$$\begin{aligned}
 (5) \quad s(t) &= \int_t^\infty \beta(x)\varphi(x-t) \exp\left(-\int_0^t \lambda(s(u)) du\right) dx \\
 &+ \int_0^t \beta(x)s(t-x) \exp\left(-\int_{t-x}^t \lambda(s(u)) du\right) dx,
 \end{aligned}$$

Define

$$F(t) = s(t) \exp\left(\int_0^t \lambda(s(u)) du\right) \quad \text{and} \quad g(t) = \int_t^\infty \beta(x)\varphi(x-t) dx,$$

and rewrite the preceding nonlinear integral equation for  $s(t)$  into a linear renewal equation for  $F(t)$ :

$$(6) \quad F(t) = g(t) + \int_0^t \beta(x)F(t-x) dx.$$

The mean number  $\mu = \int_0^\infty \beta(x) dx$  of offsprings born to a member surviving to a great age is crucial for the general behavior of  $F(t)$ . The following three results are part of the standard textbook literature (see [1], [5]).

(2.1) Suppose that  $\mu < 1$  and  $g(t) \rightarrow 0$  as  $t \rightarrow \infty$ , then  $F(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

(2.2) Suppose that  $\mu = 1$  and  $g(t)$  is integrable, then

$$\lim_{t \rightarrow \infty} F(t) = \int_0^\infty g(t) dt / \int_0^\infty x\beta(x) dx,$$

where the fraction is defined to be zero if the denominator is infinity.

(2.3) Suppose that  $\mu > 1$  and  $e^{-\alpha t}g(t)$  is integrable over  $[0, \infty)$  where  $\alpha$ , the Malthusian parameter for  $\beta(x)$ , is a real constant such that

$$\int_0^\infty e^{-\alpha t}\beta(t) dt = 1.$$

Then

$$F(t) \sim e^{\alpha t} \left( \int_0^\infty g(x) e^{-\alpha x} dx \right) \left( \int_0^\infty x\beta(x) e^{-\alpha x} dx \right)^{-1},$$

where  $\sim$  indicates the fact that the ratio of these two functions tends to one as  $t \rightarrow \infty$ .

**3. Global stability.** We now study the asymptotic behavior of the solution  $s(t)$  of the integral equation (5). Since  $\varphi(x) = p(x, 0)$  is the age-density function for the initial population, it should be nonnegative and integrable and its integral over  $[0, \infty)$  is the

initial population size. If the mean number  $\mu$  of offspring born to a member is less than 1, we would certainly expect the population to die out. The proof for this is immediate.

**THEOREM 1.** *If  $\mu = \int_0^\infty \beta(x) dx < 1$ ,  $\varphi(x)$  and  $\lambda(s)$  are nonnegative and  $\varphi(x)$  is integrable, then the solution  $s(t)$  of (5) tends to zero as  $t \rightarrow \infty$ .*

*Proof.* Since  $\beta$  and  $\varphi$  are both integrable,  $g(t) \rightarrow 0$  as  $t \rightarrow \infty$ . This implies by (2.1) that

$$F(t) = s(t) \cdot \exp \left( \int_0^t \lambda(s(u)) du \right) \rightarrow 0$$

as  $t \rightarrow \infty$ . Thus  $s(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

Consider the case  $\mu = 1$ . We shall show that if  $g(t)$  is differentiable and  $\lambda(s) \neq 0$  for all  $s \neq 0$ , then  $s(t)$  also approaches zero as  $t \rightarrow \infty$ . Assuming  $g(t)$  is differentiable and differentiating (6), we obtain

$$F'(t) = g'(t) + \beta(t)g(0) + \int_0^t \beta(x)F'(t-x) dx.$$

This is again a renewal equation. Since  $\int_0^\infty g(t) dt$ , being less than  $\mu$  times the initial population size  $\int_0^\infty \varphi(x) dx$ , is integrable, (2.2) implies  $\lim_{t \rightarrow \infty} F'(t)/F(t) = 0$ . Since

$$F(t) = s(t) \cdot \exp \left( \int_0^t \lambda(s(u)) du \right), \quad F'(t)/F(t) = [s'(t) + s(t)\lambda(s(t))]/s(t),$$

or equivalently,

$$(7) \quad s'(t) = s(t)[\varepsilon(t) - \lambda(s(t))],$$

where  $\varepsilon(t) = F'(t)/F(t) \rightarrow 0$  as  $t \rightarrow \infty$ . We are now ready to prove the next theorem.

**THEOREM 2.** *If  $\mu = \int_0^\infty \beta(x) dx = 1$ ,  $\lambda(s)$  is continuous,  $g(t) = \int_0^\infty \beta(x)\varphi(x-t) dx$  is differentiable, and  $\lambda(s) > 0$  for all  $s > 0$ , then the solution  $s(t)$  of (5) tends to zero as  $t \rightarrow \infty$ .*

*Proof.* Let  $\hat{s} = \limsup s(t)$  and  $\bar{s} = \liminf s(t)$ . Suppose that  $\hat{s} > \bar{s} \geq 0$ . Then it is seen that there exists a sequence  $\{t_n\}$  such that  $t_n \rightarrow \infty$ ,  $s(t_n) \rightarrow \hat{s} > 0$  and  $s'(t_n) > 0$ . Since  $\varepsilon(t_n) \rightarrow 0$  and  $\lambda(s(t_n)) \rightarrow \lambda(\hat{s}) > 0$  as  $n \rightarrow \infty$ , it follows from (7) that  $s'(t_n) < 0$  for all  $n$  sufficiently large, but this contradicts the choice of  $t_n$ . Thus  $\hat{s} = \bar{s}$  and  $\lim_{t \rightarrow \infty} s(t) = c$  exists. Suppose that  $c > 0$ , then (7) implies  $\lim_{t \rightarrow \infty} s'(t) = c[-\lambda(c)] < 0$ , contradicting the fact that  $s(t) \geq 0$  for all  $t$ . Thus  $c = 0$  and this completes the proof.

Consider the case  $\mu > 1$ . We shall show that  $s(t)$  approaches equilibrium. In the rest of this paper, we shall assume that  $\lambda(x)$  is nonnegative and continuous,  $\varphi(x)$  and  $\beta(t)$  are such that  $g(t)$  is differentiable and that the equation  $\alpha - \lambda(s) = 0$  (where  $\alpha$  is the Malthusian parameter for  $\beta(x)$ ) has at most one solution. We will denote the solution of  $\alpha - \lambda(s) = 0$  by  $s^*$  if it exists.

Using (2.3), an argument similar to the one used in obtaining (7) gives us the following differential equation:

$$(8) \quad s'(t) = s(t)[\alpha - \lambda(s(t)) + \alpha\varepsilon_1(t)],$$

where  $\varepsilon_1(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

Recall that  $\hat{s} = \limsup s(t)$  and  $\bar{s} = \liminf s(t)$ .

Consider the case  $\lambda(s) < \alpha$  for all  $s > 0$ . Suppose that  $\bar{s} < \hat{s} \leq \infty$ . Then there exist a sequence  $\{t_n\}$  such that  $s(t_n) \rightarrow \bar{s}$  and  $s'(t_n) < 0$ . It then follows from (8) and our assumption about  $\lambda$  that  $s'(t_n) \geq 0$  for  $n$  large, contradicting the choice  $t_n$ . This shows  $\lim s(t) = c \leq \infty$ . Suppose  $c < \infty$ , then  $s'(t) \rightarrow c[\alpha - \lambda(c)] > 0$ . This implies  $s(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , a contradiction. Thus  $\lim s(t) = \infty$ .

If  $\inf_{s \geq 0} \lambda(s) > \alpha$ , then  $s'(t)$  will be negative for all  $t$  large and thus  $s(t)$  is bounded above. Suppose that  $0 \leq \bar{s} < \hat{s}$ , then there exist a sequence  $\{t_n\}$  such that  $t_n \rightarrow \infty$ ,  $s(t_n) \rightarrow \hat{s}$  and  $s'(t_n) = 0$ . It then follows from (8) and the boundedness of  $s(t)$  that  $\hat{s}[\alpha - \lambda(\hat{s})] = 0$  and thus  $\hat{s} = \bar{s} = 0$ . This shows  $\lim s(t) = 0$ .

Consider the most interesting case where there exists exactly one  $s^*$  such that  $\lambda(s^*) = \alpha$ . Suppose that  $s(t) \rightarrow \infty$ . Then there exist  $T$  and  $s^T$  such that  $s(t) \geq s^T > s^*$  for all  $t > T$  and  $\delta = \alpha - \lambda(s^T) + \alpha \sup_{t \geq T} \varepsilon_1(t) < 0$ . Thus  $s'(t) < s(t)\delta < 0$  for all  $t > T$ , a contradiction. Suppose that  $\bar{s} < \hat{s} = \infty$ . Pick a sequence  $\{t_n\} \rightarrow \infty$  such that  $s(t_n) \rightarrow \infty$  and  $s'(t_n) > 0$ . We can show by (8) that  $s'(t_n) < 0$  for sufficiently large  $n$ , contradicting the choice of  $t_n$ . Suppose that  $\bar{s} = 0$ . Since  $s(t) \geq 0$  for all  $t$ , there exists a sequence  $t_n \rightarrow \infty$  such that  $s'(t_n) \leq 0$  and  $s(t_n) \rightarrow 0$ . By letting  $n$  be large enough, we can make  $s(t_n)$  so close to  $\bar{s} = 0$  that  $\alpha - \lambda(s(t_n)) > \alpha - \lambda(s^*) = 0$ . But this implies by (8) that  $s'(t_n) > 0$  for all large  $n$ ; this again contradicts the choice of  $t_n$ . Suppose now that  $0 < \bar{s} < \hat{s} < \infty$ , then there exist sequences  $\{t_n\}$  and  $\{\bar{t}_n\}$  such that  $t_n \rightarrow \infty$ ,  $s(t_n) \rightarrow \hat{s}$ ,  $s'(t_n) = 0$ ;  $\bar{t}_n \rightarrow \infty$ ,  $s(\bar{t}_n) \rightarrow \bar{s}$ ,  $s'(\bar{t}_n) = 0$ . Replacing  $t$  by  $t_n$  in (8) and letting  $n \rightarrow \infty$ , we obtain that  $\hat{s}[\alpha - \lambda(\hat{s})] = 0 = \bar{s}[\alpha - \lambda(\bar{s})]$ . Our assumption about  $\lambda$  implies that  $\hat{s} = \bar{s} = s^*$ . Thus  $s(t) \rightarrow s^*$  as  $t \rightarrow \infty$ . We summarize our results in the next theorem.

**THEOREM 3.** *Suppose that  $\mu = \int_0^\infty \beta(x) dx > 1$ ,  $\lambda(s)$  is nonnegative and continuous,  $\varphi(x)$  and  $\beta(x)$  are nonnegative and are such that  $g(t)$  is differentiable; then*

$$\lim_{t \rightarrow \infty} s(t) = \begin{cases} \infty & \text{if } \lambda(s) < \alpha \text{ for all } s > 0, \\ 0 & \text{if } \inf_{s \geq 0} \lambda(s) > \alpha, \\ s^* & \lambda(s) = \alpha \text{ has exactly one root,} \end{cases}$$

where  $\alpha$  and  $s^*$  are such that

$$\int_0^\infty e^{-\alpha x} \beta(x) dx = 1$$

and  $\lambda(s^*) = \alpha$ .

The above result shows that when there is exactly one positive root satisfying

$$R(s) = \int \beta(x) e^{-\lambda(s)x} dx = 1,$$

the population generally approaches a fixed size and age-structure, independent of the initial conditions. Note that  $R(s)$  is the expected number of offsprings born to a single individual during its lifetime when the size of the population remains fixed at  $s$ . If  $R(s^*) = 1$ , then for such  $s^*$  the expected number of offsprings is one and so an equilibrium population exists. The equilibrium age distribution, i.e., solutions  $p(x, t)$  of (1) and (2) which are independent of time  $t$ , can easily be found by (4). One finds that  $p(x, \infty) = s^* \cdot \exp(-\alpha \cdot x)$ . Thus the total size of the population tends to

$$\int_0^\infty p(x, \infty) dx = s^*/\alpha$$

as  $t \rightarrow \infty$ .

**4. Applications in fish and animal population dynamics.** In analyzing the population dynamics of a fish stock and the effect of fishing on it, Schaefer (see [7]) uses the logistic population model. He considers the size of the population as the biomass  $B$ , or equivalently, the total weight of all the fish in the population. The basic assumption of Schaefer's model is that the biomass  $B$  of the population will tend to increase toward

some limit  $B_\infty$ , set by the environment, and that the rate of increase is a function of the biomass.

However very young fish (with the exception of sharks and other fish which, when born are already comparatively large) are so small and usually differ so much in food requirements, distribution, etc. from their parents, that it is both simpler, and in many ways more realistic, to omit fish below some chosen size or age from most of the analysis (see [7]). It is thus reasonable to assume that the rate of natural increase (or decrease) of a stock is determined by the magnitude of the biomass  $B'$  of the adult fish population (see [7]). This assumption is even more reasonable, say, in a fishery management model, a model used to determine the proper fishery strategy. Since, in this case, the fishing effort depends on the abundance of the adult stock  $B'$ , the fishing mortality (one major constituent part of the total mortality of fish, the other one is natural mortality) is in some way proportional to fishing effort.

Let  $p(x, t)$  be the density function of the number of fish of age  $x$  at time  $t$ , and

$$c(x) = \text{the average weight of adult fish of age } x, \\ = 0 \text{ for young fish.}$$

Since size/age relationship plays an important roll in studying fish population dynamics, a size/age curve can usually be obtained in the literature on fish population dynamics (see [3], [4], [7], [14]).

Fishery scientists generally suppose that fish fecundity (i.e., egg productivity) is a function of the weight of the fish (see [2], [3], [4]). Many fishery scientists use the term "relative fecundity," that is, the number of eggs per unit weight of fish (see [2], [3], [4]). This method of expressing the productivity assumes that the relation between fecundity and weight is linear. In many cases, e.g., long rough dab [2], haddock, plaice [3] and trout for the Northwest river of Tasmania [14], it has been proved to be so.

In [3], the authors (see p. 62 of [3]) obtain an expression for the fertile egg-productivity of a fish population in terms of its age- and size-structure and abundance. Let  $r$  be the relative fecundity of females and assume that the mortality and growth coefficients are the same for both sexes, so that the proportion of females at any age can be denoted by a constant  $s$ . The annual egg-production of the population is then given by

$$s \cdot r \cdot B' = \int_0^\infty s \cdot rc(x)p(x, t) dt.$$

Thus  $\beta(x)$  is proportional to  $c(x)$  and the constant of proportionality is  $s \cdot r$ .

Our model is, in some way, an improvement to the Schaefer's model mentioned at the beginning of this section. The results of § 3 imply that, with the complication of age structure, the total biomass of the adult population grows just like a logistic curve, at least, when  $t$  gets large.

Suppose for some population, e.g., animal or bird populations, the young individual becomes adult at a certain age, say  $k$ . Suppose each individual lives for at most  $l$  years and that the reproductive rate is about the same throughout their adult ages. For example, for Wyoming antelope,  $k = 1.5$  (18 months),  $l = 9$  and the reproductive rate is about the same among the different age groups from 1.5 to 9 years of age. (See [15]). To model such a population, we put

$$\beta(x) = \begin{cases} 0 & \text{if } x < k \text{ or } x > l, \\ \beta & \text{if } k \leq x \leq l, \end{cases}$$

and let  $c(x) = \beta(x)/\beta$ . Then

$$s(t) = \int_0^{\infty} c(x)p(x, t) dx = \int_k^t p(x, t) dx$$

is just the size of the adult population at time  $t$ . In some species the young individuals differ very much in food requirements and usually are too weak to have a noticeable impact on the surrounding environment; it is thus reasonable to assume that the death rate increases if the size of the adult population increases, i.e.,  $\lambda$  is a function of  $s(t)$  rather than of  $\int_0^{\infty} p(x, t) dx$ . Certainly, the model would be much more realistic and useful if we assume that the death probability is a function of both the age and the size of the adult population. In this case, it is reasonable to believe that under appropriate conditions on  $\lambda$ , e.g.,  $\lambda(x, s)$  is a nondecreasing function on  $s$  for each fixed  $x$ , the "size"  $s(t)$  of the population will also behave asymptotically just like the logistic curve.

#### REFERENCES

- [1] K. B. ATHREYA AND P. E. NEY, *Branching Processes*, Springer-Verlag, New York, 1972.
- [2] T. B. BAGENAL, *A short review of fish fecundity*, The Biological Basis of Freshwater Fish Populations, S. D. Gerking, ed., Adlard and Son Ltd., Dorking, Great Britain, 1967.
- [3] R. J. H. BEVERTON AND S. T. HOLT, *On the Dynamics of Exploited Fish Populations*, Her Majesty's Stationery Office, London, 1957.
- [4] D. H. CUSHING, *Washington sea grant program*, Division of Marine Resources, Univ. of Washington, July 1973.
- [5] W. FELLER, *An Introduction to Probability Theory and Its Application*, Vol. II, John Wiley, New York, 1966.
- [6] D. H. GRIFFEL, *Age-dependent population growth*, J. Inst. Math. Appl., 17 (1976), pp. 141–152.
- [7] J. A. GULLAND, *The analysis of data and development of models*, Fish Population Dynamics, J. A. Gulland, ed., John Wiley and Sons, New York, 1977.
- [8] M. E. GURTIN AND R. C. MACCAMY, *Non-linear age-dependent population dynamics*, Arch. Rational Mech. Anal., 54 (1974), pp. 281–300.
- [9] N. KEYFITZ, *An Introduction to the Mathematics of Population*, Addison-Wesley, Reading, MA, 1968.
- [10] A. J. LOTKA, *Elements of Mathematical Biology*, Dover, New York, 1956.
- [11] E. C. PIELOU, *An Introduction to Mathematical Ecology*, Wiley-Interscience, New York, 1969.
- [12] C. RORRES, *Stability of an age specific population with density dependent fertility*, Theoret. Population Biology, 10 (1976), pp. 26–46.
- [13] H. VONFOERSTER, *Some remarks on changing populations*, The Kinetics of Cellular Proliferation, Grune and Stratton, New York, 1959.
- [14] A. H. WEATHERLEY, *Growth and Ecology of Fish Populations*, Academic Press, New York, 1972.
- [15] Wyoming Game and Fish Commission Publication, special antelope issue, Wyoming Game and Fish Dept., Cheyenne, WY, June, 1966.

## SOME HYPERGEOMETRIC ORTHOGONAL POLYNOMIALS\*

JAMES A. WILSON†

**Abstract.** Explicit formulas and orthogonality relations are given for some polynomials which include as special or limiting cases the classical polynomials, related polynomials with discrete orthogonalities, some polynomials of Pollaczek and the 6-*j* symbols of angular momentum.

**1. Introduction.** This paper contains derivations and discussion of some polynomial orthogonality relations which include as special or limiting cases the orthogonalities for the 6-*j* symbols of quantum mechanics, the classical polynomials and many related families of orthogonal polynomials. In [7], we give properties of these polynomials which extend the recurrence relations, differential equations and Rodrigues formulas for the classical polynomials.

The polynomials may be expressed as hypergeometric series:

$$\begin{aligned}
 p_n(t^2) &= p_n(t^2; a, b, c, d) \\
 (1.1) \quad &= (a+b)_n(a+c)_n(a+d)_n \sum_{k=0}^n \frac{(-n)_k(a+b+c+d+n-1)_k(a-t)_k(a+t)_k}{(a+b)_k(a+c)_k(a+d)_k k!} \\
 &= (a+b)_n(a+c)_n(a+d)_n {}_4F_3\left(\begin{matrix} -n, a+b+c+d+n-1, a-t, a+t; \\ a+b, a+c, a+d \end{matrix}; 1\right),
 \end{aligned}$$

where  $(a)_k = a(a+1) \cdots (a+k-1)$  if  $k \geq 1$  and  $(a)_0 = 1$ . (We use the same  ${}_4F_3$  notation for  $p_n$  even if one of  $a+b$ ,  $a+c$ , or  $a+d$  is a negative integer  $-N$  with  $N \geq n$ .) This is a polynomial of degree  $n$  in  $t^2$  since  $(a-t)_k(a+t)_k = \prod_{j=0}^{k-1} ((a+j)^2 - t^2)$ . The  ${}_4F_3$  is balanced, meaning that it is a finite series and the sum of the denominator parameters equals the sum of the numerator parameters plus one. Such series satisfy a transformation formula (Bailey [1, p. 56]):

$$(1.2) \quad {}_4F_3\left(\begin{matrix} -n, b, c, d; 1 \\ e, f, g \end{matrix}\right) = \frac{(f-b)_n(g-b)_n}{(f)_n(g)_n} {}_4F_3\left(\begin{matrix} -n, b, e-c, e-d; 1 \\ e, b-f-n+1, b-g-n+1 \end{matrix}\right),$$

provided  $e+f+g = -n+b+c+d+1$ . (This formula, when iterated, contains the symmetries of the 6-*j* symbols.) In terms of the  ${}_4F_3$  polynomials, (1.2) says that

$$(1.3) \quad p_n(x; a, b, c, d) = p_n(x; b, a, c, d),$$

so that  $p_n$  is symmetric in all four parameters.

There are various orthogonality relations for  $\{p_n\}$  (with respect to positive measures on the real line) corresponding to various conditions on  $a$ ,  $b$ ,  $c$  and  $d$ . These relations are derived in § 3 from the complex orthogonality in § 2.

**2. Complex orthogonality.** We prove the complex orthogonality relation

$$(2.1) \quad \frac{1}{2\pi i} \int_C f(z) p_m(z^2) p_n(z^2) dz = \delta_{mn} M h_n$$

\* Received by the editors August 1, 1979, and in revised form October 15, 1979.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.



with

$$f(z) = \frac{\Gamma(a+z)\Gamma(a-z)\Gamma(b+z)\Gamma(b-z)\Gamma(c+z)\Gamma(c-z)\Gamma(d+z)\Gamma(d-z)}{\Gamma(2z)\Gamma(-2z)},$$

$$M = \frac{2\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)}$$

and

$$h_n = \frac{n!(a+b+c+d+n-1)_n(a+b)_n(a+c)_n(a+d)_n(b+c)_n(b+d)_n(c+d)_n}{(a+b+c+d)_{2n}}.$$

Here  $a, b, c,$  and  $d$  are complex. The contour  $C$  is the imaginary axis deformed so as to separate the increasing sequences of poles of  $f(z)$   $\cdot (\{a+k\}_{k=0}^\infty, \{b+k\}_{k=0}^\infty, \{c+k\}_{k=0}^\infty, \{d+k\}_{k=0}^\infty)$  from the decreasing sequences  $(\{-a-k\}_{k=0}^\infty, \dots, \{-d-k\}_{k=0}^\infty)$ . Of course, we need to assume that these two sets of poles are disjoint, i.e.,

$$(2.2) \quad 2a, a+b, a+c, \dots, c+d, 2d \notin \{0, -1, -2, \dots\}.$$

The case  $m = n = 0,$

$$(2.3) \quad \frac{1}{2\pi i} \int_C \frac{\Gamma(a+z)\Gamma(a-z)\Gamma(b+z)\Gamma(b-z)\Gamma(c+z)\Gamma(c-z)\Gamma(d+z)\Gamma(d-z)}{\Gamma(2z)\Gamma(-2z)} dz = \frac{2 \cdot \Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c)\Gamma(b+d)\Gamma(c+d)}{\Gamma(a+b+c+d)}$$

is an integral analog of the  ${}_5F_4$  summation theorem (Bailey [1, p. 27]) which may be written

$$(2.4) \quad {}_5F_4 \left( \begin{matrix} 2a, a+1, a+b, a+c, a+d; \\ a, a-b+1, a-c+1, a-d+1 \end{matrix} \quad 1 \right) = \frac{\Gamma(a-b+1)\Gamma(a-c+1)\Gamma(a-d+1)\Gamma(-a-b-c-d+1)}{\Gamma(2a+1)\Gamma(-b-c+1)\Gamma(-b-d+1)\Gamma(-c-d+1)}$$

or

$$\sum_{z=a, a+1, \dots} \frac{\Gamma(1-2z)\Gamma(1+2z)}{\Gamma(1-a-z)\Gamma(1-a+z) \cdots \Gamma(1-d-z)\Gamma(1-d+z)} = \frac{\Gamma(1-a-b-c-d)}{\Gamma(1-a-b)\Gamma(1-a-c)\Gamma(1-a-d)\Gamma(1-b-c)\Gamma(1-b-d)\Gamma(1-c-d)}$$

provided  $\text{Re}(a+b+c+d) < 1$  for convergence. (Extending the sum over  $z = \pm a, \pm(a+1), \dots$  improves the resemblance to the integral formula.) Similar Mellin-Barnes integrals are found in Bailey [1], and in fact it is possible to derive (2.3) from his formula (1) on p. 47.

To prove (2.3), we need an asymptotic estimate for the integrand  $f(z)$ . We use the reflection formula  $\Gamma(z) = \pi/\Gamma(1-z) \sin \pi z$  along with Stirling's formula

$$\Gamma(a+z) = \sqrt{2\pi} \cdot z^{a+z-1/2} e^{-z} \left( 1 + O\left(\frac{1}{z}\right) \right)$$

as  $z \rightarrow \infty$  in  $S_\theta = \{z: |\arg z| < \theta\}, 0 < \theta < \pi,$  and the estimates

$$\sin 2\pi z = O(e^{2\pi|\Im mz|})$$

in the entire plane, and

$$\sin \pi(a - z) = O(e^{-\pi|\Im mz|})$$

in the plane excluding  $\varepsilon$ -neighborhoods of the poles. The implied constants in these formulas may be chosen independently of  $a$  if  $a$  takes values in a bounded set. These give

$$\begin{aligned} f(z) &= \frac{-2\pi^3 z \sin 2\pi z \Gamma(a + z) \cdots \Gamma(d + z)}{\sin \pi(a - z) \cdots \sin \pi(d - z) \Gamma(1 - a + z) \cdots \Gamma(1 - d + z)} \\ (2.5) \quad &= z^{2(a+b+c+d)-3} O(e^{-2\pi|\Im mz|}) \\ &= O(|z|^{2\operatorname{Re}(a+b+c+d)-3} e^{-2\pi|\Im mz|}) \end{aligned}$$

as  $z \rightarrow \infty$  in  $S_\theta$  excluding  $\varepsilon$ -neighborhoods of the poles. Since  $f(z)$  is an even function, (2.5) holds as  $z \rightarrow \infty$  in the plane excluding  $\varepsilon$ -neighborhoods of the poles.

In particular, (2.5) holds as  $z \rightarrow \infty$  on  $C$ , so  $\int_C f(z) dz$  is convergent. Furthermore, since the implied constant in (2.5) is independent of  $a$  (in bounded sets), the integral defines an analytic function of  $a$  in  $\{a: a, a + 1, a + 2, \dots \text{ are to the right of } C \text{ and } -a, -a - 1, \dots \text{ are to the left of } C\}$ . We will use Cauchy’s theorem to prove (2.3) under the condition  $\operatorname{Re}(a + b + c + d) < 1$ . This condition may then be removed by analytic continuation.

Consider  $\int_{C_1+C_2} f(z) dz$ , where  $C_1 = C_1(\omega)$  is the piece of  $C$  from  $-i\omega$  to  $+i\omega$ , and  $C_2(\omega)$  is the path consisting of the three line segments from  $i\omega$  to  $\omega + i\omega$  to  $\omega - i\omega$  to  $-i\omega$ . We will let  $\omega \rightarrow \infty$  through value  $\omega_0, \omega_0 + 1, \omega_0 + 2, \dots$ , where  $\omega_0$  is chosen so that the contours  $C_2(\omega_0 + k)$  avoid the poles of  $f(z)$ . Then

$$\begin{aligned} \left| \int_{C_2} f(z) dz \right| &\leq 4\omega \max \{|f(z)|: z \text{ on } C_2\} \\ &= O(\omega^{2\operatorname{Re}(a+b+c+d)-2}). \end{aligned}$$

So with  $\operatorname{Re}(a + b + c + d) < 1$ ,  $\int_{C_2} f(z) dz$  vanishes as  $\omega \rightarrow \infty$ .

It follows that

$$\frac{1}{2\pi i} \int_C f(z) dz = \lim_{\omega \rightarrow \infty} \frac{1}{2\pi i} \int_{C_1+C_2} f(z) dz,$$

which is minus the sum of the residues at the poles to the right of  $C$ . The residue at  $z = a + k$  is

$$\begin{aligned} &\frac{\Gamma(2a + k)((-1)^{k+1}/k!) \Gamma(a + b + k) \Gamma(b - a - k)}{\Gamma(c + a + k) \Gamma(c - a - k) \Gamma(d + a + k) \Gamma(d - a - k)} \\ &\quad \frac{\Gamma(2a + 2k) \Gamma(-2a - 2k)}{\Gamma(-2a)} \\ &= \frac{-\Gamma(a + b) \Gamma(a + c) \Gamma(a + d) \Gamma(b - a) \Gamma(c - a) \Gamma(d - a)}{\Gamma(-2a)} \\ &\quad \cdot \frac{(2a)_k (a + 1)_k (a + b)_k (a + c)_k (a + d)_k}{(1)_k (a)_k (a - b + 1)_k (a - c + 1)_k (a - d + 1)_k}. \end{aligned}$$

Here we are assuming that the poles are simple, but in (2.3) this condition is removable.

Minus the sum of these residues for  $k = 0, 1, 2, \dots$  is given by the  ${}_5F_4$  formula (2.4) as

$$\frac{\Gamma(a+b)\Gamma(a+c)\Gamma(a+d)\Gamma(b-a)\Gamma(c-a)\Gamma(d-a)}{\Gamma(-2a)} \cdot \frac{\Gamma(a-b+1)\Gamma(a-c+1)\Gamma(a-d+1)\Gamma(-a-b-c-d+1)}{\Gamma(2a+1)\Gamma(-b-c+1)\Gamma(-b-d+1)\Gamma(-c-d+1)} = \frac{1}{2}R \cdot S_a,$$

where  $R$  is the right-hand side of (2.3), and

$$S_a = \frac{\sin(-2\pi a) \sin \pi(b+c) \sin \pi(b+d) \sin \pi(c+d)}{\sin \pi(b-a) \sin \pi(c-a) \sin \pi(d-a) \sin \pi(a+b+c+d)}.$$

If we define  $S_b, S_c, S_d$  symmetrically, then adding the contributions from all the poles  $a+k, b+k, c+k, d+k$  for  $k = 0, 1, 2, \dots$  gives

$$\frac{1}{2\pi i} \int_C f(z) dx = \frac{1}{2}R \cdot (S_a + S_b + S_c + S_d).$$

Finally, a tedious trigonometric computation or a contour integration argument shows  $S_a + S_b + S_c + S_d = 2$ . This proves (2.3).

To prove the orthogonality (2.1), first note that, by the symmetry (1.3),

$$p_m(z^2) = (-1)^m (a+b+c+d+m-1)_m (b-z)_m (b+z)_m + \sum_{j=0}^{m-1} \gamma_j (b-z)_j (b+z)_j.$$

For  $j = 0, 1, \dots, n$ ,

$$\begin{aligned} \frac{1}{2\pi i} \int_C f(z) p_n(z^2) (b-z)_j (b+z)_j dz &= (a+b)_n (a+c)_n (a+d)_n \\ &\cdot \sum_{k=0}^n \frac{(-n)_k (n+a+b+c+d-1)_k}{(a+b)_k (a+c)_k (a+d)_k (1)_k} \\ &\cdot \frac{1}{2\pi i} \int_C f(z) (a-z)_k (a+z)_k (b-z)_j (b+z)_j dz. \end{aligned}$$

The integral here may be written

$$\frac{1}{2\pi i} \int_C \frac{\Gamma(a+k+z)\Gamma(a+k-z)\Gamma(b+j+z)\Gamma(b+j-z)\Gamma(c+z)\Gamma(c-z)\Gamma(d+z)\Gamma(d-z)}{\Gamma(2z)\Gamma(-2z)} dz$$

and, by (2.3), its value is

$$\begin{aligned} &\frac{2 \cdot \Gamma(a+b+k+j)\Gamma(a+c+k)\Gamma(a+d+k)\Gamma(b+c+j)\Gamma(b+d+j)\Gamma(c+d)}{\Gamma(a+b+c+d+k+j)} \\ &= \frac{2 \cdot \Gamma(a+b+j)\Gamma(a+c)\Gamma(a+d)\Gamma(b+c+j)}{\Gamma(b+d+j)\Gamma(c+d)(a+b+j)_k (a+c)_k (a+d)_k} \cdot \frac{1}{\Gamma(a+b+c+d+j)(a+b+c+d+j)_k}. \end{aligned}$$

Therefore,

$$\begin{aligned}
 & \frac{1}{2\pi i} \int_C f(z) p_n(z^2) (b-z)_j (b+z)_j dz \\
 &= 2 \cdot \frac{\Gamma(a+b+j)\Gamma(a+c+n)\Gamma(a+d+n)\Gamma(b+c+j)\Gamma(b+d+j)\Gamma(c+d)(a+b)_n}{\Gamma(a+b+c+d+j)} \\
 (2.6) \quad & \cdot {}_3F_2\left(\begin{matrix} -n, n+a+b+c+d-1, a+b+j; \\ a+b, a+b+c+d+j \end{matrix}; 1\right) \\
 &= 2 \cdot \frac{\Gamma(a+b+j)\Gamma(a+c+n)\Gamma(a+d+n)\Gamma(b+c+j)\Gamma(b+d+j)\Gamma(c+d+n)(-j)_n}{\Gamma(a+b+c+d+j+n)}.
 \end{aligned}$$

We have evaluated the  ${}_3F_2$  by the formula of Pfaff and Saalschütz (Bailey [1, p. 9]):

$${}_3F_2\left(\begin{matrix} -n, b, c; \\ d, e \end{matrix}; 1\right) = \frac{(d-c)_n (e-c)_n}{(d)_n (e)_n}$$

provided  $d + e = -n + b + c + 1$ . The result of (2.6) is zero if  $j < n$ . Therefore

$$\frac{1}{2\pi i} \int_C f(z) p_m(z^2) p_n(z^2) dz = 0 \quad \text{if } m < n,$$

while

$$\begin{aligned}
 & \frac{1}{2\pi i} \int_C f(z) p_n(z^2)^2 dz \\
 &= (-1)^n (a+b+c+d+n-1)_n \frac{1}{2\pi i} \int_C f(z) p_n(z^2) (b-z)_n (b+z)_n dz \\
 &= \frac{2 \cdot n! (a+b+c+d+n-1)_n \Gamma(a+b+n) \Gamma(a+c+n) \cdots \Gamma(c+d+n)}{\Gamma(a+b+c+d+2n)}
 \end{aligned}$$

as required.

**3. Real orthogonalities.** We wish to obtain from (2.1) orthogonality relations with respect to positive measures on the real line. Note that, when the real parts of  $a, b, c$ , and  $d$  are positive,  $C$  may be taken to be the imaginary axis. If, furthermore,  $a, b, c$ , and  $d$  are real except for conjugate pairs, then for imaginary  $z$ ,

$$f(z) = \left| \frac{\Gamma(a+z)\Gamma(b+z)\Gamma(c+z)\Gamma(d+z)}{\Gamma(2z)} \right|^2.$$

Letting  $z = it$  in the integral gives the orthogonality relation

$$\begin{aligned}
 (3.1) \quad & \frac{1}{2\pi} \int_0^\infty \left| \frac{\Gamma(a+it)\Gamma(b+it)\Gamma(c+it)\Gamma(d+it)}{\Gamma(2it)} \right|^2 p_n(-t^2) p_m(-t^2) dt \\
 &= \delta_{m,n} n! (n+a+b+c+d-1)_n \frac{\Gamma(a+b+n)\Gamma(a+c+n) \cdots \Gamma(c+d+n)}{\Gamma(a+b+c+d+2n)}
 \end{aligned}$$

( $a, b, c$ , and  $d$  positive real except for complex conjugate pairs with positive real parts). With these conditions on the parameters, the polynomials  $p_n(t^2)$  are real for real values of  $t^2$ . This is clear from the definition (1.1) in the case where  $a$  and  $b$  are real (and either  $c$  and  $d$  are real or  $c = \bar{d}$ ). If  $a = \bar{b}$  and  $c = \bar{d}$ , then it is clear that  $p_n(t^2; b, a, c, d) = p_n(t^2; a, b, c, d)$ . But then the symmetry (1.3) shows that  $p_n(t^2; a, b, c, d)$  is real.

Return now to (2.1) and consider the case where  $a < 0$  and  $a + b, a + c, a + d$  are positive except possibly for one pair of complex conjugates with positive real parts. These conditions will yield an orthogonality (3.3) with respect to a positive weight function consisting of a continuous function plus some point masses. By Cauchy's theorem

$$\begin{aligned}
 & \frac{1}{2\pi i} \int_C f(z)p_n(z^2)p_m(z^2) dz \\
 &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} f(z)p_n(z^2)p_m(z^2) dz \\
 & \quad - (\text{sum of residues of integrand at } z = a + k, \text{ with } a \leq a + k < 0) \\
 (3.2) \quad & \quad + (\text{sum of residues at } z = -(a + k), \text{ with } a \leq a + k < 0) \\
 &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} f(z)p_n(z^2)p_m(z^2) dz - 2 (\text{sum of residues at } z = a + k, a \leq a + k < 0) \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(it)p_n(-t^2)p_m(-t^2) dt \\
 & \quad + 2 \frac{\Gamma(a + b)\Gamma(a + c)\Gamma(a + d)\Gamma(b - a)\Gamma(c - a)\Gamma(d - a)}{\Gamma(-2a)} \\
 & \quad \cdot \sum_{\substack{k=0,1,\dots \\ a+k < 0}} \frac{(2a)_k(a + 1)_k(a + b)_k(a + c)_k(a + d)_k}{(1)_k(a)_k(a - b + 1)_k(a - c + 1)_k(a - d + 1)_k} p_n((a + k)^2)p_m((a + k)^2).
 \end{aligned}$$

Therefore, (2.1) becomes

$$\begin{aligned}
 & \frac{1}{2\pi} \int_0^\infty \left| \frac{\Gamma(a + it) \cdots \Gamma(d + it)}{\Gamma(2it)} \right|^2 p_n(-t^2)p_m(-t^2) dt \\
 & \quad + \frac{\Gamma(a + b)\Gamma(a + c)\Gamma(a + d)\Gamma(b - a)\Gamma(c - a)\Gamma(d - a)}{\Gamma(-2a)} \\
 (3.3) \quad & \sum_{\substack{k=0,1,\dots \\ a+k < 0}} \frac{(2a)_k(a + 1)_k(a + b)_k(a + c)_k(a + d)_k}{(1)_k(a)_k(a - b + 1)_k(a - c + 1)_k(a - d + 1)_k} p_n((a + k)^2)p_m((a + k)^2) \\
 &= \delta_{m,n} n!(n + a + b + c + d - 1)_n \frac{\Gamma(a + b + n) \cdots \Gamma(c + d + n)}{\Gamma(a + b + c + d + 2n)}
 \end{aligned}$$

if  $a < 0$ , and  $a + b, a + c, a + d > 0$  except possibly for a pair of conjugates with positive real parts. Condition (2.2) requires that  $2a \notin \{0, -1, -2, \dots\}$ , but here this condition is removable.

Formula (2.1) also yields some purely discrete orthogonality relations for  $p_n(t^2)$ . Take  $a + b = -N + \epsilon$ ,  $N$  a positive integer. (Condition (2.2) requires  $\epsilon \neq 0$ .) Use Cauchy's theorem as in (3.2) to replace the contour  $C$  by the imaginary axis and add some residues. Then divide the equation by  $\Gamma(a + b) = \Gamma(-N + \epsilon)$ . As  $\epsilon \rightarrow 0$ , the integral

term vanishes because  $1/\Gamma(-N + \epsilon) \rightarrow 0$ , and the result may be written

$$(3.4) \quad \frac{(a-c+1)_N(a-d+1)_N}{(2a+1)_N(1-c-d)_N} \sum_{k=0}^N \frac{(2a)_k(a+1)_k(a+b)_k(a+c)_k(a+d)_k}{(1)_k(a)_k(a-b+1)_k(a-c+1)_k(a-d+1)_k} \cdot p_n((a+k)^2)p_m((a+k)^2) \\ = \delta_{m,n} \frac{n!(n+a+b+c+d-1)_n(a+b)_n(a+c)_n(a+d)_n(b+c)_n(b+d)_n(c+d)_n}{(a+b+c+d)_{2n}}$$

when  $a + b = -N$ . Interchanging  $a$  and  $b$  here is equivalent to summing in the reverse order. The case  $m = n = 0$  is the terminating series version of (2.4):

$${}_5F_4 \left( \begin{matrix} 2a, a+1, a+b, a+c, a+d; \\ a, a-b+1, a-c+1, a-d+1 \end{matrix} \middle| 1 \right) = \frac{(2a+1)_N(1-c-d)_N}{(a-c+1)_N(a-d+1)_N}$$

when  $a + b = -N$ . Formula (3.4) can also be proven directly from this  ${}_5F_4$  theorem just as the complex orthogonality (2.1) was proven from the integral formula (2.3).

Necessary and sufficient conditions on  $a, b, c, d$  for the positivity of the weights in (3.4) are quite messy, but some sufficient conditions are

$$(3.5) \quad a+b = -N, \quad b < -\frac{1}{2} < a, \quad -a < c < a+1 \quad \text{and} \quad \text{either } d > -b \text{ or } d < b+1.$$

Of course, interchanging  $a$  and  $b$  in (3.5) also gives sufficient conditions.

**4. Limiting cases.** We now describe how, as claimed in § 1, many orthogonality relations for previously known polynomials are included in the  ${}_4F_3$  orthogonalities as limiting cases. The appropriate limit processes can usually be determined by comparing the hypergeometric series representations of the polynomials. It sometimes helps to write the  ${}_4F_3$  polynomials, with a change of variable and parameters, as

$$(4.1) \quad r_n(\lambda(x); \alpha, \beta, \gamma, \delta) = {}_4F_3 \left( \begin{matrix} -n, n+\alpha+\beta+1, -x, x+\gamma+\delta+1; \\ \alpha+1, \beta+\delta+1, \gamma+1 \end{matrix} \middle| 1 \right)$$

with  $\lambda(x) = x(x + \gamma + \delta + 1)$ .

Then (3.4) becomes

$$(4.2) \quad \sum_{k=0}^N \frac{(\gamma+\delta+1)_k((\gamma+\delta+3)/2)_k(\alpha+1)_k(\beta+\delta+1)_k(\gamma+1)_k}{(1)_k((\gamma+\delta+1)/2)_k(\gamma+\delta-\alpha+1)_k(\gamma-\beta+1)_k(\delta+1)_k} \cdot r_n(\lambda(k))r_m(\lambda(k)) \\ = \delta_{m,n} M \cdot \frac{n!(n+\alpha+\beta+1)_n(\beta+1)_n(\alpha-\delta+1)_n(\alpha+\beta-\gamma+1)_n}{(\alpha+\beta+2)_{2n}(\alpha+1)_n(\beta+\delta+1)_n(\gamma+1)_n}$$

if  $\alpha + 1, \beta + \delta + 1$ , or  $\gamma + 1 = -N$ , with

$$M = \begin{cases} \frac{(\gamma+\delta+2)_N(-\beta)_N}{(\gamma-\beta+1)_N(\delta+1)_N} & \text{if } \alpha + 1 = -N, \\ \frac{(\gamma+\delta+2)_N(\delta-\alpha)_N}{(\gamma+\delta-\alpha+1)_N(\delta+1)_N} & \text{if } \beta + \delta + 1 = -N, \\ \frac{(-\delta)_N(\alpha+\beta+2)_N}{(\alpha-\delta+1)_N(\beta+1)_N} & \text{if } \gamma + 1 = -N. \end{cases}$$

Letting  $\delta \rightarrow \infty$  with  $\gamma + 1 = -N$  gives the Hahn polynomial orthogonality:

$$(4.3) \quad \sum_{x=0}^N \frac{(\alpha + 1)_x (-N)_x}{(-N - \beta)_x x!} Q_n(x; \alpha, \beta, N) Q_m(x; \alpha, \beta, N) = 0, \quad m \neq n,$$

$$Q_n(x; \alpha, \beta, N) = {}_3F_2 \left( \begin{matrix} -n, n + \alpha + \beta + 1, -x; \\ \alpha + 1, -N \end{matrix} \quad 1 \right), \quad 0 \leq n \leq N.$$

Letting  $\beta \rightarrow \infty$  in (4.2) with  $\alpha + 1 = -N$  gives the dual Hahn orthogonality (Karlin and McGregor [3])

$$\sum_{x=0}^N \frac{(\gamma + \delta + 1)_x ((\gamma + \delta + 3)/2)_x (\gamma + 1)_x (-N)_x (-1)^x}{x! ((\gamma + \delta + 1)/2)_x (\delta + 1)_x (N + \gamma + \delta + 2)_x} \cdot R_n(\lambda(x); \gamma, \delta, N) R_m(\lambda(x); \gamma, \delta, N) = 0, \quad m \neq n,$$

$$R_n(\lambda(x); \gamma, \delta, N) = {}_3F_2 \left( \begin{matrix} -n, -x, x + \gamma + \delta + 1; \\ -N, \gamma + 1 \end{matrix} \quad 1 \right), \quad 0 \leq n \leq N.$$

Actually, the dual Hahn polynomials have continuous, discrete, and mixed orthogonality relations (with positive weight functions) which come from the orthogonalities for the  ${}_4F_3$ 's as  $d \rightarrow \infty$ . For example, (3.1) becomes

$$(4.4) \quad \frac{1}{2\pi} \int_0^\infty \left| \frac{\Gamma(a + it)\Gamma(b + it)\Gamma(c + it)}{\Gamma(2it)} \right|^2 p_n(-t^2) p_m(-t^2) dt$$

$$= \delta_{m,n} n! \Gamma(a + b + n) \Gamma(a + c + n) \Gamma(b + c + n),$$

where

$$p_n(z^2) = (a + b)_n (a + c)_n {}_3F_2 \left( \begin{matrix} -n, a - z, a + z; \\ a + b, a + c \end{matrix} \quad 1 \right)$$

and  $a, b,$  and  $c$  are all positive except possibly for a pair of complex conjugates with positive real parts. The complex orthogonality (2.1) becomes

$$\frac{1}{2\pi i} \int_C \frac{\Gamma(a + z)\Gamma(a - z)\Gamma(b + z)\Gamma(b - z)\Gamma(c + z)\Gamma(c - z)}{\Gamma(2z)\Gamma(-2z)} p_n(z^2) p_m(z^2) dz$$

$$= \delta_{m,n} 2 \cdot n! \Gamma(a + b + n) \Gamma(a + c + n) \Gamma(b + c + n)$$

with  $p_n(z^2)$  as above, and  $C$  separating the increasing and decreasing sequences of poles.

It is known that, by taking limits of (4.3), we can obtain the discrete orthogonalities for the Meixner, Krawtchouk and Charlier polynomials, as well as the orthogonalities for the classical polynomials of Jacobi, Laguerre, and Hermite.

It is also interesting that the classical polynomial orthogonalities can be realized in a different way as limits of the continuous orthogonality relations (3.1) and (4.4). In (3.1), let  $a = b = (\alpha + 1)/2$  and  $c = \bar{d} = (\beta + 1)/2 + i\omega$ , and change variable  $t = \omega s$  to get:

$$\int_0^\infty \frac{1}{2\pi} \left| \frac{\Gamma((\alpha + 1)/2 + i\omega s)^2 \Gamma((\beta + 1)/2 + i\omega(s + 1)) \Gamma((\beta + 1)/2 + i\omega(s - 1))}{\Gamma(2i\omega s) \Gamma((\alpha + \beta + 2)/2 + i\omega)^2} \right|^2$$

$$\cdot {}_4F_3 \left( \begin{matrix} -n, n + \alpha + \beta + 1, (\alpha + 1)/2 + i\omega s, (\alpha + 1)/2 - i\omega s; \\ \alpha + 1, (\alpha + \beta + 2)/2 + i\omega, (\alpha + \beta + 2)/2 - i\omega \end{matrix} \quad 1 \right) {}_4F_3 \left( \begin{matrix} -m, \dots \\ \dots \end{matrix} \right) \omega ds$$

$$= \delta_{mn} \frac{n!(n + \alpha + \beta + 1)_n \Gamma(\alpha + 1) \Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 2n)} \frac{(\beta + 1)_n}{(\alpha + 1)_n}.$$

As  $\omega \rightarrow +\infty$ , the weight function is asymptotic to  $2 \cdot s^{2\alpha+1} |1-s^2|^\beta e^{-\pi\omega(|s-1|+s-1)}$  by Stirling's formula, and therefore has limit  $2s^{2\alpha+1}(1-s^2)^\beta$  if  $0 < s < 1$ , 0 if  $s > 1$ . At least formally, this gives

$$\int_0^1 s^{2\alpha}(1-s^2)^\beta {}_2F_1\left(\begin{matrix} -n, n+\alpha+\beta+1 \\ \alpha+1 \end{matrix}; s^2\right) {}_2F_1\left(\begin{matrix} -m, m+\alpha+\beta+1 \\ \alpha+1 \end{matrix}; s^2\right) d(s^2) \\ = \delta_{mn} \frac{n!(n+\alpha+\beta+1)_n \Gamma(\alpha+1)\Gamma(\beta+1)(\beta+1)_n}{\Gamma(\alpha+\beta+2n)(\alpha+1)_n},$$

which is the Jacobi orthogonality with a change of variable.

In (4.4), let  $a = b = (\alpha + 1)/2$  and  $c = \omega^2$ , and change variable  $t = \omega s$ . As  $\omega \rightarrow +\infty$ , a messy application of Stirling's formula gives

$$(4.5) \quad \int_0^\infty e^{-s^2}(s^2)^\alpha {}_1F_1\left(\begin{matrix} -n \\ \alpha+1 \end{matrix}; s^2\right) {}_1F_1\left(\begin{matrix} -m \\ \alpha+1 \end{matrix}; s^2\right) d(s^2) \\ = \delta_{mn} \frac{\Gamma(\alpha+1)n!}{(\alpha+1)_n},$$

which is the Laguerre polynomial orthogonality. The cases of formula (4.5) with  $\alpha = -\frac{1}{2}$  and  $\alpha = \frac{1}{2}$  are the orthogonalities for the Hermite polynomials of even and odd degrees, respectively. Of course, the Laguerre and Hermite orthogonalities are also limiting cases of the Jacobi polynomial orthogonality.

Another limiting case of (3.1) is Pollaczek's orthogonality relation [4]:

$$\int_{-\infty}^\infty e^{(-\pi+2\phi)x} |\Gamma(\lambda+ix)|^2 P_n^{(\lambda)}(x; \phi) P_m^{(\lambda)}(x; \phi) dx = 0, \quad m \neq n, \\ P_n^{(\lambda)}(x; \phi) = \frac{(2\lambda)_n e^{in\phi}}{n!} {}_2F_1\left(\begin{matrix} -n, \lambda+ix \\ 2\lambda \end{matrix}; 1-e^{-2i\phi}\right), \quad n \geq 0,$$

$\lambda > 0, 0 < \phi < \pi$ . We extend (3.1) to a symmetric integral on  $(-\infty, \infty)$ . Then we take  $a = \lambda + i\omega, b = \lambda - i\omega$ , and  $c = d = \omega \cot(\phi/2)$ , and substitute  $t = x - \omega$ :

$$\frac{1}{2\pi} \int_{-\infty}^\infty \left| \frac{\Gamma(\lambda+ix)\Gamma(\lambda+ix-2i\omega)\Gamma(\omega \cot(\phi/2)+i(x-\omega))^2}{\Gamma(2i(x-\omega))} \right|^2 \\ \cdot {}_4F_3\left(\begin{matrix} -n, n+2\lambda+2\omega \cot(\phi/2)-1, \lambda+ix, \lambda-ix+2i\omega \\ 2\lambda, \lambda+\omega(\cot(\phi/2)+i), \lambda+\omega(\cot(\phi/2)+i) \end{matrix}; 1\right) {}_4F_3\left(\begin{matrix} -m, \dots \\ \dots \end{matrix}\right) dx \\ n!(n+2\lambda+2\omega \cot(\phi/2)-1)_n \Gamma(2\lambda+n) \\ = 2\delta_{mn} \frac{\Gamma(2\omega \cot(\phi/2)+n) |\Gamma(\lambda+\omega(\cot(\phi/2)+i)+n)|^2}{\Gamma(2\lambda+2\omega \cot(\phi/2)+2n)}.$$

As  $\omega \rightarrow +\infty$ , we get Pollaczek's orthogonality by another application of Stirling's formula.

**5. The 6-j symbols.** The 6-j symbols  $\bar{W}\left(\begin{matrix} a, b, c \\ d, e, f \end{matrix}\right)$ , important in quantum mechanics in the coupling of angular momenta, satisfy an orthogonality relation which we will show to be equivalent to certain cases of (3.4). It appears that this orthogonality was recognized as a polynomial orthogonality only in very special cases (Biedenharn et al. [2, p. 253]).  $\bar{W}\left(\begin{matrix} a, b, c \\ d, e, f \end{matrix}\right)$  is defined for half-integers  $a, b, c, d, e, f$  which are nonnegative and



satisfy certain triangle conditions:  $a + b + c$  is an integer,  $a \leq b + c$ ,  $b \leq a + c$ , and  $c \leq a + b$ , so that  $a$ ,  $b$ , and  $c$  are the sides of some triangle; each of the triples  $(a, e, f)$ ,  $(d, b, f)$ , and  $(d, e, c)$  satisfies similar conditions. An explicit formula for the 6- $j$  symbol is

$$(5.1) \quad \bar{W} \begin{pmatrix} a, b, c \\ d, e, f \end{pmatrix} = \Delta(a, b, c) \Delta(a, e, f) \Delta(d, b, f) \Delta(d, e, c) \cdot \sum_k \frac{(-1)^k (k+1)!}{(k-a-b-c)! (k-a-e-f)! (k-d-b-f)! (k-d-e-c)! \cdot (a+b+d+e-k)! (a+c+d+f-k)! (b+c+e+f-k)!}$$

where

$$\Delta(a, b, c) = \left[ \frac{(a+b-c)! (a-b+c)! (-a+b+c)!}{(a+b+c+1)!} \right]^{1/2}$$

and the sum is over all integers  $k \geq 0$ . Only finitely many terms of the sum are nonzero. There are symmetries here which are easier to understand if we consider the tetrahedron with edges  $a, b, c, d, e$  and  $f$  and faces  $(a, b, c)$ ,  $(a, e, f)$ ,  $(d, b, f)$ , and  $(d, e, c)$ . The value of  $\bar{W} \begin{pmatrix} a, b, c \\ d, e, f \end{pmatrix}$  is preserved under any permutation of the parameters which preserves the tetrahedron.

Racah's orthogonality [5] is

$$(5.2) \quad \sum_c (2c+1) \bar{W} \begin{pmatrix} a, b, c \\ d, e, f \end{pmatrix} \bar{W} \begin{pmatrix} a, b, c \\ d, e, f' \end{pmatrix} = \frac{\delta_{ff'}}{2f+1}.$$

The permissible values for  $f$  and  $f'$  and the values of the summation variable are determined by the triangle conditions. The inequalities involved are

$$(5.3) \quad |a-e|, |b-d| \leq f, f' \leq a+e, b+d$$

and

$$(5.4) \quad |a-b|, |d-e| \leq c \leq a+b, d+e.$$

By the tetrahedron symmetries, there is no loss of generality in assuming that

$$(5.5) \quad \max(|a-b|, |d-e|) = |a-b| = a-b.$$

We need to consider two cases, distinguished by the upper limit on  $c$ .

Consider first the case where

$$(5.6) \quad d+e \leq a+b.$$

With conditions (5.5) and (5.6) we must have  $\max(|a-e|, |b-d|) = a-e$  and  $b+d \leq a+e$ , so (5.3) and (5.4) reduce to  $a-e \leq f, f' \leq b+d$ , and  $a-b \leq c \leq d+e$ . Let  $N = b+d+e-a$  and replace the variable  $c$  by  $d+e-x$ , so the orthogonality is on the  $N+1$  points where  $x = 0, 1, \dots, N$ . Replace  $f$  by  $b+d-n$ , so the orthogonal functions are indexed by  $n, 0 \leq n \leq N$ . We claim that

$$(5.7) \quad \bar{W} \begin{pmatrix} a, b, d+e-x \\ d, e, b+d-n \end{pmatrix} = C \frac{\Delta(a, b, d+e-x) \Delta(d, e, d+e-x)}{x! (N-x)! (2d-x)! (a+b-d-e+x)!} \cdot {}_4F_3 \left( \begin{matrix} -n, n-2b-2d-1, -x, x-2d-2e-1; \\ -a-b-d-e-1, -2d, -N \end{matrix} \right),$$

where the factor  $C$  is independent of  $x$ . In fact, this is just the  ${}_4F_3$  transformation (1.2), but since a little care is needed to avoid zero denominators in the computation, we give details. By applying (5.1) to  $\bar{W}\left(\begin{matrix} a, b, d+e-x \\ d, e, b+d-n \end{matrix}\right)$  and substituting  $a+b+d+e-j$  for  $k$ , we get

$$\begin{aligned} & \bar{W}\left(\begin{matrix} a, b, d+e-x \\ d, e, b+d-n \end{matrix}\right) \\ &= C \cdot \Delta(a, b, d+e-x)\Delta(d, e, d+e-x) \\ & \cdot \sum_{j=0}^n \frac{(-1)^{a+b+d+e-j}(a+b+d+e-j+1)!}{(x-j)!(n-j)!(n+a-b-d+e-j)!(x+a+b-d-e-j)!} \\ & \cdot j!(2d-n-x+j)!(N-n-x+j)! \\ &= C \cdot \lim_{\varepsilon \rightarrow 0} \Delta(a, b, d+e-x)\Delta(d, e, d+e-x) \\ & \cdot \sum_{j=0}^n \frac{(-1)^{a+b+d+e-j}(a+b+d+e-j+1)!}{\Gamma(x+\varepsilon-j+1)(n-j)!(n+a-b-d+e-j)!} \\ & \cdot \frac{1}{\Gamma(x+\varepsilon+a+b-d-j+1)j!\Gamma(2d-n-x-\varepsilon+j+1)\Gamma(N-n-x-\varepsilon+j+1)} \\ &= C \cdot \lim_{\varepsilon \rightarrow 0} \frac{\Delta(a, b, d+e-x)\Delta(d, e, d+e-x)}{\Gamma(x+\varepsilon+1)\Gamma(x+\varepsilon+a+b-d-e+1)} \\ & \cdot \Gamma(2d-n-x-\varepsilon+1)\Gamma(N-n-x-\varepsilon+1) \\ & \cdot {}_4F_3\left(\begin{matrix} -n, -n-a+b+d-e, -x-\varepsilon, -x-\varepsilon-a-b+d+e; \\ -a-b-d-e-1, 2d-n-x-\varepsilon+1, N-n-x-\varepsilon+1 \end{matrix}; 1\right). \end{aligned}$$

By transformation (1.2), this is

$$\begin{aligned} & C \cdot \lim_{\varepsilon \rightarrow 0} \frac{\Delta(a, b, d+e-x)\Delta(d, e, d+e-x)}{\Gamma(x+\varepsilon+1)\Gamma(x+\varepsilon+a+b-d-e+1)\Gamma(2d-x-\varepsilon+1)\Gamma(n-x-\varepsilon+1)} \\ & \cdot {}_4F_3\left(\begin{matrix} -n, n-2b-2d-1, -x-\varepsilon, x+\varepsilon-2d-2e-1; \\ -a-b-d-e-1, -2d, -N \end{matrix}; 1\right) \\ &= \frac{C \cdot \Delta(a, b, d+e-x)\Delta(d, e, d+e-x)}{x!(x+a+b-d-e)!(2d-x)!(N-x)!} \\ & \cdot {}_4F_3\left(\begin{matrix} -n, n-2b-2d-1, -x, x-2d-2e-1; \\ -a-b-d-e-1, -2d, -N \end{matrix}; 1\right). \end{aligned}$$

This establishes (5.7). If  $a' = -d - e - \frac{1}{2}$ ,  $b' = a - b + \frac{1}{2}$ ,  $c' = e - d + \frac{1}{2}$ , and  $d' = -a - b - \frac{1}{2}$ , then the  ${}_4F_3$  we have is  $C \cdot p_n((a'+x)^2; a', b', c', d')$ . It is now an easy matter to compare the weight functions (or apply a general theorem) to see that (3.4) and (5.2) represent the same orthogonality.

The case where  $a+b \leq d+e$  is dealt with similarly. The limits on  $f$  and  $c$  become  $d-b \leq f, f' \leq b+d$  and  $a-b \leq c \leq a+b$ . Let  $N = 2b$ ; replace  $f$  by  $b+d-n$ ; and replace  $c$  by  $a+b-x$ . Then corresponding to (5.7) is

$$\begin{aligned} \bar{W}\left(\begin{matrix} a, b, a+b-x \\ d, e, b+d-n \end{matrix}\right) &= \frac{C \cdot \Delta(a, b, a+b-x)\Delta(d, e, a+b-x)}{x!(N-x)!(d+e-a-b+x)!(a+b+d-e-x)!} \\ & \cdot {}_4F_3\left(\begin{matrix} -n, n-2b-2d-1, -x, x-2a-2b-1; \\ -a-b-d-e-1, e-a-b-d, -2b \end{matrix}; 1\right). \end{aligned}$$

The  ${}_4F_3$  here is  $Cp_n((a'+x)^2; a', b', c', d')$  with  $a' = -a - b - \frac{1}{2}$ ,  $b' = a - b + \frac{1}{2}$ ,  $c' = e - d + \frac{1}{2}$ , and  $d' = -d - e - \frac{1}{2}$ .

In both cases, the weight functions satisfy the positivity conditions (3.5).

**Acknowledgment.** I thank the referee for pointing out that the reduction of the integral in (2.3) to the sum of four very well poised  ${}_5F_4$ 's is contained in L. J. Slater [6, (4.5.1.2)].

#### REFERENCES

- [1] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.
- [2] L. C. BIEDENHARN, J. M. BLATT AND M. E. ROSE, *Some properties of the Racah and associated coefficients*, Rev. Modern Phys., 24 (1952), pp. 249–257.
- [3] S. KARLIN AND J. L. MCGREGOR, *The Hahn polynomials, formulas and an application*, Scripta Math., 26 (1961), pp. 33–46.
- [4] F. POLLACZEK, *Sur une famille de polynômes orthogonaux qui contient les polynômes d'Hermite et de Laguerre comme cas limites*, C. R. Acad. Sci. Paris, 230 (1950), pp. 1563–1565.
- [5] G. RACAHA, *Theory of complex spectra II*, Phys. Rev., 62 (1942), pp. 438–462.
- [6] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.
- [7] J. A. WILSON, *Hypergeometric series recurrence relations and properties of some orthogonal functions*, to appear.

## ON INTEGRAL REPRESENTATIONS FOR LAMÉ AND OTHER SPECIAL FUNCTIONS\*

R. SHAIL†

**Abstract.** In this paper we develop the theory of a class of linear integral representations of the Lamé functions of the second kind, originally introduced by Arscott (1964). The kernels in this type of representation involve Legendre functions of the second kind. The complete class of 24 representations, three for each of the eight species of Lamé function of the second kind, is derived, and certain hitherto undetermined multipliers in the formulae are calculated. It is then shown how a knowledge of the values of the Lamé  $F$ -functions in certain basic regions enables one to compute their values throughout the complex plane.

The latter part of the paper is devoted to giving a simple and economic derivation of the class of Liouville nonlinear integral equations for the Lamé functions. Some generalizations and the applicability of the method to other classes of special functions are indicated.

**1. Introduction.** In the study of the properties of Lamé polynomials a number of authors [1], [2], [3] have considered certain homogeneous integral equations of the second kind satisfied by the polynomials, the kernels of the integral equations being Legendre polynomials. Typical of this type of formula is the relation

$$(1) \quad uE_{2n}^m(\alpha) = \lambda \int_{-2K}^{2K} P_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta,$$

where  $k$  is the modulus of the Jacobian elliptic functions, the characteristic value  $\lambda$  depends on  $n, m$ , and  $K(k)$  is the complete elliptic integral of the first kind. Arscott [3] has pointed out that there are 24 distinct kernels for integral equations similar to (1), leading to three representations for each of the eight types of Lamé polynomial, and has also noted that various alternative paths of integration are possible.

The derivation of (1) and its companion formulae depends on the general Theorem I of [3], which shows how to obtain from one solution of Lamé's equation further solutions in the form of integral transforms. The kernels of the transforms satisfy an equation which, by means of a coordinate transformation of the sphero-conal type, can be identified with the equation for spherical harmonics of some degree, and hence the Legendre polynomial in (1) is explained.

In the same paper [3], Arscott also gives four instances of analogous representations of the Lamé functions of the second kind; an example is

$$(2) \quad uF_{2n}^m(\alpha) = \lambda \int_K^{K+iK'} Q_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta.$$

In (2) the kernel is a Legendre function of the second kind,  $\lambda$  is an undetermined constant, and  $K' = K(k')$ , where  $k'$  is the modulus complementary to  $k$ . In order to prevent the argument of the  $Q$ -function becoming  $\pm 1$  on the linear path of integration, the restriction  $\operatorname{Re} \alpha \neq (2p + 1)K$ ,  $p$  an integer, is imposed on the complex variable  $\alpha$ . Arscott also conjectures that, as in the case of Lamé polynomials, there are three independent formulae for each of the eight types of Lamé function of the second kind, and expresses the opinion that such formulae may lead to practical methods of their computation. However, apart from some work by Sleeman [4] on the series-expansion of Lamé functions, little use seems to have been made of formulae such as (2) until the present author [5] found them to be ideal for the evaluation of certain limits arising in the calculation of stress-intensity factors in elastostatic elliptic crack problems.

\* Received by the editors May 22, 1979, and in revised form October 5, 1979.

† Department of Mathematics, University of Surrey, Guildford, Surrey, England.

One aim of this paper is to develop the theory of representations of the form (2) in a number of directions, but here it is appropriate to point out a fallacy in Arscott's work [3]. It is clear that if the  $Q_{2n}$ -function is that defined in (11), then the right-hand side of (2) is a periodic function of  $\alpha$ , with two independent periods  $4K$  and  $2iK'$ ; however the Lamé function of the second kind is *not* periodic, and hence (2) cannot be valid for all  $\alpha$ , subject to the sole restriction  $\text{Re } \alpha \neq (2p + 1)K$ . (A similar curious error occurs in Arscott's book [6, p. 94] in connection with Mathieu functions.) To overcome this difficulty, *basic regions* of the complex  $\alpha$ -plane are established in which the representation (2) and its companions are correct, and the full list of 24 formulae for the Lamé functions of the second kind, valid in these regions, is given. The hitherto unsolved problem of determining the  $\lambda$ -multipliers is treated. We then show that, in regions of the  $\alpha$ -plane congruent with the basic regions, the representation of the  $F$ -function is the sum of an integral of the form (2) together with a certain multiple of the corresponding Lamé polynomial. Thus, from a knowledge of the value of the  $F$ -function at a point in the basic region, its value at any congruent point in the  $\alpha$ -plane can be computed.

The second distinct topic considered in this paper concerns a class of nonlinear integral equations for the Lamé polynomials and functions of the second kind. Let  $(\alpha, \beta, \gamma)$  denote ellipsoidal coordinates and  $R$  the distance between the points  $(\alpha, \beta, \gamma)$  and  $(\alpha', \beta', \gamma')$ . As long ago as 1846, Liouville [7] discovered the formula

$$(3) \quad F_n^m(\gamma) = \frac{ik^{2\tau+3}l(-1)^{\sigma+\tau}(2n+1)}{8\pi E_n^m(\alpha)E_n^m(\beta)E_n^m(\gamma')} \int \int_{\mathcal{S}} \frac{1}{R} (\text{sn}^2 \alpha' - \text{sn}^2 \beta') E_n^m(\alpha') E_n^m(\beta') d\alpha' d\beta',$$

where  $l$  is the scale constant in the ellipsoidal coordinates, and

$$\int \int_{\mathcal{S}} \phi(\alpha', \beta') d\alpha' d\beta' = \int_{-2K}^{2K} \int_{K-2iK'}^{K+2iK'} \phi(\alpha', \beta') d\alpha' d\beta',$$

i.e., integration covering the ellipsoid  $\gamma = \gamma'$  twice. Further integral equations of this type have been treated by Arscott [3] and Sleeman [8], each of these authors using Liouville's original method. This method stems from a further double integral transform theorem (Theorem II of [3]), which is used to show that the Lamé function  $F_n^m$  is proportional to the double integral on the right-hand side of (3). The constant of proportionality is then determined by an application of Green's theorem.

In the later sections of this paper we show, that by considering certain trivial potential problems for an ellipsoid, the Liouville-type representations of the Lamé function  $F_n^m$  can be written down immediately, avoiding the rather circuitous arguments of earlier workers. A further type of nonlinear integral equation is also given. The method is also applicable to coordinate systems other than ellipsoidal, and also to the special functions which arise in the solution by separation of variables of Helmholtz's equation. We illustrate this by obtaining Sleeman's [8] integral equations for the ellipsoidal wave functions.

**2. Mathematical preliminaries.** Lamé's differential equation is

$$(4) \quad \frac{d^2 W}{d\zeta^2} + \{h - n(n+1) \text{sn}^2 \zeta\} W = 0,$$

where  $k$  is the modulus of the elliptic function  $\text{sn}(\zeta, k)$ . For finite periodic solutions, the parameter  $n$  is a nonnegative integer, and the eigenvalue  $h$  has one of  $2n + 1$  values for which (4) has a Lamé polynomial as a first solution. Such polynomials have the form

$$(5) \quad \text{sn}^\rho \zeta \text{cn}^\sigma \zeta \text{dn}^\tau \zeta U_p(\text{sn}^2 \zeta),$$

where  $\rho, \sigma, \tau = 0$  or  $1$ ,  $U_p(\text{sn}^2 \zeta)$  is a polynomial of degree  $p$  in the argument  $\text{sn}^2 \zeta$ , and  $2p + \rho + \sigma + \tau = n$ . We adopt the notation of Arscott [6], and denote the eight types of Lamé polynomial, obtained by giving  $\rho, \sigma, \tau$  their possible values, by  $uE_{2n}^m(\zeta)$ ,  $scE_{2n+2}^m(\zeta)$ ,  $sE_{2n+1}^m(\zeta)$ ,  $cE_{2n+1}^m(\zeta)$ ,  $dE_{2n+1}^m(\zeta)$ ,  $sdE_{2n+2}^m(\zeta)$ ,  $cdE_{2n+2}^m(\zeta)$ , and  $scdE_{2n+3}^m(\zeta)$ , where  $m = 0, 1, \dots, n$ . In this notation a letter  $s, c$  or  $d$  is prefixed to the  $E$ -symbol according as to whether the corresponding function  $\text{sn} \zeta, \text{cn} \zeta$ , or  $\text{dn} \zeta$  appears in (5), and  $u$  (denoting unity) corresponds to  $\rho = \sigma = \tau = 0$ . The suffix in the  $E$ -symbol denotes the total degree of the polynomial, and  $m$  specifies the number of zeros in the interval  $0 < \zeta < K$ . We further adopt the normalization that the coefficient of the highest power of  $\text{sn} \zeta$  in the polynomial  $U_p(\text{sn}^2 \zeta)$  is to be unity.

Let  $E_n^m(\zeta)$  be one of the eight types of periodic Lamé polynomial; then (4) has a corresponding second nonperiodic solution  $F_n^m(\zeta)$ , defined by

$$(6) \quad F_n^m(\zeta) = (-1)^{\sigma+\tau}(2n+1)k^{2\tau+1}E_n^m(\zeta) \int_{iK'}^{\zeta} \frac{d\xi}{[E_n^m(\xi)]^2}.$$

The normalization adopted for  $E_n^m$  and (6) means that  $F_n^m(\zeta) \sim (\text{sn} \zeta)^{-n-1}$  as  $\zeta \rightarrow iK'$ , and the Wronskian relation is

$$(7) \quad E_n^m(\zeta)F_n^{m'}(\zeta) - E_n^{m'}(\zeta)F_n^m(\zeta) = (-1)^{\sigma+\tau}(2n+1)k^{2\tau+1}.$$

At a zero of  $E_n^m(\xi)$  the residue of the integrand in (6) is zero [16, p. 562]; hence the contour of integration in (6) can be deformed across a pole of the integrand without altering the value of the integral.

An ellipsoidal wave function of the first kind is a doubly periodic single-valued solution of the differential equation

$$(8) \quad \frac{d^2W}{d\zeta^2} - (a + bk^2 \text{sn}^2 \zeta + qk^4 \text{sn}^4 \zeta)W = 0,$$

where the parameters  $a, b$  are suitably chosen functions of  $q$ . The ellipsoidal wave functions of the first kind, which are in a one-one correspondence with the Lamé polynomials to which they reduce when  $q = 0$ , are denoted by  $el_n^m(\zeta)$ , and have the general form

$$(9) \quad el_n^m(\zeta) = \text{sn}^\rho \zeta \text{cn}^\sigma \zeta \text{dn}^\tau \zeta F(\text{sn}^2 \zeta),$$

where  $\rho, \sigma, \tau = 0$  or  $1$ , and  $F(\text{sn}^2 \zeta)$  denotes an integral function of  $\text{sn}^2 \zeta$ . Let  $el_n^m(\zeta)$  and  $E_n^m(\zeta)$  (the Lamé polynomial to which  $el_n^m(\zeta)$  reduces as  $q \rightarrow 0$ ) be written as

$$el_n^m(\zeta) = \text{sn}^\rho \zeta \text{cn}^\sigma \zeta \text{dn}^\tau \zeta \sum_{j=0}^{\infty} A_j(q) \text{sn}^{2j} \zeta$$

and

$$E_n^m(\zeta) = \text{sn}^\rho \zeta \text{cn}^\sigma \zeta \text{dn}^\tau \zeta \sum_{j=0}^N A_j(0) \text{sn}^{2j} \zeta,$$

where  $A_N(0) = 1$ . We then normalize  $el_n^m(\zeta)$  by the condition that  $A_N(q) = 1$  for all  $q$ .

A second independent solution of (8) which reduces to  $F_n^m(\zeta)$  as  $q \rightarrow 0$  is denoted by  $hl_n^m(\zeta)$ , and is normalized in such a way that

$$(10) \quad el_n^m(\zeta)hl_n^{m'}(\zeta) - el_n^{m'}(\zeta)hl_n^m(\zeta) = (-1)^{\sigma+\tau}(2n+1)k^{2\tau+1}.$$

A more precise specification of the second independent solution of concern to us will be given at a later stage.

We conclude this section by collecting together certain information regarding Legendre functions. The complete solution of Legendre's equation

$$\frac{d}{d\zeta} \left\{ (1-\zeta^2) \frac{dW}{d\zeta} \right\} + n(n+1)W = 0,$$

where  $n$  is a nonnegative integer, is

$$W = AP_n(\zeta) + BQ_n(\zeta),$$

$P_n(\zeta)$  being the usual Legendre polynomial of order  $n$ .

Adopting Copson's definition [9],  $Q_n(\zeta)$  is a single-valued function in the complex  $\zeta$ -plane, cut along the real axis from  $-1$  to  $1$ , given by

$$(11) \quad Q_n(\zeta) = \frac{1}{2} P_n(\zeta) \log \frac{\zeta+1}{\zeta-1} - W_{n-1}(\zeta),$$

where the logarithm has its principal value and  $W_{n-1}(\zeta)$  is a polynomial in  $\zeta$  of degree  $n-1$ . When  $\mu$  is real with  $|\mu| < 1$ , we have that

$$(12) \quad Q_n(\mu + i0) - Q_n(\mu - i0) = -\pi i P_n(\mu).$$

Defining  $\mathcal{Q}_n(\mu)$  by

$$(13) \quad \mathcal{Q}_n(\mu) = \frac{1}{2} P_n(\mu) \log \frac{1+\mu}{1-\mu} - W_{n-1}(\mu),$$

then

$$(14) \quad Q_n(\mu \pm i0) = \mathcal{Q}_n(\mu) \mp \frac{1}{2} \pi i P_n(\mu),$$

and

$$(15) \quad \mathcal{Q}_n(\mu) = \frac{1}{2} \{ Q_n(\mu + i0) + Q_n(\mu - i0) \}.$$

We also have the asymptotic formula that as  $|\zeta| \rightarrow \infty$ ,

$$(16) \quad Q_n(\zeta) \sim \frac{\pi^{1/2} \Gamma(n+1)}{2^{n+1} \Gamma(n+3/2)} \frac{1}{\zeta^{n+1}}$$

and the Wronskian relation

$$(17) \quad P_n(\zeta)Q'_n(\zeta) - P'_n(\zeta)Q_n(\zeta) = \frac{1}{1-\zeta^2}.$$

**3. Basic regions and representations of the  $F_n^m$ -functions.** For clarity of exposition we state first the fundamental representation theorem of Arscott [3].

**THEOREM 1.** *Let  $w(\beta)$  satisfy Lamé's equation (4), and let  $G(\alpha, \beta)$  satisfy the partial differential equation*

$$(18) \quad \frac{\partial^2 G}{\partial \alpha^2} - \frac{\partial^2 G}{\partial \beta^2} = n(n+1)k^2(\text{sn}^2 \alpha - \text{sn}^2 \beta)G,$$

where  $w, G$  are analytic when  $\alpha, \beta$  lie in regions  $R_\alpha, R_\beta$  of their planes. Let  $C$  be a path in the  $\beta$ -plane lying wholly in  $R_\beta$  and such that

$$(19) \quad (i) \quad G(\alpha, \beta) \frac{dw}{d\beta}(\beta) - w(\beta) \frac{\partial G}{\partial \beta}(\alpha, \beta)$$

has the same value at each end of  $C$ , and

$$(20) \quad (ii) \quad W(\alpha) = \int_C G(\alpha, \beta) w(\beta) d\beta$$

exists and, if singular, converges uniformly with respect to  $\alpha \in R_\alpha$ . Then  $W(\alpha)$  is a solution of Lamé's equation with the same  $n, h$  as  $w(\beta)$ .

Equation (18) may be transformed into a recognizable standard form by means of the equations

$$(21) \quad \begin{aligned} kr \operatorname{sn} \alpha \operatorname{sn} \beta &= r \sin \theta \cos \phi, \\ ikk'^{-1} r \operatorname{cn} \alpha \operatorname{cn} \beta &= r \sin \theta \sin \phi, \\ k'^{-1} r \operatorname{dn} \alpha \operatorname{dn} \beta &= r \cos \theta, \end{aligned}$$

which relate sphero-conal coordinates  $(r, \alpha, \beta)$  to spherical polar coordinates  $(r, \theta, \phi)$ . The result is that (19) becomes

$$(22) \quad \frac{\partial^2 G}{\partial \theta^2} + \cot \theta \frac{\partial G}{\partial \theta} + \operatorname{cosec}^2 \theta \frac{\partial^2 G}{\partial \phi^2} + n(n+1)G = 0,$$

which has the spherical harmonics of degree  $n$  as solutions. It is important to notice that the same equation (22) arises for each of the possible permutations of the right-hand sides of (21), a fact which leads to three representations of each of the eight species of  $F$ -functions.

We next introduce the concept of a *basic region* of the complex  $\alpha$ -plane in which representation (2) and its companions are valid. Slight variations in the basic regions are necessary depending on which of the three possible arguments appear in the  $Q$ -function, but each region differs only in the exclusion of certain line segments, on which the argument takes the values  $\pm 1$ , from the rectangle  $R$  defined by

$$(23) \quad R = \{\alpha; |\operatorname{Re} \alpha| \leq K, 0 \leq \operatorname{Im} \alpha \leq 2K'\};$$

further, the zeros of the arguments of the  $Q$ -functions are all boundary points of the regions. The basic regions also have the properties that in order to obtain formulae for the Lamé function of the second kind in contiguous regions, either an analytic continuation across the boundary is necessary, or, if the line segment between contiguous regions consists of points at which the integral (2), say, does not exist (e.g.,  $\operatorname{Re} \alpha = K$ ), then an examination of the singularity of the  $F$ -function at the center of the contiguous region is necessary.

In the subsequent analysis  $\{\alpha_1, \alpha_2\}$  will denote the straight-line segment (including end points), in the complex  $\alpha$ -plane joining the points  $\alpha_1$  and  $\alpha_2$ . Let  $R_d$  be the *basic region* of the  $\alpha$ -plane defined by

$$(24) \quad R_d = R \setminus [ -K, K ] \cup \{ -K + 2iK', K + 2iK' \},$$

and let  $R_\beta$  be a region of the  $\beta$ -plane which contains  $\{0, K\}$ . (The subscript  $d$  in  $R_d$  indicates that the basic region is appropriate to a representation in which the argument of the Legendre function contains the  $\operatorname{dn}$ -Jacobian elliptic functions.) We further define  $R_d^*$  by

$$(25) \quad R_d^* = R_d \setminus [ \{ K, K + 2iK' \} \cup \{ -K, -K + 2iK' \} ].$$

The solution

$$Q_{2n}(\cos \theta) = Q_{2n}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta)$$



of (22) is an analytic function of its argument for  $\alpha \in R_d^*$  and  $\beta \in \{0, K\}$ , and it is a straightforward matter to verify, using Theorem 1, that

$$(26) \quad W(\alpha) = \int_0^K Q_{2n}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta, \quad \alpha \in R_d^*,$$

is a solution of Lamé's equation with the same  $n, h$  as  $uE_{2n}(\alpha)$ . Using (16), an examination of the integrand in (26) shows that  $W(\alpha)$  behaves like  $(\operatorname{sn} \alpha)^{-(2n+1)}$  as  $\alpha \rightarrow iK', \alpha \in R_d^*$ , which is precisely the behavior of  $uF_{2n}^m(\alpha)$  in this limit. Thus, for some  $\lambda$ , depending only on  $n$  and  $m$ , we have that

$$(27) \quad uF_{2n}^m(\alpha) = \lambda \int_0^K Q_{2n}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta, \quad \alpha \in R_d^*.$$

Consider next the appropriate form which (27) takes when  $\alpha \in R_d \setminus R_d^* = \{K, K + 2iK'\} \cup \{-K, -K + 2iK'\}$ . Let  $\alpha = K + iuK'$ , where  $0 \leq u \leq 2$ . Then, as long as  $K + iuK' - w \in R_d^*$ ,

$$\lim_{w \rightarrow 0} Q_{2n}(k'^{-1} \operatorname{dn}(K + iuK' - w) \operatorname{dn} \beta) = Q_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta - i0),$$

where  $k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta$  is real and numerically less than one for  $0 \leq u \leq 2, 0 \leq \beta \leq K$ . Thus, from (27) and (14) we have

$$(28) \quad \begin{aligned} uF_{2n}^m(K + iuK') &= \lambda \int_0^K Q_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta - i0) uE_{2n}^m(\beta) d\beta \\ &= \lambda \left\{ \int_0^K \mathcal{Q}_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta \right. \\ &\quad \left. + \frac{1}{2} \pi i \int_0^K P_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta \right\} \end{aligned}$$

and, using the results in [3], the second integral in (28) is a multiple of  $uE_{2n}^m(K + iuK')$ . In a similar manner,

$$(29) \quad \begin{aligned} uF_{2n}^m(-K + iuK') &= \lambda \left\{ \int_0^K \mathcal{Q}_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta \right. \\ &\quad \left. - \frac{1}{2} \pi i \int_0^K P_{2n}(k'^{-1} \operatorname{dn}(K + iuK') \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta \right\}, \end{aligned}$$

and (27), (28) and (29) constitute the required integral representation of  $uF_{2n}^m(\alpha)$  for all  $\alpha \in R_d$ . Due to the discontinuity of  $Q_{2n}(\zeta)$  across the cut in the  $\zeta$ -plane, the right-hand side of (26) is discontinuous across  $\operatorname{Re} \alpha = \pm K$ , and we investigate in a later section the analytic continuation of the right member of (26) across  $\operatorname{Re} \alpha = \pm K$  into regions contiguous with the basic region  $R_d$ .

Two further representations for  $uF_{2n}^m(\alpha)$  follow on permuting the right-hand sides of (21). However, as Arscott [3] has pointed out, it is necessary to choose different paths of integration for the application of Theorem 1. Further, with the change in argument of the  $Q$ -functions, the validity conditions change, leading to different basic regions.

Consider the sphero-conal transformation in which  $\cos \theta = k \operatorname{sn} \alpha \operatorname{sn} \beta$ . Let  $R_s$  be the basic region of the  $\alpha$ -plane defined by

$$(30) \quad R_s = R \setminus [\{K, K + 2iK'\} \cup \{-K, -K + 2iK'\}],$$

where the subscript  $s$  is to be interpreted analogously to the  $d$  in  $R_d$ . Further  $R_\beta$  is a region of the  $\beta$ -plane which contains the straight-line segment  $\{K, K + iK'\}$ , and

$$R_s^* = R_s \setminus \{[-K, K] \cup [-K + 2iK', K + 2iK']\}.$$

Theorem 1 and similar reasoning to that used in deriving (27) now show that

$$(31) \quad uF_{2n}^m(\alpha) = \lambda \int_K^{K+iK'} Q_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta, \quad \alpha \in R_s^*$$

for some  $\lambda$  possibly different from that in (27). For  $\alpha \in R_s \setminus R_s^*$ , the formulae equivalent to (28) and (29) are

$$(32) \quad uF_{2n}^m(uK) = \lambda \left\{ \int_K^{K+iK'} \mathcal{Q}_{2n}(k \operatorname{sn} uK \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta \right. \\ \left. - \frac{1}{2} \pi i \int_K^{K+iK'} P_{2n}(k \operatorname{sn} uK \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta \right\},$$

and

$$(33) \quad uF_{2n}^m(uK + 2iK') = \lambda \left\{ \int_K^{K+iK'} \mathcal{Q}_{2n}(k \operatorname{sn} uK \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta \right. \\ \left. + \frac{1}{2} \pi i \int_K^{K+iK'} P_{2n}(k \operatorname{sn} uK \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta \right\}$$

for  $|u| \leq 1$ . Again the second integrals in (32) and (33) are multiples of  $uE_{2n}^m(uK)$ .

The third sphero-conal transformation sets  $\cos \theta = (ik/k') \operatorname{cn} \alpha \operatorname{cn} \beta$ . The appropriate basic region is now  $R_c$  given by

$$(34) \quad R_c = R \setminus \{[0, 2iK'] \cup [-K + iK', K + iK']\},$$

and we set

$$R_c^* = R_c \setminus \partial R_c.$$

The integration path suggested by Arscott [3] is the piecewise linear path  $\Gamma$  in the  $\beta$ -plane, defined by

$$\Gamma = \{0, K\} \cup \{K, K + iK'\}.$$

The resulting representation is

$$(35) \quad uF_{2n}^m(\alpha) = \lambda \int_\Gamma Q_{2n}\left(\frac{ik}{k'} \operatorname{cn} \alpha \operatorname{cn} \beta\right) uE_{2n}^m(\beta) d\beta, \quad \alpha \in R_c^*$$

for some scalar  $\lambda$ . As before, the points  $\alpha \in R \setminus R_c$  are excluded from the basic region since the argument of the  $Q_{2n}$ -function can then assume the values  $\pm 1$  on the integration path. For  $\alpha \in \partial R_c$ , formulae such as (28) and (29) are easily written down, (35) being discontinuous as  $\alpha$  varies across  $\partial R_c$ .

To generate the complete set of 24 representations of the eight types of Lamé functions in the appropriate basic regions, we use as kernels in Theorem 1 the four simplest spherical harmonics of the second kind, i.e.,  $Q_n(\cos \theta)$ ,  $Q_n^1(\cos \theta) \sin \phi$ ,  $Q_n^1(\cos \theta) \cos \phi$ ,  $Q_n^2(\cos \theta) \sin 2\phi$ , which are respectively equivalent to  $Q_n(\cos \theta)$ ,  $\sin \theta \sin \phi Q_n'(\cos \theta)$ ,  $\sin \theta \cos \phi Q_n'(\cos \theta)$  and  $\sin^2 \theta \sin \phi \cos \phi Q_n''(\cos \theta)$ , in each of which  $n$  may be either even or odd. (Here and

subsequently the prime denotes differentiation with respect to the argument of the  $Q$ -function.)

Expressing these functions in terms of  $\alpha$  and  $\beta$  by means of (21) and the other formulae obtained by permuting the right-hand sides of (21), we have 24 distinct functions  $G(\alpha, \beta)$  suitable as kernels in Theorem 1. Writing  $S = k \operatorname{sn} \alpha \operatorname{sn} \beta$ ,  $C = ikk'^{-1} \operatorname{cn} \alpha \operatorname{cn} \beta$ ,  $D = k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta$ , these 24 kernels can be allocated to the eight types of Lamé function of the second kind as can be seen in Table 1.

TABLE 1

	I $uF_{2n}(\alpha)$	II $sF_{2n+1}(\alpha)$	III $cF_{2n+1}(\alpha)$	IV $dF_{2n+1}(\alpha)$
(i)	$Q_{2n}(S)$	$Q_{2n+1}(S)$	$Q_{2n+1}(C)$	$Q_{2n+1}(D)$
(ii)	$Q_{2n}(C)$	$SQ'_{2n+1}(C)$	$CQ'_{2n+1}(S)$	$DQ'_{2n+1}(S)$
(iii)	$Q_{2n}(D)$	$SQ'_{2n+1}(D)$	$CQ'_{2n+1}(D)$	$DQ'_{2n+1}(C)$
	V $scF_{2n+2}(\alpha)$	VI $sdF_{2n+2}(\alpha)$	VII $cdF_{2n+2}(\alpha)$	VIII $scdF_{2n+3}(\alpha)$
(i)	$CQ'_{2n+2}(S)$	$DQ'_{2n+2}(S)$	$DQ'_{2n+2}(C)$	$CDQ''_{2n+3}(S)$
(ii)	$SQ'_{2n+2}(C)$	$SQ'_{2n+2}(D)$	$CQ'_{2n+2}(D)$	$SDQ''_{2n+3}(C)$
(iii)	$SCQ''_{2n+2}(D)$	$SDQ''_{2n+2}(C)$	$CDQ''_{2n+2}(S)$	$SCQ''_{2n+3}(D)$

In this Table 1, I(i), I(ii), I(iii) and II(ii) are the kernels obtained by Arscott.

It is easily shown that the kernels as given in Table 1 do correspond to the Lamé functions under which they are tabulated, provided that the limits of integration are chosen as in (26), (31) and (35), the arguments of the  $Q$ -functions being  $D$ ,  $S$  and  $C$ , respectively. Thus, for values of  $\alpha$  within the respective basic regions we have the following 24 representations, the values of  $\lambda$  differing from equation to equation.

$$\begin{aligned}
 \text{I (i)} \quad uF_{2n}^m(\alpha) &= \lambda \int_K^{K+iK'} Q_{2n}(S) uE_{2n}^m(\beta) d\beta, \\
 \text{(ii)} \quad uF_{2n}^m(\alpha) &= \lambda \int_\Gamma Q_{2n}(C) uE_{2n}^m(\beta) d\beta, \\
 \text{(iii)} \quad uF_{2n}^m(\alpha) &= \lambda \int_0^K Q_{2n}(D) uE_{2n}^m(\beta) d\beta, \\
 \text{II (i)} \quad sF_{2n+1}^m(\alpha) &= \lambda \int_K^{K+iK'} Q_{2n+1}(S) sE_{2n+1}^m(\beta) d\beta, \\
 \text{(ii)} \quad sF_{2n+1}^m(\alpha) &= \lambda \int_\Gamma SQ'_{2n+1}(C) sE_{2n+1}^m(\beta) d\beta, \\
 \text{(iii)} \quad sF_{2n+1}^m(\alpha) &= \lambda \int_0^K SQ'_{2n+1}(D) sE_{2n+1}^m(\beta) d\beta, \\
 \text{III (i)} \quad cF_{2n+1}^m(\alpha) &= \lambda \int_\Gamma Q_{2n+1}(C) cE_{2n+1}^m(\beta) d\beta, \\
 \text{(ii)} \quad cF_{2n+1}^m(\alpha) &= \lambda \int_K^{K+iK'} CQ'_{2n+1}(S) cE_{2n+1}^m(\beta) d\beta, \\
 \text{(iii)} \quad cF_{2n+1}^m(\alpha) &= \lambda \int_0^K CQ'_{2n+1}(D) cE_{2n+1}^m(\beta) d\beta,
 \end{aligned}$$

IV (i)  $dF_{2n+1}^m(\alpha) = \lambda \int_0^K Q_{2n+1}(D) dE_{2n+1}^m(\beta) d\beta,$

(ii)  $dF_{2n+1}^m(\alpha) = \lambda \int_K^{K+iK'} DQ'_{2n+1}(S) dE_{2n+1}^m(\beta) d\beta,$

(iii)  $dF_{2n+1}^m(\alpha) = \lambda \int_{\Gamma} DQ'_{2n+1}(C) dE_{2n+1}^m(\beta) d\beta,$

V (i)  $scF_{2n+2}^m(\alpha) = \lambda \int_K^{K+iK'} CQ'_{2n+2}(S) scE_{2n+2}^m(\beta) d\beta,$

(ii)  $scF_{2n+2}^m(\alpha) = \lambda \int_{\Gamma} SQ'_{2n+2}(C) scE_{2n+2}^m(\beta) d\beta,$

(iii)  $scF_{2n+2}^m(\alpha) = \lambda \int_0^K SCQ''_{2n+2}(D) scE_{2n+2}^m(\beta) d\beta,$

VI (i)  $sdF_{2n+2}^m(\alpha) = \lambda \int_K^{K+iK'} DQ'_{2n+2}(S) sdE_{2n+2}^m(\beta) d\beta,$

(ii)  $sdF_{2n+2}^m(\alpha) = \lambda \int_0^K SQ'_{2n+2}(D) sdE_{2n+2}^m(\beta) d\beta,$

(iii)  $sdF_{2n+2}^m(\alpha) = \lambda \int_{\Gamma} SDQ''_{2n+2}(C) sdE_{2n+2}^m(\beta) d\beta,$

VII (i)  $cdF_{2n+2}^m(\alpha) = \lambda \int_{\Gamma} DQ'_{2n+2}(C) cdE_{2n+2}^m(\beta) d\beta,$

(ii)  $cdF_{2n+2}^m(\alpha) = \lambda \int_0^K CQ'_{2n+2}(D) cdE_{2n+2}^m(\beta) d\beta,$

(iii)  $cdF_{2n+2}^m(\alpha) = \lambda \int_K^{K+iK'} CDQ''_{2n+2}(S) cdE_{2n+2}^m(\beta) d\beta,$

VIII (i)  $scdF_{2n+3}^m(\alpha) = \lambda \int_K^{K+iK'} CDQ''_{2n+3}(S) scdE_{2n+3}^m(\beta) d\beta,$

(ii)  $scdF_{2n+3}^m(\alpha) = \lambda \int_{\Gamma} SDQ''_{2n+3}(C) scdE_{2n+3}^m(\beta) d\beta,$

(iii)  $scdF_{2n+3}^m(\alpha) = \lambda \int_0^K SCQ''_{2n+3}(D) scdE_{2n+3}^m(\beta) d\beta.$

**4. Calculation of the  $\lambda$ -multipliers.** In this section we show how to calculate the  $\lambda$ -multiplier in the representations of the  $F$ -functions given in § 3. Two methods will be indicated, and illustrated by detailed consideration of formulae I(i), (ii) and (iii). In what follows, the multiplier will be denoted simply by  $\lambda$ , but it must be remembered that its value is a function of  $n, m$  and  $k$ , and can vary between representations of the same  $F$ -function.

Consider first formula I(i), namely

$$uF_{2n}^m(\alpha) = \lambda \int_K^{K+iK'} Q_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta, \quad \alpha \in R_s.$$

Differentiation with respect to  $\alpha$  gives

$$(36) \quad uF_{2n}^{m'}(\alpha) = \lambda k \operatorname{cn} \alpha \operatorname{dn} \alpha \int_K^{K+iK'} Q'_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) \operatorname{sn} \beta uE_{2n}^m(\beta) d\beta,$$

where the prime denotes differentiation with respect to argument. We now set  $\alpha = iv, v \in \mathbb{R}$ , and let  $v \rightarrow 0+$ ; in this limit (36) yields

$$(37) \quad uF_{2n}^{m'}(0) = \lambda k Q'_{2n}(i0) \int_K^{K+iK'} \operatorname{sn} \beta uE_{2n}^m(\beta) d\beta.$$

Now from (7), (17) and the result  $uE_{2n}^{m'}(0) = 0$ , we have that

$$(38) \quad \text{and} \quad uF_{2n}^{m'}(0) = (4n + 1)k/uE_{2n}^m(0) \\ Q'_{2n}(i0) = \{P_{2n}(0)\}^{-1} = (-1)^n 2^{2n} (n!)^2 / (2n)!.$$

Thus, from (37) and (38) we have

$$(39) \quad \lambda = (4n + 1)P_{2n}(0)/uE_{2n}^m(0) \int_K^{K+iK'} \operatorname{sn} \beta uE_{2n}^m(\beta) d\beta.$$

In tables prepared by Arscott and Khabaza [10] the coefficients of powers of  $\operatorname{sn}^{2r} \beta$  in the expansion of  $uE_{2n}^m(\beta)$  are given for a range of  $n, m$  and  $k$ . Thus, using the reduction formulae given in [11] for integrals of powers of  $\operatorname{sn} \beta$ , the integral on the right-hand side of (39) can be computed.

Further,  $uE_{2n}^m(0)$  can be found from [11], and hence  $\lambda$  is furnished explicitly by (39). As an example, suppose that  $n = m = 0$ . Then  $uE_0^0(\beta) = 1$  and

$$\int_K^{K+iK'} \operatorname{sn} \beta uE_0^0(\beta) d\beta = \frac{i\pi}{2k}.$$

It follows that  $\lambda = 2k/\pi i$  and

$$uF_0^0(\alpha) = \frac{2k}{\pi i} \int_K^{K+iK'} Q_0(k \operatorname{sn} \alpha \operatorname{sn} \beta) d\beta,$$

that is

$$(40) \quad \alpha - iK' = \frac{1}{\pi i} \int_K^{K+iK'} \log \left( \frac{k \operatorname{sn} \alpha \operatorname{sn} \beta + 1}{k \operatorname{sn} \alpha \operatorname{sn} \beta - 1} \right) d\beta, \quad \alpha \in \mathbb{R}_s.$$

Equation (40) is an example of the rich variety of definite integrals which our representations yield.

A second method of evaluating  $\lambda$  consists of examining the form of each side of (36) as  $\alpha \rightarrow iK'$ . In this limit,  $\operatorname{sn} \alpha \rightarrow \infty$ , and (16) shows that

$$(41) \quad Q_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) \sim \frac{\pi^{1/2}(2n)!}{(2k)^{2n+1}\Gamma(2n + 3/2)} \frac{1}{(\operatorname{sn} \alpha \operatorname{sn} \beta)^{2n+1}}.$$

Now  $uF_{2n}^m(\alpha) \sim (\operatorname{sn} \alpha)^{-2n-1}$  as  $\alpha \rightarrow iK'$ ; thus combining this asymptotic form with I(i) and (41), we have that

$$(42) \quad \lambda = (2k)^{2n+1}\Gamma\left(2n + \frac{3}{2}\right)/\pi^{1/2}(2n)! \int_K^{K+iK'} (\operatorname{sn} \beta)^{2n+1} uE_{2n}^m(\beta) d\beta.$$

The integral in (42) is a little more awkward to evaluate than that in (39); hence (39) is

preferred. When  $n = m = 0$ ,

$$\int_K^{K+iK'} \operatorname{ns} \beta \, d\beta = [\log \{\operatorname{sn} \beta / (\operatorname{cn} \beta + \operatorname{dn} \beta)\}]_K^{K+iK'} = \frac{1}{2}\pi i,$$

and (42) reproduces the value  $\lambda = 2k/\pi i$ .

We next turn to I(ii), viz.,

$$uF_{2n}^m(\alpha) = \lambda \int_0^K Q_{2n}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) uE_{2n}^m(\beta) \, d\beta, \quad \alpha \in R_d.$$

After differentiating with respect to  $\alpha$ , the appropriate limiting process is to set  $\alpha = K + iK' - u$ ,  $u \in \mathbb{R}$ , and let  $u \rightarrow 0+$ , giving

$$uF_{2n}^{m'}(K + iK') = i\lambda Q'_{2n}(-i0) \int_0^K uE_{2n}^m(\beta) \operatorname{dn} \beta \, d\beta.$$

Use of the Wronskians (7) and (17) now shows that

$$(43) \quad \lambda = k(4n + 1)P_{2n}(0)/iuE_{2n}^m(K + iK') \int_0^K uE_{2n}^m(\beta) \operatorname{dn} \beta \, d\beta.$$

Since  $uE_{2n}^m(\alpha) = \sum_{r=0}^n a_r \operatorname{sn}^{2r} \alpha$ , where the coefficients  $a_r$  (with  $a_n = 1$ ) are available from [10], the integral in (43) is reduced to a sum of elementary trigonometric integrals by the substitution  $\psi = \operatorname{am} \beta$ . Thus, when  $n = m = 0$ , (43) gives  $\lambda = 2k/\pi i$ , and therefore

$$(44) \quad uF_0^0(\alpha) = \frac{2k}{\pi i} \int_0^K Q_0(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) \, d\beta,$$

that is,

$$(45) \quad \alpha - iK' = \frac{1}{\pi i} \int_0^K \log \left( \frac{k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta + 1}{k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta - 1} \right) \, d\beta, \quad \alpha \in R_d.$$

Putting  $\alpha = K + iuK' - w$ ,  $0 < u < 1$  in (44), letting  $w \rightarrow 0+$  and using (14), (13), the integral in (44) becomes

$$\frac{1}{\pi i} \left\{ \int_0^K \log \left( \frac{1 + a \operatorname{dn} \beta}{1 - a \operatorname{dn} \beta} \right) \, d\beta + \pi i \int_0^K \, d\beta \right\},$$

where  $a = k'^{-1} \operatorname{dn}(K + iuK')$ . It follows that

$$(46) \quad \int_0^K \log \left( \frac{1 + a \operatorname{dn} \beta}{1 - a \operatorname{dn} \beta} \right) \, d\beta = \pi(1 - u)K',$$

a result which is in accord with entry 801.07 of [11]. Thus, (45) constitutes a generalization of that entry.

To apply the second method of evaluating  $\lambda$ , it is only necessary to note that as  $\alpha \rightarrow iK'$ ,  $k'^{-1} \operatorname{dn} \alpha \sim (-ik'/k') \operatorname{sn} \alpha$ . The arguments leading to (42) now show that

$$(47) \quad \lambda = (2k/k')^{2n+1} \Gamma(2n + \frac{3}{2}) (-1)^n / i\pi^{1/2} (2n)! \int_0^K (\operatorname{nd} \beta)^{2n+1} uE_{2n}^m(\beta) \, d\beta.$$

Again, the definite integral in (47) is more complicated than that in (43).

The final member of the trio of representations of  $uF_{2n}^m(\alpha)$  is I(ii). The appropriate analytic process is again differentiation with respect to  $\alpha$ , now followed by setting  $\alpha = K - u + iv$ ,  $u, v \in \mathbb{R}$ , and taking the double limit  $u, v \rightarrow 0+$  (thereby remaining within

$R_c$ ). Then

$$\lim_{u,v \rightarrow 0^+} \int_0^K Q'_{2n} \left( \frac{ik}{k'} \operatorname{cn} \alpha \operatorname{cn} \beta \right) uE_{2n}^n(\beta) \operatorname{cn} \beta \, d\beta = Q'_{2n}(i0) \int_0^K uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta$$

and

$$\begin{aligned} \lim_{u,v \rightarrow 0^+} \int_K^{K+iK'} Q'_{2n} \left( \frac{ik}{k'} \operatorname{cn} \alpha \operatorname{cn} \beta \right) uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta \\ = Q'_{2n}(-i0) \int_K^{K+iK'} uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta. \end{aligned}$$

Since  $Q'_{2n}(i0) = Q'_{2n}(-i0) = \{P_{2n}(0)\}^{-1}$ , it follows that

$$\lim_{u,v \rightarrow 0^+} \int_{\Gamma} Q'_{2n} \left( \frac{ik}{k'} \operatorname{cn} \alpha \operatorname{cn} \beta \right) uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta = \frac{1}{P_{2n}(0)} \int_{\Gamma} uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta,$$

and

$$(48) \quad uF_{2n}^{m'}(K) = \frac{-ik\lambda}{P_{2n}(0)} \int_{\Gamma} uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta.$$

Use of the Wronskian now gives the result that

$$(49) \quad \lambda = i(4n+1)P_{2n}(0)/uE_{2n}^m(K) \int_{\Gamma} uE_{2n}^m(\beta) \operatorname{cn} \beta \, d\beta,$$

and the integral in the denominator of (49) can be written as a sum of integrals of the form  $\int_{\Gamma} \operatorname{cn}^{2r+1} \beta \, d\beta$ .

In the special case  $n = m = 0$ , the result

$$\int_{\Gamma} \operatorname{cn} \beta \, d\beta = \frac{\pi}{2k}$$

shows that  $\lambda = 2ik/\pi$ , a value different from that obtained for the two previous representations. A form for  $\lambda$  alternative to (49) is readily obtained by means of the limit  $\alpha \rightarrow iK'$ , and involves integrals of the form

$$\int_{\Gamma} (\operatorname{nc} \beta)^{2n+1} uE_{2n}^m(\beta) \, d\beta.$$

In the derivation of (39), (43) and (49), differentiation with respect to  $\alpha$  was first carried out in order to make use, in the Wronskian, of the vanishing of  $uE_{2n}^{m'}(\alpha)$  at  $\alpha = 0, K + iK'$  and  $K$  respectively. However, if  $E_{2n}^m$  itself vanishes for the appropriate value of  $\alpha$ , the prior differentiation is unnecessary. Consider, for example, representation VI(ii), viz.,

$$sdF_{2n+2}^m(\alpha) = \lambda k \operatorname{sn} \alpha \int_0^K \operatorname{sn} \beta Q'_{2n+2}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) sdE_{2n+2}^m(\beta) \, d\beta, \quad \alpha \in R_d.$$

Here  $sdE_{2n+2}^m(K + iK') = 0$  since  $\operatorname{dn}(K + iK') = 0$ . Thus, setting  $\alpha = K + iK' - u$  and letting  $u \rightarrow 0^+$ ,

$$(50) \quad sdF_{2n+2}^m(K + iK') = \lambda Q'_{2n+2}(-i0) \int_0^K \operatorname{sn} \beta sdE_{2n+2}^m(\beta) \, d\beta$$

with  $Q'_{2n+2}(-i0) = \{P_{2n+2}(0)\}^{-1}$ , and the form of  $sdE_{2n+2}^m(\beta)$  means that the integral in

(50) can be reduced to a sum of elementary trigonometric integrals by the substitution  $\psi = \text{am } \beta$ . The Wronskian (7) shows that

$$sdE_{2n+2}^{m'}(K + iK') sdF_{2n+2}^m(K + iK') = (4n + 5)k^3,$$

whence, from (50),

$$(51) \quad \lambda = (4n + 5)k^3 P_{2n+2}(0) / sdE_{2n+2}^{m'}(K + iK') \int_0^K \text{sn } \beta sdE_{2n+2}^m(\beta) d\beta.$$

As an example, take  $n = m = 0$  with  $sdE_2^0(\alpha) = \text{sn } \alpha \text{ dn } \alpha$ ; then

$$sdE_2^0(K + iK') = ik'/k, \quad \int_0^K \text{sn}^2 \beta \text{ dn } \beta d\beta = \frac{1}{4} \pi$$

and  $P_2(0) = -\frac{1}{2}$ . Thus,  $\lambda = 10ik^4/\pi k'$  and we have the explicit representation

$$(52) \quad sdF_2^0(\alpha) = \frac{10ik^5}{\pi k'} \text{sn } \alpha \int_0^K \text{sn } \beta Q'_{2n+2}(k'^{-1} \text{ dn } \alpha \text{ dn } \beta) sdE_2^0(\beta) d\beta.$$

(The functional form of  $sdF_2^0(\alpha)$  is given in [5].)

It is now a straightforward matter to codify the rules for computing the  $\lambda$ -multiplier in each of the 24 cases. For the contours of integration  $\{K, K + iK'\}$ ,  $\{0, K\}$  and  $\Gamma$ , respectively, the limits required are  $\alpha = iv$  with  $v \rightarrow 0+$ ,  $\alpha = K + iK' - u$  with  $u \rightarrow 0+$ , and  $\alpha = K - u + iv$  with  $u, v \rightarrow 0+$ . A prior differentiation is necessary for these contours in cases where  $E_n^m(0)$ ,  $E_n^m(K + iK')$  or  $E_n^m(K)$  are zero; thus differentiation is required for representations I(ii, iii), II(ii, iii), III(ii, iii), IV(ii, iii), V(iii), VI(iii) and VII(iii), i.e., precisely half the full set of representations.

**5. Representations outside the basic regions.** Let  $I(\alpha)$  denote the integral on the right-hand side of representation (26). Then  $I(\alpha + 2K) = I(\alpha)$ , and further  $I(\alpha)$  is discontinuous across  $\text{Re } \alpha = K$ . Specifically, if  $\alpha = K + iuK' - w$ , where  $0 < u < 2$  and  $w \in \mathbb{R}$ , then

$$(53) \quad \lim_{w \rightarrow 0-} I(K + iuK' - w) = \int_0^K Q_{2n}(k'^{-1} \text{ dn}(K + iuK') \text{ dn } \beta + i0) uE_{2n}^m(\beta) d\beta.$$

Thus, using (28), (53) and (12) we obtain

$$(54) \quad I(K + iuK' - 0) = I(K + iuK' + 0) + \pi i \int_0^K P_{2n}(k'^{-1} \text{ dn}(K + iuK') \text{ dn } \beta) uE_{2n}^m(\beta) d\beta,$$

exhibiting the discontinuity of  $I(\alpha)$  across  $\text{Re } \alpha = K$ . In a similar manner,  $I(\alpha)$  is discontinuous across  $\text{Re } \alpha = (2N + 1)K$  for all integer  $N$ . It is now apparent from (54) that the analytic continuation of the right-hand side of (26) across  $\text{Re } \alpha = K, 0 < \text{Im } \alpha < 2K'$ , is

$$(55) \quad \lambda \left\{ \int_0^K Q_{2n}(k'^{-1} \text{ dn } \alpha \text{ dn } \beta) uE_{2n}^m(\beta) d\beta + \pi i \int_0^K P_{2n}(k'^{-1} \text{ dn } \alpha \text{ dn } \beta) uE_{2n}^m(\beta) d\beta \right\}$$

for  $K < \text{Re } \alpha < 3K$ , the left-hand limit of (55) as  $\text{Re } \alpha \rightarrow K + 0$  being equal to the right-hand limit of (26) as  $\text{Re } \alpha \rightarrow K - 0$  for  $0 < \text{Im } \alpha < 2K'$ . Thus, for  $\alpha \in R_d$ , (26) and



(55) show that

$$\begin{aligned}
 uF_{2n}^m(\alpha + 2K) &= uF_{2n}^m(\alpha) + \pi i \lambda \int_0^K P_{2n}(k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta) uE_{2n}^m(\beta) d\beta, \\
 (56) \qquad &= uF_{2n}^m(\alpha) + \pi i \frac{\lambda}{\mu} uE_{2n}^m(\alpha),
 \end{aligned}$$

where  $\mu$  is the characteristic number in the appropriate homogeneous integral equation for  $uE_{2n}^m(\alpha)$ .

Equation (56) enables one to compute values of  $uF_{2n}^m(\alpha)$  in the contiguous region  $K < \operatorname{Re} \alpha < 3K, 0 < \operatorname{Im} \alpha < 2K'$ , from its values in the basic region. By repeated application of the above arguments we have that, for any integer  $N$ ,

$$(57) \qquad uF_{2n}^m(\alpha + 2NK) = uF_{2n}^m(\alpha) + N\pi i \frac{\lambda}{\mu} uE_{2n}^m(\alpha), \quad \alpha \in R_d.$$

A slightly different form of (56) can be obtained by direct transformation of the integral representation (6), which gives, for  $\alpha \in R_d$ ,

$$(58) \qquad uF_{2n}^m(\alpha + 2K) = (4n + 1)kuE_{2n}^m(\alpha + 2K) \int_{iK'}^{\alpha + 2K} \frac{d\xi}{[uE_{2n}^m(\xi)]^2}.$$

The path of integration  $\{iK', \alpha + 2K\}$  can be deformed into the union of  $\{iK', iK' + 2K\}$  and  $\{iK' + 2K, \alpha + 2K\}$ , and for the second segment the transformation  $\xi = 2K + \eta$  is applied. Equation (58) can then be rewritten as

$$\begin{aligned}
 (59) \qquad uF_{2n}^m(\alpha + 2K) &= (4n + 1)kuE_{2n}^m(\alpha) \left\{ \int_{iK'}^{iK' + 2K} \frac{d\xi}{[uE_{2n}^m(\xi)]^2} + \int_{iK'}^{\alpha} \frac{d\eta}{[uE_{2n}^m(\eta)]^2} \right\} \\
 &= uF_{2n}^m(\alpha) + (4n + 1)kuE_{2n}^m(\alpha) \int_{iK'}^{iK' + 2K} \frac{d\xi}{[uE_{2n}^m(\xi)]^2}.
 \end{aligned}$$

Comparing (56) and (59) we have the remarkable relationship

$$(60) \qquad \frac{\lambda}{\mu} = \frac{(4n + 1)k}{\pi i} \int_{iK'}^{iK' + 2K} \frac{d\xi}{[uE_{2n}^m(\xi)]^2}$$

between the multipliers appearing in the representations of the Lamé functions of the first and second kinds, respectively. It is worth remarking that (59) can also be obtained by investigating the singularity of  $uF_{2n}^m(\alpha)$  at  $\alpha = iK' + 2K$ .

Consider next the values of the  $uF$ -functions in the contiguous region  $|\operatorname{Re} \alpha| \leq K, 2K' < \operatorname{Im} \alpha < 4K'$ . Since the integral representation (26) does not exist for  $\operatorname{Im} \alpha = 2K'$ , it is not possible to use analytic continuation; however the direct transformation of (6) can still be used. Let  $\alpha \in R_d$ ; then

$$(61) \qquad uF_{2n}^m(\alpha + 2iK') = (4n + 1)kuE_{2n}^m(\alpha + 2iK') \int_{iK'}^{\alpha + 2iK'} \frac{d\xi}{[uE_{2n}^m(\xi)]^2},$$

and we decompose the path of integration as

$$\{iK', \alpha + 2iK'\} = \{iK', 3iK'\} \cup \{3iK', \alpha + 2iK'\}.$$

On the second segment the transformation  $\xi = 2iK' + \eta$  is applied, and the final result, comparable with (59), is

$$(62) \qquad uF_{2n}^m(\alpha + 2iK') = uF_{2n}^m(\alpha) + (4n + 1)kuE_{2n}^m(\alpha) \int_{iK'}^{3iK'} \frac{d\xi}{[uE_{2n}^m(\xi)]^2}.$$

To obtain a form of (62) without the awkward integral we can start with representation  $I(i)$  and effect the analytic continuation across  $\text{Im } \alpha = 2K'$ . Writing the  $\lambda$ -multiplier in  $I(i)$  as  $\tilde{\lambda}$ , we find that for  $\alpha \in R_s$ ,

$$(63) \quad uF_{2n}^m(\alpha + 2iK') = uF_{2n}^m(\alpha) + \pi i \frac{\tilde{\lambda}}{\tilde{\mu}} uE_{2n}^m(\alpha),$$

where  $\tilde{\mu}$  is the characteristic number in the homogeneous integral equation

$$uE_{2n}^m(\alpha) = \tilde{\mu} \int_K^{K+iK'} P_{2n}(k \operatorname{sn} \alpha \operatorname{sn} \beta) uE_{2n}^m(\beta) d\beta.$$

Comparing (62) and (63) it follows that

$$(64) \quad \frac{\tilde{\lambda}}{\tilde{\mu}} = \frac{(4n+1)}{\pi i} \int_{iK'}^{3iK'} \frac{d\xi}{[uE_{2n}^m(\xi)]^2}.$$

By repeated application of (56) and (63) it also follows that for  $\alpha \in R_s \cap R_d$ ,

$$(65) \quad uF_{2n}^m(\alpha + 2NK + 2MiK') = uF_{2n}^m(\alpha) + \pi i \left\{ \frac{N\lambda}{\mu} + \frac{M\tilde{\lambda}}{\tilde{\mu}} \right\} uE_{2n}^m(\alpha),$$

a formula which is valid by continuity throughout the whole of the  $\alpha$ -plane.

It is apparent that the process of analytic continuation may be applied to  $I(ii)$  across all sides of the basic region  $R_c$ . Further the arguments of this section can be applied to the remaining 21 formulae of the complete set of representations; the details are left to the reader.

**6. Potential problems and Liouville's integral equations.** As outlined in the Introduction, the object of this section is to give a simple method of deriving nonlinear integral equations for the Lamé functions of the Liouville type. The method depends on some simple results in potential theory.

Let  $S$  be a simple closed surface in three-dimensional space; let  $T$  denote the infinite region exterior to  $S$ , and  $T^*$  the region interior to  $S$ . Denote by  $V(r)$  a potential function which satisfies Laplace's equation  $\nabla^2 V = 0$  in  $T \cup T^*$  and is  $O(|r|^{-1})$  as  $|r| \rightarrow \infty$ . Defining  $R = |r - r'|$ , then

$$(66) \quad \nabla^2 \left( \frac{1}{R} \right) = -4\pi\delta(r - r'),$$

and we write

$$V(r) = \begin{cases} V^{(o)}(r), & r \in T, \\ V^{(i)}(r), & r \in T^*. \end{cases}$$

Using Green's theorem applied to the volume  $T$  and (66), it follows that for  $r \in T$ ,

$$(67) \quad V(r) = -\frac{1}{4\pi} \int_S \left\{ \frac{1}{R} \frac{\partial}{\partial n'} V^{(o)}(r') - V^{(o)}(r') \frac{\partial}{\partial n'} \left( \frac{1}{R} \right) \right\} dS',$$

where  $\partial/\partial n'$  denotes differentiation along the outward drawn normal to  $S$  with respect to primed coordinates. A second application of Green's theorem to the volume  $T^*$  with  $r \in T$  shows that

$$(68) \quad \int_S \left\{ V^{(i)}(r') \frac{\partial}{\partial n'} \left( \frac{1}{R} \right) - \frac{1}{R} \frac{\partial}{\partial n'} V^{(i)}(r') \right\} dS' = \int_{T^*} \left\{ V(r') \nabla'^2 \left( \frac{1}{R} \right) - \frac{1}{R} \nabla'^2 V \right\} dv' \equiv 0.$$

Suppose now that the potential function  $V$  is chosen to be continuous across  $S$ , i.e.,  $V^{(o)}(r) = V^{(i)}(r)$  for  $r \in S$ . Then (67) and (68) can be combined to give the result that

$$(69) \quad V(r) = \int_S \frac{\sigma(r')}{R} dS', \quad r \in T,$$

with

$$\sigma(r) = -\frac{1}{4\pi} \frac{\partial}{\partial n} \{V^{(o)}(r) - V^{(i)}(r)\}, \quad r \in S.$$

Further, it is easy to verify that (69) also hold for  $r \in T^*$ . Alternatively we can choose the potential function  $V$  to have a continuous normal derivative across  $S$ ; in this case (67) and (68) combine to give

$$(70) \quad V(r) = \int_S \sigma(r') \frac{\partial}{\partial n'} \left( \frac{1}{R} \right) dS', \quad r \in T,$$

with

$$\sigma(r) = \frac{1}{4\pi} \{V^{(o)}(r) - V^{(i)}(r)\}, \quad r \in S.$$

As with (69), (70) also hold for  $r \in T^*$ .

To derive integral equations of the Liouville type we introduce ellipsoidal coordinates  $(\alpha, \beta, \gamma)$  which are related to Cartesian coordinates  $(x, y, z)$  by the equations

$$(71) \quad \begin{aligned} x &= k^2 l \operatorname{sn} \alpha \operatorname{sn} \beta \operatorname{sn} \gamma, \\ y &= -k^2 l k'^{-1} \operatorname{cn} \alpha \operatorname{cn} \beta \operatorname{cn} \gamma, \\ z &= i l k'^{-1} \operatorname{dn} \alpha \operatorname{dn} \beta \operatorname{dn} \gamma. \end{aligned}$$

In (71)  $k$  is the modulus of the Jacobian elliptic functions;  $k'$ , the complementary modulus; and  $l$ , an arbitrary constant. To obtain all values of  $(x, y, z)$  it is necessary for  $\alpha, \beta, \gamma$  to vary in the ranges  $\alpha$  from  $-2K$  to  $2K$ ,  $\beta$  from  $K$  to  $K + 2iK'$ , and  $\gamma$  from  $iK'$  to  $K + iK'$ . The coordinate surface  $\gamma = \text{constant}$  is an ellipsoid (for more details of this coordinate system and ellipsoidal harmonics see Arscott [6]).

Let  $S$  be the ellipsoid  $\gamma = \gamma'$ , a constant, and consider the potential function  $V(\alpha, \beta, \gamma)$  which vanishes like the inverse distance at infinity (i.e., as  $\gamma \rightarrow iK'$ ), which is continuous across  $\gamma = \gamma'$ , and which takes on  $\gamma = \gamma'$  prescribed value

$$V(\alpha, \beta, \gamma') = E_n^m(\alpha) E_n^m(\beta) = E p_n^m(\alpha, \beta),$$

where  $E_n^m$  is an arbitrary Lamé polynomial, and the Lamé product notation of Arscott [6] has been introduced. In this geometry, the points of  $T$  correspond to  $\gamma \in (\gamma', iK')$ , and points of  $T^*$  to  $\gamma \in [K + iK', \gamma']$ , where  $\gamma'$  has the form  $\eta + iK'$ ,  $0 < \eta \leq K$ .

The form of the potential function  $V(\alpha, \beta, \gamma)$  satisfying the preceding specifications is given by

$$(72) \quad \begin{aligned} V^{(i)}(\alpha, \beta, \gamma) &= E p_n^m(\alpha, \beta) E_n^m(\gamma) / E_n^m(\gamma'), \quad \gamma \in [K + iK', \gamma'], \\ V^{(o)}(\alpha, \beta, \gamma) &= E p_n^m(\alpha, \beta) F_n^m(\gamma) / F_n^m(\gamma'), \quad \gamma \in [\gamma', iK']. \end{aligned}$$

To apply (69) to the potential (72) we let the point  $P' \in S$  have ellipsoidal coordinates  $(\alpha', \beta', \gamma')$ ; then

$$dS' \frac{\partial}{\partial n'} = -i k^2 l (\operatorname{sn}^2 \alpha' - \operatorname{sn}^2 \beta') d\alpha' d\beta' \frac{\partial}{\partial \gamma'}.$$

Thus, using (72) we have

$$(73) \quad \left( \frac{\partial V^{(o)}}{\partial n'} - \frac{\partial V^{(i)}}{\partial n'} \right) dS' = -ik^2 l (\text{sn}^2 \alpha' - \text{sn}^2 \beta') Ep_n^m(\alpha', \beta') \cdot \left[ \frac{\partial}{\partial \gamma} \left\{ \frac{F_n^m(\gamma)}{F_n^m(\gamma')} - \frac{E_n^m(\gamma)}{E_n^m(\gamma')} \right\} \right]_{\gamma=\gamma'}$$

Carrying out the differentiation and using the Wronskian (7), the term in square brackets in (73) can be reduced to

$$(74) \quad (-1)^{\sigma+\tau} (2n+1) k^{2\tau+1} / E_n^m(\gamma') F_n^m(\gamma')$$

It now follows from (69), (73), (74) and the definition of the integration region given in the Introduction that

$$(75) \quad V(\alpha, \beta, \gamma) = \frac{ik^{2\tau+3} l (-1)^{\sigma+\tau} (2n+1)}{8\pi E_n^m(\gamma') F_n^m(\gamma')} \iint_{\mathcal{R}} \frac{1}{R} Ep_n^m(\alpha', \beta') (\text{sn}^2 \alpha' - \text{sn}^2 \beta^2) d\alpha' d\beta'$$

Thus, from (72) and (75) we have that for  $\gamma \in (\gamma', iK')$ ,

$$(76) \quad F_n^m(\gamma) = \frac{ik^{2\tau+3} l (-1)^{\sigma+\tau} (2n+1)}{8\pi E_n^m(\alpha) E_n^m(\beta) E_n^m(\gamma')} M(\alpha, \beta, \gamma),$$

and for  $\gamma \in (K + iK', \gamma')$ ,

$$(77) \quad E_n^m(\gamma) = \frac{ik^{2\tau+3} l (-1)^{\sigma+\tau} (2n+1)}{8\pi E_n^m(\alpha) E_n^m(\beta) F_n^m(\gamma')} M(\alpha, \beta, \gamma),$$

where  $M(\alpha, \beta, \gamma)$  denotes the double integral on the right-hand side of (75). Formulae (76) and (77) are the results of Liouville [7], Arscott [3], and Sleeman [8].

We can also derive further nonlinear integral equations using (70). The potential function  $V(\alpha, \beta, \gamma)$  defined by

$$(78) \quad \begin{aligned} V^{(i)}(\alpha, \beta, \gamma) &= Ep_n^m(\alpha, \beta) \frac{E_n^m(\gamma)}{E_n^m(\gamma')}, & \gamma \in (K + iK', \gamma') \\ V^{(o)}(\alpha, \beta, \gamma) &= Ep_n^m(\alpha, \beta) \frac{F_n^m(\gamma)}{F_n^m(\gamma')}, & \gamma \in (\gamma', iK') \end{aligned}$$

has a continuous normal derivative across  $\gamma = \gamma'$ . Also, for  $(\alpha', \beta', \gamma') \in S$ , (78) and the Wronskian (7) show that

$$V^{(i)}(\alpha', \beta', \gamma') - V^{(o)}(\alpha', \beta', \gamma') = (-1)^{\sigma+\tau} (2n+1) k^{2\tau+1} \frac{Ep_n^m(\alpha', \beta')}{E_n^m(\gamma') F_n^m(\gamma')}$$

Thus, using (70) we obtain

$$(79) \quad V(\alpha, \beta, \gamma) = \frac{ik^{2\tau+3} l (-1)^{\sigma+\tau} (2n+1)}{8\pi E_n^m(\gamma') F_n^m(\gamma')} \iint_{\mathcal{R}} \frac{\partial}{\partial \gamma'} \left( \frac{1}{R} \right) Ep_n^m(\alpha', \beta') (\text{sn}^2 \alpha' - \text{sn}^2 \beta') d\alpha' d\beta',$$

and we have the integral equations

$$(80) \quad E_n^m(\gamma) = \frac{ik^{2\tau+3} l (-1)^{\sigma+\tau} (2n+1)}{8\pi E_n^m(\alpha) E_n^m(\beta) F_n^m(\gamma')} N(\alpha, \beta, \gamma)$$

for  $\gamma \in [K + iK', \gamma']$ , and

$$(81) \quad F_n^m(\gamma) = \frac{ik^{2\tau+3}l(-1)^{\sigma+\tau}(2n+1)}{8\pi E_n^m(\alpha)E_n^m(\beta)E_n^m(\gamma')} N(\alpha, \beta, \gamma)$$

for  $\gamma \in (\gamma', iK')$ , where  $N(\alpha, \beta, \gamma)$  denotes the double integral on the right-hand side of (79). Equations (80) and (81) seem to be new, although some special cases were implicit in the work of Sleeman [12] on low-frequency scalar diffraction by ellipsoids.

**7. Some generalizations.** There are a number of ways in which the arguments of § 6 can be generalized and here we briefly indicate some extensions.

It was pointed out by Arscott [3] that (77) is only valid for  $P(\alpha, \beta, \gamma)$  at the origin (i.e.,  $\alpha = 0, \beta = K, \gamma = K' + iK', R = l(k^2 \text{sn}^2 \alpha' - k^2 \text{cn}^2 \beta' - \text{dn}^2 \gamma')^{1/2} = r$ , say) in the case of the  $u$ -Lamé functions, leading to the formula

$$(82) \quad uF_{2n}^m(\gamma') = \frac{ik^3l(4n+1)}{8\pi uE_{2n}^m(0)uE_{2n}^m(K)uE_{2n}^m(K+iK')} \cdot \int_{\varphi} \int \frac{1}{r} uE_{2n}^m(\alpha', \beta')(\text{sn}^2 \alpha' - \text{sn}^2 \beta') d\alpha' d\beta'$$

((75) is of course valid for all the Lamé functions). To obtain nonnegative formulae of the type (82) for the remaining seven types of Lamé function it is necessary to replace the inverse distance Green's function by those corresponding to higher multipoles. To illustrate how this fits in with the arguments of § 6, observe first that

$$\nabla^2 \left( \frac{x-x'}{R^3} \right) = 4\pi \frac{\partial}{\partial x} \delta(r-r'),$$

where  $r = (x, y, z)$  and  $r' = (x', y', z')$ . Using the function  $(x-x')/R^3$  instead of  $1/R$  in the derivation of (70) now shows that

$$\frac{\partial V}{\partial x}(x, y, z) = -\frac{1}{4\pi} \int_S \frac{x-x'}{R^3} \left\{ \frac{\partial V^{(o)}}{\partial n'} - \frac{\partial V^{(i)}}{\partial n'} \right\} dS',$$

valid for all  $r$ . If we now use for  $V$  the potential function (72) with Lamé functions of the  $s$ -type and evaluate  $\partial V/\partial x$  at the origin, Arscott's formula for  $sF_{2n}^m(\gamma')$  (equation (5.10) of [3]) is reproduced.

We can also consider coordinate systems other than ellipsoidal. For example, oblate spheroidal coordinates  $(\xi, \eta, \phi)$  are related to Cartesian coordinates  $(x, y, z)$  by the transformation

$$(83) \quad \begin{aligned} x &= a\{(1+\xi^2)(1-\eta^2)\}^{1/2} \cos \phi, \\ y &= a\{(1+\xi^2)(1-\eta^2)\}^{1/2} \sin \phi, \\ z &= a\xi\eta, \end{aligned}$$

where  $a$  is a constant,  $\xi \geq 0$ , and  $-1 \leq \eta \leq 1$ . The surface  $\xi = \text{constant}$  is an oblate spheroid whose axis of symmetry is the  $z$ -axis. A potential function, suitable for the application of (69), which is continuous across  $\xi = \xi'$ , is provided by

$$(84) \quad \begin{aligned} V^{(i)} &= \frac{P_n^m(i\xi)}{P_n^m(i\xi')} P_n^m(\eta) \cos m\phi, & 0 \leq \xi \leq \xi', \\ V^{(o)} &= \frac{Q_n^m(i\xi)}{Q_n^m(i\xi')} P_n^m(\eta) \cos m\phi, & \xi \geq \xi', \end{aligned}$$

where  $n$  is a nonnegative integer and  $m$  is an integer with  $0 \leq n \leq m$ . On the surface  $\xi = \xi'$  we have that

$$dS' \frac{\partial}{\partial n'} = a(1 + \xi'^2) d\eta' d\phi' \frac{\partial}{\partial \xi'}.$$

Thus (83), (69) and the Wronskian for the associated Legendre functions lead to the results that for  $0 \leq \xi \leq \xi'$ ,

$$(85) \quad P_n^m(i\xi)P_n^m(\eta) \cos m\phi = \frac{ia(-1)^{m+1}(n+m)!}{4\pi(n-m)!Q_n^m(i\xi')} \int_0^{2\pi} d\phi' \int_{-1}^1 \frac{\cos m\phi'}{R} P_n^m(\eta') d\eta'$$

and for  $\xi > \xi'$ ,

$$(86) \quad Q_n^m(i\xi)P_n^m(\eta) \cos m\phi = \frac{ia(-1)^{m+1}(n+m)!}{4\pi(n-m)!P_n^m(i\xi')} \int_0^{2\pi} d\phi' \int_{-1}^1 \frac{\cos m\phi'}{R} P_n^m(\eta') d\eta'.$$

In (85) and (86),  $R$  is the distance between the points  $P, P'$  with oblate spheroidal coordinates  $(\xi, \eta, \phi), (\xi', \eta', \phi')$ . If these points have cylindrical polar coordinates  $(\rho, z, \phi)$  and  $(\rho', z', \phi')$ , then

$$\begin{aligned} \int_0^{2\pi} \frac{\cos m\phi'}{R} d\phi' &= 2\pi \cos m\phi \int_0^\infty e^{-k|z-z'|} J_m(k\rho) J_m(k\rho') dk \\ &= 2\pi \cos m\phi K_m(\rho, \rho'; z, z'), \end{aligned}$$

say, and (85), (86) become

$$(87) \quad P_n^m(i\xi)Q_n^m(\eta) = \frac{ia(-1)^{m+1}(\eta+m)!}{2(n-m)!Q_n^m(i\xi')} \int_{-1}^1 K_m(\rho, \rho'; z, z') P_n^m(\eta') d\eta',$$

$$0 \leq \xi \leq \xi',$$

and

$$(88) \quad Q_n^m(i\xi)P_n^m(\eta) = \frac{ia(-1)^{m+1}(n+m)!}{2(n-m)!P_n^m(i\xi')} \int_{-1}^1 K_m(\rho, \rho'; z, z') P_n^m(\eta') d\eta',$$

$$\xi \geq \xi'.$$

To relate (87) and (88) to known formulae, consider the special case  $\xi' = 0$  (i.e., the spheroid has degenerated into the disk-shaped region  $z' = 0, 0 \leq \rho' \leq a, 0 \leq \phi' \leq 2\pi$ ), with  $n, m$  even integers equal to  $2r$  and  $2s$ , say. In the limit  $\xi \rightarrow 0+$ ,  $\eta = (1 - \rho^2/a^2)^{1/2}$ , and using the result that

$$Q_{2r}^{2s}(i0+) = \frac{1}{2}\pi i(-1)^{s+1} P_{2r}^{2s}(0),$$

(88) gives the result that

$$(89) \quad P_{2r}^{2s}\{(1 - \rho^2/a^2)^{1/2}\} = \frac{2(-1)^s(2r+2s)!}{\pi a(2r-2s)! \{P_{2r}^{2s}(0)\}^2} \int_0^a \frac{\rho'}{(1 - \rho'^2/a^2)^{1/2}} \cdot K_{2s}(\rho, \rho'; 0, 0) P_{2r}^{2s}\{(1 - \rho'^2/a^2)^{1/2}\} d\rho'.$$

Formula (89) is a generalization of a result first written down by Popov for the case  $s = 0$ . Popov's original proof utilizes properties of special functions, and an alternative derivation is given in [14]. In a similar manner we may put  $\eta = 0$ , i.e.,  $\xi = (\rho^2/a^2 - 1)^{1/2}$ ,

in (88) thus obtaining the equation

$$(90) \quad Q_{2r}^{2s}\{i(\rho^2/a^2 - 1)^{1/2}\} = -\frac{i(2r+2s)!}{a(2r-2s)!\{P_{2r}^{2s}(0)\}^{1/2}} \int_0^a \frac{\rho'}{(1-\rho'^2/a^2)^{1/2}} \cdot K_{2s}(\rho, \rho'; 0, 0) P_{2r}^{2s}\{(1-\rho'^2/a^2)^{1/2}\} d\rho'.$$

As a final extension of § 6, suppose that  $V$ , instead of being a harmonic function, is everywhere regular and satisfies the Helmholtz equation  $(\nabla^2 + \chi^2)V = 0$  throughout space. As  $r = |r| \rightarrow \infty$ , we also impose on  $V$  a Sommerfeld condition of the form

$$r\left(\frac{\partial V}{\partial r} - i\chi V\right) \rightarrow 0 \quad \text{as } r \rightarrow \infty.$$

We can now repeat the arguments leading to equations (69) using the appropriate Green's function  $e^{i\chi R}/R$  instead of the inverse distance Green's function, to obtain the representations

$$(91) \quad V(r) = \int_S \frac{e^{i\chi R}}{R} \sigma(r') dS',$$

where

$$(92) \quad \sigma(r') = -\frac{1}{4\pi} \left\{ \frac{\partial V^{(o)}}{\partial n'} - \frac{\partial V^{(i)}}{\partial n'} \right\}$$

and  $R = |r - r'|$ . In (92)  $V^{(i)}$  is a wave function regular within  $S$ , and  $V^{(o)}$  is a wave function exterior to  $S$  satisfying the Sommerfeld condition and such that  $V^{(o)} = V^{(i)}$  for points on  $S$ .

To apply (91) to the ellipsoidal wave functions we put  $q = -\chi^2 l^2$  in (8). In ellipsoidal coordinates the Helmholtz equation then has solutions of the form

$$el_n^m(\alpha)el_n^m(\beta)el_n^m(\gamma)$$

which are regular at the origin, and solutions of the form

$$el_n^m(\alpha)el_n^m(\beta)hl_n^{(1)m}(\gamma)$$

behaving like  $e^{i\chi k \operatorname{sn} \gamma}/k \operatorname{sn} \gamma$ , i.e.,  $e^{ikr}/r$  as  $r \rightarrow \infty$ . In Arscott's terminology [6],  $hl_n^{(1)m}(\gamma)$  is an ellipsoidal wave function of the third kind. A continuous solution  $V$  suitable for use in (91) is now furnished by

$$(93) \quad \begin{aligned} V^{(i)} &= el_n^m(\alpha)el_n^m(\beta) \frac{el_n^m(\gamma)}{el_n^m(\gamma')}, & \gamma \in (K + iK', \gamma'), \\ V^{(o)} &= el_n^m(\alpha)el_n^m(\beta) \frac{hl_n^{(1)m}(\gamma)}{hl_n^{(1)m}(\gamma')}, & \gamma \in [\gamma', iK'], \end{aligned}$$

the surface  $S$  being the ellipsoid  $\gamma = \gamma'$ . Inserting (93) into (91) and (92), and using the Wronskian (10) now leads directly to the integral representations

$$(94) \quad hl_n^{(1)m}(\gamma) = \frac{ik^{2\tau+3}l(-1)^{\sigma+\tau}(2n+1)}{8\pi el_n^m(\alpha)el_n^m(\beta)hl_n^m(\gamma')} M(\alpha, \beta, \gamma, \chi), \quad \gamma \in [\gamma', iK'],$$

and

$$(95) \quad el_n^m(\gamma) = \frac{ik^{2\tau+3}l(-1)^{\sigma+\tau}(2n+1)}{8\pi el_n^m(\alpha)el_n^m(\beta)hl_n^{(1)m}(\gamma')} M(\alpha, \beta, \gamma, \chi), \quad \gamma \in (K + iK', \gamma'),$$

where

$$(96) \quad M(\alpha, \beta, \gamma, \chi) = \iint_{\mathcal{S}} \frac{e^{ixR}}{R} el_n^m(\alpha') el_n^m(\beta') (\text{sn}^2 \alpha' - \text{sn}^2 \beta') d\alpha' d\beta'.$$

Equations (95) and (96) were first obtained by Sleeman [8]. Integral equations analogous to (80), (81) and (5.10) of [3] are also easily written down (see [17]).

**8. Conclusions.** This paper has explored two different species of integral representation of the Lamé function of the second kind. In §§ 3, 4 and 5, the theory of the class of representations of  $F_n^m(\alpha)$  first introduced by Arscott [3] has been developed and brought to a reasonable state of completeness; indeed it is now a practical proposition to use the representations for the numerical computation of the functions. A lack of this capability has undoubtedly hindered the application of Lamé functions to problems of interest in applied mathematics, the analytic form of all but the lowest-order functions being hopelessly complicated.

In §§ 6 and 7 a concise and simple treatment is given of Liouville-type nonlinear integral equations for the Lamé and ellipsoidal wave functions. The method used seems to have been overlooked in the literature, although it would be unwise to claim that it is original. Indeed, it is related to the standard method of expansion of Green's functions in terms of eigenfunctions, as described in [15]. However, the ease with which the integral equations and generalizations are derived, when compared with the procedure used by earlier writers, makes it worthy of note.

**Acknowledgments.** The writer's interest in Lamé functions was first aroused when he was working on elliptic punch and crack problems [5] in elastostatics. At that time he was able to benefit from the advice and help of his colleague, Dr. R. S. Taylor, now with Her Majesty's Inspectorate of Schools. The full class of 24 representations of  $F_n^m(\alpha)$  given in § 3 first appeared in Dr. Taylor's Ph.D. thesis (University of Surrey, 1970), and it was the original intention that this paper should be co-authored by Dr. Taylor. However, the scope of the paper has widened considerably in the course of my investigations, and subsequently Dr. Taylor felt unable to appear as a co-author. I accepted this characteristically generous gesture with some reluctance, but would like to place on record my thanks to Dr. Taylor, and to emphasize his contribution to this work. I am also grateful to Professor F. M. Arscott for helpful comments on this paper.

#### REFERENCES

- [1] E. T. WHITTAKER, *On Lamé's equation and ellipsoidal harmonics*, Proc. London Math. Soc., 14 (1915), pp. 260–268.
- [2] A. ERDÉLYI, *Integral equations for Lamé functions*, Proc. Edinburgh Math. Soc., 7 (1942), pp. 3–15.
- [3] F. M. ARSCOTT, *Integral equations and relations for Lamé functions*, Quart. J. Math. Oxford Ser., 15 (1964), pp. 103–115.
- [4] B. D. SLEEMAN, *The expansion of Lamé functions into series of associated Legendre functions of the second kind*, Proc. Cambridge Phil. Soc., 62 (1966), pp. 441–452.
- [5] R. SHAIL, *Lamé polynomial solutions to some elliptic crack and punch problems*, Internat. J. Engrg. Sci., 16 (1978), pp. 551–563.
- [6] F. M. ARSCOTT, *Periodic Differential Equations*, Pergamon Press, Oxford, England 1964.
- [7] J. LIOUVILLE, *Lettres sur diverses questions*, J. Math. Pures Appl., 11 (1846), pp. 217–236.
- [8] B. D. SLEEMAN, *Integral equations and relations for Lamé functions and ellipsoidal wave functions*, Proc. Cambridge Philos. Soc., 64 (1968), pp. 113–126.
- [9] E. T. COPSON, *Theory of Functions of a Complex Variable*, Oxford University Press, Oxford, England, 1960.



- [10] F. M. ARSCOTT AND I. M. KHABAZA, *Tables of Lamé Polynomials*, Pergamon Press, Oxford, England 1962.
- [11] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Scientists*, Springer-Verlag, Berlin, 1971.
- [12] B. D. SLEEMAN, *The low-frequency scalar Dirichlet scattering by a general ellipsoid*, J. Inst. Maths Appl. 3 (1967), pp. 291–312.
- [13] G. I. A. POPOV, *The contact problem of the theory of elasticity for the case of a circular area of contact*, Prikl. Mat. Mech, 26 (1963), pp. 207–225.
- [14] A. H. ENGLAND AND R. SHAIL, *Orthogonal polynomial solutions to some mixed boundary-value problems in elasticity, II*, Quart. J. Mech. Appl. Math. 30 (1977), pp. 397–414.
- [15] P. M. MORSE AND H. FESHBACH, *Methods of Theoretical Physics.*, McGraw-Hill, New York, 1953.
- [16] E. T. WHITTAKER AND G. N. WATSON. *A Course of Modern Analysis*, Cambridge University Press, London, England, 1963.
- [17] B. A. HARGRAVE AND B. D. SLEEMAN, *Asymptotic evaluation of certain integral formulae for ellipsoidal wave functions*, Proc. Roy. Soc. Edinburgh, Sect. A, A72 (1973/74), pp. 257–269.

## AN APPLICATION OF SPLINE APPROXIMATION WITH VARIABLE KNOTS TO OPTIMAL ESTIMATION OF THE DERIVATIVE\*

CHARLES K. CHUI† AND PHILIP W. SMITH‡

**Abstract.** In studying the optimal  $L_p$  estimation of  $f'$  at  $x_0 = t_{n+k}$  from the data  $f(t_1), \dots, f(t_{n+k-1})$ , where  $t_1 < \dots < t_{n+k-1} = t_{n+k}$ , one naturally arrives at the problem of best  $L_q$  approximation of  $N_{n,k}$  from the span of  $N_{i,k}$ ,  $i = 1, \dots, n-1$ , where  $\{N_{i,k}\}_{i=1}^n$  are the normalized B-splines with the knot sequence  $\{t_i\}_{i=1}^{n+k}$  and  $1/p + 1/q = 1$ . We prove that the  $L_q$  error  $d_q(c(\mathbf{t})N_{n,k}, \text{Sp}\{N_{i,k}\}_{i=1}^{n-1})$ , where  $c(\mathbf{t}) = 1/(k-1)! \prod_{i=1}^{k-2} (1-t_{n+i})$ , is nonincreasing in  $\mathbf{t}$  for  $1 \leq q \leq \infty$ , and actually is decreasing for  $1 \leq q < \infty$ , as the knots  $t_1, \dots, t_{n+k-2}$  are moved to the right toward  $t_{n+k-1}$ . This agrees with the general conjecture of G. G. Lorentz, namely, "like best approximates like." Consequently, in approximating  $f'(x_0)$ , it is advisable to take the nodes  $t_1, \dots, t_{n+k-2}$  as close to  $x_0$  as possible as long as the process is stable. However, it is also shown that this conclusion is no longer valid if one approximates  $f''(x_0) - \alpha f'(x_0)$ , say for  $\alpha = 10$  and  $k = 3$ .

**1. Introduction.** In applications, it is often essential to approximate the derivative of a function  $f$  at some point say  $x_0$  from the data  $\{f(t_i)\}_{i=1}^{n+k-1}$ . For convenience, we consider  $0 \leq t_1 < t_2 < \dots < t_{n+k-1} = t_{n+k} := x_0 = 1$ . The reader will see that the location of  $x_0$  does not have to be restricted to be at  $t_{n+k-1}$ . We are interested in approximations of the following type:

$$(1.1) \quad f'(1) \sim \sum_{i=1}^{n+k-1} \gamma_i f(t_i).$$

This will be called a differentiation formula. Of course, to obtain an optimal differentiation formula, one must know certain properties of  $f$ , and the problem is, in general, very difficult. To simplify the problem, the approximation theorist usually considers optimality over a class of functions. In this paper, we consider the Sobolev space  $H_p^k$ ,  $1 \leq p \leq \infty$ , of functions on  $[0, 1]$ . For the problem to be meaningful, we require that  $k \geq 2$  and  $n \geq 1$ .

Let  $L = L(\boldsymbol{\gamma}) = L(\boldsymbol{\gamma}, \mathbf{t})$ , where  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{n+k-1})$  and  $\mathbf{t} = (t_1, \dots, t_{n+k-1})$ , be a continuous linear functional on  $H_p^k$  defined by  $Lf = f'(1) - \sum_{i=1}^{n+k-1} \gamma_i f(t_i)$ ,  $f \in H_p^k$ . Set  $|L| = \sup \{ |Lf| : f \in H_p^k, \|f^{(k)}\|_p \leq 1 \}$  and let  $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_{n+k-1}^*)$  be chosen such that  $|L(\boldsymbol{\gamma}^*)| = \inf \{ |L(\boldsymbol{\gamma})| : \boldsymbol{\gamma} \in \mathbb{R}^{n+k-1} \}$ . We will show that  $|L(\boldsymbol{\gamma}^*, \mathbf{t})|$  is decreasing in  $\mathbf{t}$ . This is achieved by recasting the above problem in terms of B-splines as follows. Let  $\{N_{i,k}\}_{i=1}^n$  be the normalized B-splines with the knot sequence  $\{t_i\}_{i=1}^{n+k}$  and let  $d_q(N_{n,k}, \text{Sp}\{N_{i,k}\}_{i=1}^{n-1})$  be the  $L_q$  distance from the span of  $\{N_{i,k}\}_{i=1}^{n-1}$  to  $N_{n,k}$ , where  $1/p + 1/q = 1$ . It is shown in the next section that

$$(1.2) \quad |L(\boldsymbol{\gamma}^*)| = c(\mathbf{t}) d_q(N_{n,k}, \text{Sp}\{N_{i,k}\}_{i=1}^{n-1}),$$

where  $c(\mathbf{t})$  is defined by

$$(1.3) \quad c(\mathbf{t}) = \frac{1}{(k-1)!} \prod_{i=1}^{k-2} (1-t_{n+i}).$$

\* Received by the editors October 4, 1978. This research was supported in part by the U.S. Army Research Office under Grants DAHC-04-75-G-0186 and DAAG 29-78-G-0097.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843.

Hence,  $c(\mathbf{t})$  is a decreasing function of  $(t_{n+1}, \dots, t_{n+k-2})$ . We will also show that  $c(\mathbf{t})d_q(N_{n,k}, \text{Sp}\{N_{i,k}\}_{i=1}^{n-1})$  is a nonincreasing function of  $(t_1, \dots, t_{n+k-2})$  for all  $1 \leq q \leq \infty$ , and, in fact, it is strictly decreasing for  $1 \leq q < \infty$ . The inequality will be considered in § 3, while the strict inequality will be proved in § 4. It is clear that, for  $q = \infty$ , strict inequality is not possible.

These results on approximation of  $N_{n,k}$  from the span of  $\{N_{i,k}\}_{i=1}^{n-1}$  are related to the problem of Lorentz on approximation of  $x^n$  from the span of  $\{x^{\lambda_1}, \dots, x^{\lambda_k}\}$  on  $[0, 1]$ . This polynomial result is proved in [4] and a generalization is considered in [11]. The more general result on Descartes systems is obtained by the second author in [12]. The difficulty in working with B-splines is that they do not form a Descartes system. To overcome this difficulty much additional analysis is necessary, and, in fact, strict inequality is not even possible for the  $L_\infty$  case.

We have now seen that in approximating  $f'(1)$  from the data  $\{f(t_i)\}_{i=1}^{n+k-1}$ , where  $t_1 < \dots < t_{n+k-1} = 1$ , it is advisable, in the sense discussed above, to take the nodes  $t_1, \dots, t_{n+k-1}$  as close to 1 as possible. In fact this phenomenon holds for a more general linear functional. In particular, it is trivially true for the approximation of  $f'(1) + \alpha f(1)$ ,  $\alpha$  constant. One might say that this result is intuitively obvious, since to obtain the information at a point, it would be advisable to take the data at the nodes as close to the given point of interest as possible. Surprisingly, this intuition is wrong in approximating, say,  $f''(1) - 10 f'(1)$ , when  $k = 3$ . This result will be discussed in § 5.

**2. Differentiation formula.** Let  $H_p^k$ ,  $1 \leq p < \infty$ , be the Sobolev space of functions which are  $k$ -fold integrals of functions in  $L_p[0, 1]$ , and let  $0 \leq t_1 < t_2 < \dots < t_{n+k-1} = t_{n+k} = 1$ . We are interested in approximating  $f'(1) = f'(t_{n+k})$  from the data  $\{f(t_i)\}_{i=1}^{n+k-1}$  via formulae of the type

$$f'(1) \sim \sum_{i=1}^{n+k-1} \gamma_i f(t_i).$$

For the problem to be meaningful, we will require that  $f \in H_p^k$  for  $k \geq 2$  and  $n \geq 1$ . Of course, the point  $t_{n+k} = 1$  is just chosen for convenience and could be replaced by any value to the left of 1 which is greater than zero.

The error

$$(2.1) \quad Lf := f'(1) - \sum_{i=1}^{n+k-1} \gamma_i f(t_i)$$

is a continuous linear functional on  $H_p^k$ . Since we will want to compare these differentiation formulae, we may from time to time use  $L(\boldsymbol{\gamma})$  or even  $L(\boldsymbol{\gamma}, \mathbf{t})$  in place of  $L$ , where  $\mathbf{t} := (t_1, \dots, t_{n+k-1})$  and  $\boldsymbol{\gamma} := (\gamma_1, \dots, \gamma_{n+k-1})$ , unless it is clear from the context which formulae  $L$  of type (2.1) we are denoting. Since any choice of  $\boldsymbol{\gamma}$  in (2.1) yields a differentiation formula, it is natural to ask which choice of  $\boldsymbol{\gamma}$  yields good or even a best differentiation formula. We choose to measure the effectiveness of these formulae by

$$|L| := \sup \{ |Lf| : f \in H_p^k, \|f^{(k)}\|_p \leq 1 \}.$$

Then a best formula given by  $\boldsymbol{\gamma}^*$  would satisfy

$$(2.2) \quad \eta = \inf \{ |L(\boldsymbol{\gamma})| : \boldsymbol{\gamma} \in \mathbb{R}^{n+k-1} \} = |L(\boldsymbol{\gamma}^*)|.$$

Throughout this paper, we will write  $\eta = |L(\boldsymbol{\gamma}^*)| = |L(\boldsymbol{\gamma}^*, \mathbf{t})|$ .

Since  $f^{(k)} = (f+p)^{(k)}$  for any  $p$  in  $\pi_k$ , the space of polynomials of degree less than  $k$ , it is clear that any  $L$  satisfying  $|L| < \infty$  must annihilate  $\pi_k$  and we write  $L \in \pi_k^\perp$ . Hence, we

may apply the Peano kernel theorem to obtain

$$(2.3) \quad \eta = \inf_{L \in \pi_k^\perp} \sup \left\{ \int_0^1 K_L g: g \in L_p[0, 1] \text{ and } \|g\|_p \leq 1 \right\} = \inf_{L \in \pi_k} \|K_L\|_q,$$

where  $K_L$  is the appropriate Peano kernel for  $L$ , e.g.,

$$(2.4) \quad K_L(t) = L \left[ \frac{(\cdot - t)_+^{k-1}}{(k-1)!} \right]$$

and  $1/p + 1/q = 1$ .

Throughout the rest of the paper, we denote by  $\{N_{i,k}: i = 1, \dots, n\}$ , the set of normalized B-splines [1] with knot sequence  $t = (t_1, \dots, t_{n+k})$ . The following lemma relates  $K_L$  to the span of the  $N_{i,k}$ 's.

LEMMA 2.1. *Let*

$$(2.5) \quad A = \left\{ N_{n,k} - \sum_{i=1}^{n-1} c_i N_{i,k}: c_i \text{ real} \right\}$$

and

$$(2.6) \quad B = \{K_L(t): L = L(\gamma) \in \pi_k^\perp\}.$$

Then

$$(2.7) \quad B = \left[ \frac{1}{(k-1)!} \prod_{i=1}^{k-2} (1 - t_{n+i}) \right] A.$$

*Proof.* For each  $b \in B$ , we have

$$b(t) = K_{L(\gamma)}(t) = \frac{(1-t)_+^{k-2}}{(k-2)!} - \sum_{i=1}^{n+k-1} \frac{\gamma_i}{(k-1)!} (t_i - t)_+^{k-1},$$

so that (recalling the definition of  $N_{n,k}$ )

$$b = \left[ \frac{1}{(k-1)!} \prod_{i=1}^{k-2} (1 - t_{n+i}) \right] a + p,$$

where  $a \in A$  and  $p \in \pi_k$ . Since both  $a$  and  $b$  vanish on  $(1, \infty)$ ,  $p$  must be identically zero. Conversely, any element in the right side of (2.7) can be written as

$$g(t) = \frac{(1-t)_+^{k-2}}{(k-2)!} - \sum_{i=1}^{n+k-1} \frac{d_i}{(k-1)!} (t_i - t)_+^{k-1}$$

with  $g(t) = 0$  for  $t \leq t_i$ . Integration by parts now allows us to conclude that  $g(t) = K_{L(\mathbf{d})}(t)$ , where  $\mathbf{d} = (d_1, \dots, d_{n+k-1})$  and  $L(\mathbf{d}) \in \pi_k^\perp$ . This completes the proof of the lemma.

From this lemma we now conclude that the infimum in (2.2) is attained and

$$(2.8) \quad \eta = \inf \{c(\mathbf{t})\|a\|_q: a \in A\},$$

where

$$(2.9) \quad c(\mathbf{t}) = \frac{1}{(k-1)!} \prod_{i=1}^{k-2} (1 - t_{n+i}).$$

That is, the minimization problem (2.2) reduces to the problem of best approximating  $N_{n,k}$  from the subspace spanned by  $\{N_{i,k}\}_{i=1}^{n-1}$  in  $L_q[0, 1]$ . It is, of course, a classical theorem that for  $1 < q < \infty$  the solutions are unique.

All the above facts have been known for quite a while. We are interested in comparing best differentiation formulae based on different interpolation points (nodes)  $\mathbf{t}$ . Specifically, let  $\mathbf{t}^i := 0 \leq t_1^i < \dots < t_{n+k-1}^i = t_{n+k}^i = 1, i = 1, 2$ , be two sets of points, and let  $L(\boldsymbol{\gamma}^{*i}, \mathbf{t}^i)$  correspond to the error functionals for best differentiation formulae as defined above for some  $p$ . In this paper, we will show that

$$(2.10) \quad |L(\boldsymbol{\gamma}^{*1}, \mathbf{t}^1)| \leq |L(\boldsymbol{\gamma}^{*2}, \mathbf{t}^2)|$$

whenever  $t_j^1 \geq t_j^2$  for  $j = 1, \dots, n+k$ ; and, for  $1 \leq p < \infty$ , equality holds only if  $\mathbf{t}^1 = \mathbf{t}^2$ . This result verifies the ‘‘obvious fact’’ that one obtains a better differentiation formula by using information nearer the point of interest (in this case  $t_{n+k} = 1$ ). The statement and proof of (2.10) is analogous to the problem of approximating  $x^n$  from span  $\{x^{\lambda_1}, \dots, x^{\lambda_k}\}$  on  $[0, 1]$  where one asks for the best  $\lambda_i$ ’s given that  $0 \leq \lambda_1 < \dots < \lambda_k < n, \lambda_i$  integers. Lorentz [8] proposed this problem in  $L_\infty[0, 1]$  and conjectured the correct answer ( $\lambda_i^* = N - k - 1 + i$ ) as was subsequently proved in [4].

**3. Comparison of best differentiation formulae.** A set of functions  $\{\varphi_i\}_{i=1}^n \subset C(a, b)$  will be called totally positive (see Karlin [7]) on the interval  $(a, b)$  provided

$$(3.1) \quad \det \begin{bmatrix} \varphi_{\lambda_1}(t_1) & \varphi_{\lambda_2}(t_1) & \dots & \varphi_{\lambda_k}(t_1) \\ \vdots & \vdots & & \vdots \\ \varphi_{\lambda_1}(t_k) & \varphi_{\lambda_2}(t_k) & \dots & \varphi_{\lambda_k}(t_k) \end{bmatrix} \geq 0,$$

whenever  $1 \leq \lambda_1 < \dots < \lambda_k \leq n, a < t_1 < \dots < t_k < b$ , and  $1 \leq k \leq n$ . Let  $\{\varphi_i\}_{i=1}^n \subset C(a, b)$  be totally positive. In view of the discussion in § 2, we will be interested in approximating  $\varphi_n$  by subspaces spanned by subsets of  $\{\varphi_i\}_{i=1}^{n-1}$ . Let  $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_k)$  be given where  $1 \leq \lambda_1 < \lambda_2 < \dots < \lambda_k < n$  and set  $S(\boldsymbol{\lambda}) = \text{span} \{\varphi_{\lambda_i}\}_{i=1}^k$ . We define, for  $1 \leq q \leq \infty$ ,

$$\text{dist}_q(\varphi_n, S(\boldsymbol{\lambda})) = \inf \{ \|\varphi_n - s\|_{L_q(a,b)} : s \in S(\boldsymbol{\lambda}) \}.$$

The following lemma is a direct consequence of the proof of Theorem 3 in [12] and a standard smoothing technique (cf. [10]).

LEMMA 3.1. *Let  $\{\varphi_i\}_{i=1}^n \subset L_q(a, b)$  be a totally positive system on  $(a, b)$ . If  $\boldsymbol{\lambda}$  and  $\boldsymbol{\gamma}$  are  $k$ -tuples as above with  $1 \leq \gamma_i \leq \lambda_i < n$ , and if  $1 \leq q \leq \infty$  then*

$$(3.2) \quad \text{dist}_q(\varphi_n, S(\boldsymbol{\lambda})) \leq \text{dist}_q(\varphi_n, S(\boldsymbol{\gamma})).$$

Since we will need to use the smoothing technique several times, we sketch its use here. For any  $\delta > 0$  define

$$(3.3) \quad \varphi_i(x, \delta) = (1/\delta\sqrt{2\pi}) \int_a^b \varphi_i(t) \exp(-(x-t)^2/2\delta^2) dt.$$

The functions  $\{\varphi_i(\cdot, \delta)\}_{i=1}^n$  converge in  $L_q[a, b]$  (and locally uniformly) to  $\{\varphi_i(\cdot)\}_{i=1}^n$  as  $\delta \rightarrow 0^+$ . Furthermore, for each  $\delta > 0$ , the set of functions  $\{\varphi_i(\cdot, \delta)\}_{i=1}^n$  is an extended totally positive system (see Karlin [7]) and in particular a Descartes system. Following the proof of Theorem 3 of [12] we see that (3.2) is valid for Descartes systems and, hence, (3.2) holds by letting  $\delta \rightarrow 0^+$ . This smoothing technique has been used by Micchelli under similar circumstances in [10].

We now present a lemma which will shortly allow us to make pairwise comparisons of certain differentiation formulae.

LEMMA 3.2. Let  $0 \leq t_1 < t_2 < \dots < t_{m+k-1} = t_{m+k} = 1$  and  $1 \leq r < r+1 \leq m+k-2$ . Then there exist  $k$ -th order B-splines  $\hat{N}_1, \dots, \hat{N}_m$  satisfying:

- (i)  $\text{span} \{\hat{N}_i\}_{i=1}^m = \text{span} \{N_{i,k}\}_{i=1}^m$ ,
- (ii) for  $i = r$  and  $r+1$ ,  $t_i$  is a knot of  $\hat{N}_i$  and  $\hat{N}_{i+1}$  respectively, for some index  $i_r$ , but is not a knot of any other  $\hat{N}_j$ , and
- (iii)  $\{\hat{N}_1, \dots, \hat{N}_m\}$  is a totally positive system.

*Proof.* Our proof of the lemma is constructive. The construction of the  $\hat{N}_i$ 's is a little different when  $t_r$  and  $t_{r+1}$  are close to the final or the initial knots. We therefore divide the proof of the lemma into three cases.

Case 1. Suppose  $r > k$ . Set

$$\begin{aligned}
 \hat{N}_i &= N_{i,k} \quad \text{if } 1 \leq i \leq r-k \text{ or } r+2 \leq i \leq m, \\
 \hat{N}_i(t) &= (t_{i+k} - t_i)[t_{r-k}, \dots, t_{r-1}, t_{i+k}]_s (s-t)_+^{k-1} \\
 &\quad \text{if } r-k < i \leq \min(r+1, m-1), \\
 \hat{N}_m(t) &= (1 - t_{r-k+1})[t_{r-k+1}, \dots, t_{r-1}, 1]_s (s-t)_+^{k-1} \\
 &\quad \text{if } m < r+2.
 \end{aligned}
 \tag{3.4}$$

It is clear that for  $i = r$  and  $r+1$ ,  $t_i$  is a knot of  $\hat{N}_{i-k}$  but of no other  $\hat{N}_j$ ,  $j \neq i-k$ . This gives condition (ii). Since  $\{N_{i,k}\}_{i=1}^m$  is a totally positive system (cf. [3, 7]), in order to verify (i) and (iii), it is sufficient to prove that there exist positive constants  $\gamma_i$ 's, such that, for  $i = r-k+1, \dots, r+1$ ,

$$N_{i,k} = \gamma_i \left( \hat{N}_i + \sum_{j=r-k}^{i-1} c_{j,i} \hat{N}_j \right)
 \tag{3.5}$$

for some constants  $c_{j,i}$ . We only prove (3.5) for  $r+2 \leq m$ . The case  $r+2 > m$  follows similarly. Note that if  $r+2 \leq m$ , the supports of  $\hat{N}_i$ , where  $r-k+1 \leq i \leq r+1$ , and  $\hat{N}_{r-k}$  are contained in the interval  $[t_{r-k}, 1]$  for some constants  $c_{j,i}$ . To prove (3.5), we note that the supports of  $\hat{N}_i$ ,  $r-k+1 \leq i \leq r+1$ , and  $\hat{N}_{r-k}$  are contained in the interval  $[t_{r-k}, 1]$ . Hence, by choosing

$$d_{i,1} = \lim_{t \rightarrow t_{r-k}^+} \hat{N}_i^{(k-1)}(t) / \hat{N}_{r-k}^{(k-1)}(t),$$

we have

$$(\hat{N}_i - d_{i,1} \hat{N}_{r-k})^{(i)}(t_{r-k}) = 0$$

for  $j = 0, \dots, k-1$ . Since  $\hat{N}_i - d_{i,1} \hat{N}_{r-k}$  is a polynomial of degree  $\leq k-1$  on  $[t_{r-k}, t_{r-k+1}]$ , it must be identically zero there. This shows that  $(\hat{N}_i - d_{i,1} \hat{N}_{r-k})$  is a  $k$ th order spline with (possible) knots at  $t_{r-k+1}, \dots, t_r, t_{i+k}$  and has support in  $[t_{r-k+1}, t_{i+k}]$ . Since this support consists of a minimal number of knots, it must be a constant multiple of the normalized B-spline with the same knots. But on  $[t_{i+k-1}, t_{i+k}]$  it is equal to  $\hat{N}_i$  so that it is positive there. Hence, we have

$$\hat{N}_i - d_{i,1} \hat{N}_{r-k} = \gamma_{i,1} [t_{r-k+1}, \dots, t_r, t_{i+k}]_s (s-t)_+^{k-1}
 \tag{3.6}$$

with  $\gamma_{i,1} > 0$ . Similarly, there is a positive constant  $\gamma_{i,2}$ , such that

$$\hat{N}_i - d_{i,2} [\hat{N}_{r-k+1} - d_{i,1} \hat{N}_{r-k}] = \gamma_{i,2} [t_{r-k+2}, \dots, t_{r+1}, t_{i+k}]_s (s-t)_+^{k-1}.
 \tag{3.7}$$

This process can be repeated until (3.5) is obtained.

Case 2. Suppose next that  $1 < r \leq k$ . In this case, we define our  $\hat{N}_i$ 's as in the following:

$$\begin{aligned}\hat{N}_1(t) &= (t_{k+2} - t_1)[t_1, \dots, t_r, t_{r+2}, \dots, t_{k+2}]_s (s-t)_+^{k-1}, \\ \hat{N}_2(t) &= (t_{k+2} - t_1)[t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2}]_s (s-t)_+^{k-1}, \\ \hat{N}_3(t) &= (t_{k+3} - t_1)[t_1, \dots, t_{r-1}, t_{r+2}, \dots, t_{k+3}]_s (s-t)_+^{k-1}, \\ &\vdots \\ \hat{N}_{r+1}(t) &= (t_{r+k-1} - t_1)[t_1, \dots, t_{r-1}, t_{2r}, \dots, t_{r+k+1}]_s (s-t)_+^{k-1}, \\ \hat{N}_{r+2}(t) &= (t_{r+k+2} - t_{r+2})[t_{r+2}, \dots, t_{r+k+2}]_s (s-t)_+^{k-1} = N_{r+2,k}(t), \\ &\vdots \\ \hat{N}_m(t) &= N_{m,k}(t).\end{aligned}$$

From the above construction, it is clear that (ii) is satisfied. Indeed,  $t_r$  is a knot of  $\hat{N}_1$  and  $t_{r+1}$  is a knot of  $\hat{N}_2$ , but they are not knots of any other  $\hat{N}_i$ . Following the argument in Case 1, to verify (i) and (iii) it is sufficient to prove that (3.5) is satisfied for some positive constants  $\gamma_i$ ,  $1 \leq i \leq r+1$ . The procedure in Case 1 applies; that is, we should verify (3.6), (3.7), etc. Note that the only difficulty is to verify (3.6) for  $\hat{N}_1$  and  $\hat{N}_2$  since  $\text{supp } \hat{N}_2 \subset \dots \subset \text{supp } \hat{N}_{r+1}$ . That is, we have to prove that there exists a  $\gamma > 0$  such that

$$(3.8) \quad (\hat{N}_2 - c_1 \hat{N}_1)(t) = \gamma [t_2, \dots, t_{k+2}]_s (s-t)_+^{k-1}.$$

Let  $V(\dots)$  denote the Vandermonde determinant. By Cramer's rule, it is clear that for  $t_{k+1} < t < t_{k+2}$ , we have

$$\hat{N}_2(t) = (t_{k+2} - t_1)(t_{k+2} - t)^{k-1} \frac{V(t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+1})}{V(t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2})}$$

and

$$\hat{N}_1(t) = (t_{k+2} - t_1)(t_{k+2} - t)^{k-1} \frac{V(t_1, \dots, t_r, t_{r+2}, \dots, t_{k+1})}{V(t_1, \dots, t_r, t_{r+2}, \dots, t_{k+2})}.$$

Hence, again for  $t_{k+1} < t < t_{k+2}$ , we have

$$(3.9) \quad \frac{\hat{N}_2(t)}{\hat{N}_1(t)} = \frac{t_{k+2} - t_r}{t_{k+2} - t_{r+1}} > 1.$$

For  $t_1 < t < t_2$ , we use the identity

$$(s-t)_+^{k-1} = (s-t)^{k-1} + (-1)^k (t-s)_+^{k-1}$$

to conclude that, for  $t_1 < t < t_2$ ,

$$\begin{aligned}\hat{N}_2(t) &= (t_{k+2} - t_1)[t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2}]_s \{(s-t)^{k-1} + (-1)^k (t-s)_+^{k-1}\} \\ &= (-1)^k (t_{k+2} - t_1)[t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2}]_s (t-s)_+^{k-1} \\ &= (-1)^{2k} (t_{k+2} - t_1)(t-t_1)^{k-1} \frac{V(t_2, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2})}{V(t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2})},\end{aligned}$$

where if  $r=2$ ,  $V(t_2, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2})$  means  $V(t_3, \dots, t_{k+2})$  and  $V(t_1, \dots, t_{r-1}, t_{r+1}, \dots, t_{k+2})$  means  $V(t_1, t_3, \dots, t_{k+2})$  etc., and

$$\begin{aligned}\hat{N}_1(t) &= (t_{k+2} - t_1)[t_1, \dots, t_r, t_{r+2}, \dots, t_{k+2}]_s \{(s-t)^{k-1} + (-1)^k (t-s)_+^{k-1}\} \\ &= (-1)^{2k} (t_{k+2} - t_1)(t-t_1)^{k-1} \frac{V(t_2, \dots, t_r, t_{r+2}, \dots, t_{k+2})}{V(t_1, \dots, t_r, t_{r+2}, \dots, t_{k+2})}.\end{aligned}$$

Hence, for  $t_1 < t < t_2$ , we have

$$(3.10) \quad \frac{\hat{N}_2(t)}{\hat{N}_1(t)} = \frac{t_r - t_1}{t_{r+1} - t_1} < 1.$$

That is, we have

$$(\hat{N}_2 - c_1 \hat{N}_1)(t) = 0$$

for all  $t \in (t_1, t_2)$ , where  $c_1 = (t_r - t_1)/(t_{r+1} - t_1)$  is a constant between 0 and 1. By using (3.9), we have  $(\hat{N}_2 - c_1 \hat{N}_1)(t) > 0$  for all  $t \in (t_{k+1}, t_{k+2})$ . But  $\hat{N}_2 - c_1 \hat{N}_1$  has minimal support and is, therefore, a positive multiple of the normalized B-spline with the same knots. That is, we have (3.8) with  $\gamma > 0$ .

Case 3. Suppose finally  $r = 1$ . We then define

$$\hat{N}_1(t) = (t_{k+2} - t_1)[t_1, t_3, \dots, t_{k+2}]_s (s - t)_+^{k-1},$$

and  $\hat{N}_i = N_{i,k}$  for  $i = 2, \dots, n$ . Condition (ii) is clearly satisfied, and arguing in the same manner as in Case 1, we also have conditions (i) and (iii). This completes the proof of the lemma.

Now let  $\mathbf{t}^i := 0 \leq t_1^i < \dots < t_{n+k-1}^i = t_{n+k}^i = 1$  for  $i = 1, 2$  be two sets of points satisfying

$$(3.11) \quad \begin{aligned} (i) \quad & t_j^1 = t_j^2 \quad \text{if } j \neq r \quad \text{and} \\ (ii) \quad & t_r^1 > t_r^2. \end{aligned}$$

With the help of Lemmas 2.1, 3.1, and 3.2 we can deduce

LEMMA 3.3. Let  $\mathbf{t}^1$  and  $\mathbf{t}^2$  be as above satisfying (3.10). Further let  $L(\boldsymbol{\gamma}^{*1}, \mathbf{t}^1)$  and  $L(\boldsymbol{\gamma}^{*2}, \mathbf{t}^2)$  correspond to best  $L_p$  differentiation formulae with  $1 \leq p \leq \infty$ . Then

$$(3.12) \quad |L(\boldsymbol{\gamma}^{*1}, \mathbf{t}^1)| \leq |L(\boldsymbol{\gamma}^{*2}, \mathbf{t}^2)|.$$

*Proof.* Lemma 2.1 reduces the problem to comparing  $d_q(N_{n,k}^1, S^1)$  with  $d_q(N_{n,k}^2, S^2)$ , where the superscripts reference the knot sequence used and  $S^i = \text{span}\{N_{j,k}^i\}_{j=1}^{n-1}$ . First note that if  $r < n$  then  $N_{n,k}^1 = N_{n,k}^2$  and using Lemmas 3.1 and 3.2 the result (3.12) is immediate. If  $n \leq r \leq n + k - 2$ , then the result still follows since we may combine the two knot sequences to form

$$\begin{aligned} \mathbf{t}^* &:= 0 \leq t_1^* < \dots < t_{n+k}^* = t_{n+k+1}^* = 1 \\ &:= 0 \leq t_1^2 < \dots < t_r^2 < t_r^1 < \dots < t_{n+k-1}^1 = t_{n+k}^1 = 1. \end{aligned}$$

Let  $\{\hat{N}_i\}_{i=1}^{n+1}$  be the B-splines constructed in Lemma 3.2 on the knot sequence  $\mathbf{t}^*$  with  $m = n + 1$  and  $t_r^2$  only contributing to  $\hat{N}_{i_r}$ , and  $t_r^1$  only contributing to  $\hat{N}_{i_r+1}$  with  $i_r + 1 < n + 1$ , and let  $\hat{S}_i = \text{span}\{\hat{N}_j\}_{j=1, j \neq i_r+1}^n$ . Then since  $c(\mathbf{t}^i)(N_{n,k}^i)^{(k-2)}(1^-)$  is 1, we have

$$c(\mathbf{t}^i)d_q(N_{n,k}^i, S^i) = \frac{1}{\hat{N}_{n+1}^{(k-2)}(1^-)} d_q(\hat{N}_{n+1}, \hat{S}_i)$$

for  $i = 1, 2$ . But Lemmas 3.1 and 3.2 imply that  $d_q(\hat{N}_{n+1}, \hat{S}_1) \leq d_q(\hat{N}_{n+1}, \hat{S}_2)$ . Hence, recalling Lemma 2.1, we obtain (3.12). This brings us to the main result of this section.

THEOREM 3.4. Let  $1 \leq p \leq \infty$  be fixed, and suppose that two best differentiation formulae  $L(\boldsymbol{\gamma}^{*1}, \mathbf{t}^1)$  and  $L(\boldsymbol{\gamma}^{*2}, \mathbf{t}^2)$  are given with

$$(3.13) \quad \begin{aligned} (i) \quad & 0 \leq t_1^i < \dots < t_{n+k-1}^i = t_{n+k}^i = 1, \quad i = 1, 2, \quad \text{and} \\ (ii) \quad & t_j^1 \geq t_j^2, \quad j = 1, \dots, n + k - 2. \end{aligned}$$

Then  $|L(\boldsymbol{\gamma}^{*1}, \mathbf{t}^1)| \leq |L(\boldsymbol{\gamma}^{*2}, \mathbf{t}^2)|$ .



This theorem is easily proved by making pairwise comparisons and applying Lemma 3.3. A similar argument is made in [12]. In the next section, we will consider the problem of proving strict inequality.

**4. Strict inequality.** In this section, we strengthen the conclusion of Theorem 3.4. In particular, we will prove that for  $1 \leq q < \infty$  (or  $1 < p \leq \infty$ ) the inequalities in Theorem 3.4 may be replaced by strict inequalities provided  $\mathbf{t}^1 \neq \mathbf{t}^2$ . It is clear that such a conclusion cannot hold for  $q = \infty$ . The analysis here is parallel to that in [12], where similar results are obtained for Descartes systems. We begin by presenting a few basic facts about totally positive systems.

LEMMA 4.1. *Let  $\{\varphi_i\}_{i=1}^{n+1} \subset C(a, b)$  be a totally positive system on  $(a, b)$ . Suppose that  $a < \tau_1 < \dots < \tau_n < b$  exist so that*

$$\det [\varphi_j(\tau_i)]_{\substack{i=1, \dots, n \\ j=1, \dots, n}} > 0,$$

and  $\varphi_{n+1}(\tau_j) - \sum_{i=1}^n a_i \varphi_i(\tau_j) = 0$ ,  $j = 1, \dots, n$ . Then for  $\tau \in (\tau_i, \tau_{i+1})$ ,  $i = 0, \dots, n$ ,

$$(-1)^{n-i} \left[ \left( \varphi_{n+1} - \sum_{i=1}^n a_i \varphi_i \right) (\tau) \right] \geq 0$$

where  $\tau_0 = a$  and  $\tau_{n+1} = b$ . Furthermore,  $(-1)^{n-i} a_i \geq 0$ , for  $i = 1, \dots, n$ .

These results are immediate if one uses smoothing (cf. (3.3)) and recalls Descartes' rule of sign. We now show that, for  $1 \leq q < \infty$ , the best  $L_q$  approximation by totally positive systems quite often produces an error functional whose sign structure is comparable to that found in approximation by Descartes systems.

LEMMA 4.2. *Let  $1 \leq q < \infty$ ,  $\mathbf{t} := 0 \leq t_1 < \dots < t_{m+k-1} = t_{m+k} = 1$  and  $\{N_{i,k}\}_{i=1}^m$  be the corresponding normalized B-splines. Set  $S = \text{span} \{N_{i,k}\}_{i=1}^m$ , and suppose that  $s_q \in S$  is the best  $L_q(t_1, 1)$  approximation to  $N_{m,k}$  from  $S$ . If the Lebesgue measure of the set  $\{x \in (t_1, 1) : N_{m,k}(x) - s_q(x) := e_q(x) = 0\}$  is zero, then there exist  $\tau_1 < \dots < \tau_{m-1}$  satisfying*

$$(4.1) \quad \begin{array}{ll} \text{(i)} & e_q \text{ changes sign at each of the } \tau_i \text{'s, and} \\ \text{(ii)} & \det [N_{i,k}(\tau_j)]_{\substack{i=1, \dots, m-1 \\ j=1, \dots, m-1}} > 0. \end{array}$$

*Proof.* For the case  $q = 1$ , the reader should consult Micchelli [10]. For  $1 < q < \infty$ , we note that by smoothing there exist at most  $j$  ( $j < m$ ) points  $x_1 < \dots < x_j$  at which  $e_q$  changes sign. Again, by smoothing we can produce an  $s \in S$  which, if  $j < m - 1$ , is not identically zero and has exactly the same (weak) sign structure of  $e_q$ . But this means that

$$(4.2) \quad \int_{t_1}^1 (|e_q|^{q-1} \text{sign } e_q) s \neq 0,$$

contradicting the assumption that  $s_q$  is the best  $L_q$  approximation to  $\varphi_{n+1}$  on  $(t_1, 1)$ . Note that it is here that we use the fact that  $\{x : e_q(x) = 0\} = 0$ . Thus, we have the existence of the  $\{\tau_i\}_{i=1}^{m-1}$  satisfying (4.1, i). As for (4.1, ii) the arguments in Micchelli [10, Thm. 2] yield the desired conclusion.

We mention, for future reference, that the  $\{\tau_i\}_{i=1}^{m-1}$  depend continuously on the knot sequence  $\mathbf{t}$ . For  $1 < q < \infty$ , this is clear since the best approximation deform continuously as a function of  $\mathbf{t}$ . For  $q = 1$ , this can be deduced from the work of Micchelli [10].

In order to be able to use Lemma 4.2, we need to know that the error  $e_q$  never vanishes on a set of positive measure. This fact is contained in the next lemma.

LEMMA 4.3. *Let  $\mathbf{t} := 0 \leq t_1 < \dots < t_{m+k-1} = t_{m+k} = 1$  be given and let  $\{N_{i,k}\}_{i=1}^m$  be the corresponding normalized B-splines. Set  $S = \text{span} \{N_{i,k}\}_{i=1}^m$ , and let  $s_q \in S$  be the best*

$L_q(t_1, 1)$  approximation to  $N_{m,k}$  from  $S$ , where  $1 \leq q < \infty$ . Then  $e_q := N_{m,k} - s_q$  does not vanish on a set of positive measure in  $(t_1, 1)$ .

*Proof.* Suppose, on the contrary, that  $e_q$  vanished on a set of positive measure. Then, by analyticity,  $e_q$  must vanish on some interval  $(t_{j-1}, t_j)$ . Due to the linear independence and the property of the supports on the B-splines, we must have  $j \leq m$  and the coefficients of  $N_{j-k,k}, \dots, N_{j-1,k}$  in  $s_q$  must be zero. Since  $q < \infty$ , the best approximant  $s_q$  of  $N_{m,k}$  is unique, so that  $s_q$  (and hence  $e_q$ ) is also zero on  $(t_1, t_{j-1})$ . Hence,  $e_q$  vanishes on  $(t_1, t_j)$ , and the above argument also shows that we may assume that  $e_q$  does not vanish on a set of positive measure on  $(t_j, 1)$ . Also, since

$$(4.3) \quad \int_{t_1}^1 (|e_q|^{q-1} \operatorname{sgn} e_q) N_{i,k} = 0$$

for  $i = 1, \dots, m - 1$ , and since  $e_q$  has no sign change on  $(t_j, t_{j+1})$ , we must have  $j < k$ . By relabeling, we may assume without loss of generality that  $j = 2$ . Now Lemma 4.2 tells us that there are exactly  $m - 2$  points  $\tau_1 < \dots < \tau_{m-2}$  over which the B-splines  $\{N_{2,k}, \dots, N_{m-1,k}\}$  are linearly independent and which are sign changes of  $e_q$ . Now it must be that  $N_{i+1,k}(\tau_i) \neq 0$  for  $i = 1, \dots, m - 2$ , which is a well-known condition for the invertibility of the spline collocation matrix [3], so that  $\tau_i \in (t_{i+1}, t_{i+1+k})$ . On the other hand,  $|e_q|^{q-1} \operatorname{sgn} e_q$  is orthogonal to  $N_{1,k}$  so that  $\tau_1 \in (t_1, t_{1+k})$ . Let  $r$  be the first integer  $\leq m - 1$  so that  $\tau_r \in (t_r, t_{r+k})$ . Then  $\tau_r \in [t_{r+k}, t_{r+k+1})$  and  $\{N_{i,k}\}_{i=1}^{r-1}$  are linearly independent over  $\{\tau_1, \dots, \tau_{r-1}\}$  (cf. [7]), and hence, one can produce a nontrivial linear combination of  $\{N_{i,k}\}_{i=1}^r$  which has the same weak sign structure as  $e_q$ , contradicting (4.3). If no such  $r$  exists, one can similarly construct a nontrivial linear combination of  $\{N_{i,k}\}_{i=1}^{m-1}$  which has the same (weak) sign structure as  $e_q$ , again contradicting (4.3).

We are now ready to state and prove the main result of this section: namely, for  $1 \leq q < \infty$  the inequalities in Theorem 3.4 can be replaced by strict inequalities.

**THEOREM 4.4.** *Let  $1 < p \leq \infty$  and suppose that two knot sequences  $\mathbf{t}^1$  and  $\mathbf{t}^2$  are given such that (3.12) holds. Then if  $\mathbf{t}^1 \neq \mathbf{t}^2$  we have  $|L(\gamma^{*1}, \mathbf{t}^1)| < |L(\gamma^{*2}, \mathbf{t}^2)|$  where the  $L$ 's are best  $L_p$  differentiation formulae.*

*Proof.* As in the case of Theorem 3.4, the general result is proved by making pairwise comparisons. Thus, it is sufficient to assume that  $\mathbf{t}^1$  and  $\mathbf{t}^2$  satisfy (3.10) and to show strict inequality in this case. Proceeding as in the proof of Lemma 3.3 we form the auxiliary knot sequence

$$(4.4) \quad \begin{aligned} \mathbf{t}^* &:= 0 \leq t_1^* < \dots < t_{n+k}^* = t_{n+k+1}^* = 1 \\ &:= 0 \leq t_1^2 < \dots < t_r^2 < t_r^1 < \dots < t_{n+k-1}^1 = t_{n+k}^1 = 1, \end{aligned}$$

and the B-splines  $\{\hat{N}_{i,j}\}_{i=1}^{n+1}$  constructed as in Lemma 3.2 with  $m = n + 1$  on the knot sequence  $\mathbf{t}^*$  with  $t_r^2$  only contributing to  $\hat{N}_{i,r}$ , and  $t_r^1$  only contributing to  $\hat{N}_{i,r+1}$ . We define the two subspaces  $\hat{S}_i = \operatorname{span} \{\hat{N}_{j,i}\}_{j=1, j \neq i, i-1}$ ,  $i = 1, 2$ , and denote by  $e_q^i$  the error in best  $L_q(t_1, 1)$  approximation of  $\hat{N}_{n+1}$  from  $\hat{S}_i$ . We need to show, recalling Lemma 2.1, that  $\|e_q^1\|_{L_q} < \|e_q^2\|_{L_q}$ . Lemma 4.3 tells us that  $e_q^2$  does not vanish on a set of positive measure in  $(t_1, 1)$ , and so by Lemma 4.2, there are  $\{\tau_i\}_{i=1}^{n-1}$  which are the unique (ordered) sign changes of  $e_q^2$ .

It may happen that  $\{\hat{N}_{j,i}\}_{j=1, j \neq i}$  are not linearly independent over the  $\{\tau_i\}_{i=1}^{n-1}$ , but we can assume that they are by moving  $t_r^2$  sufficiently close to  $t_r^1$  preserving the order and recalling the continuity of the  $\{\tau_i\}_{i=1}^{n-1}$  as a function of the knots. The comparisons will preserve the inequalities because of Theorem 3.4. Let  $u = \hat{N}_{n+1} - s$ , where  $s \in \hat{S}_1$  is uniquely chosen by the condition

$$(4.5) \quad u(\tau_i) = 0, \quad i = 1, \dots, n - 1.$$

This can be done because of the linear independence. Now consider the difference

$$(4.6) \quad u - e_q^2 := \sum_{j=1}^n \alpha_j \hat{N}_j.$$

This difference is not identically zero since  $e_q^2$  has the knot  $t_r^2$  which is active. Furthermore, the sign of  $\alpha_{i+1}$  is determined by Lemma 4.1 and condition (4.5). Now Lemma 4.1 implies that for  $\tau \in (\tau_i, \tau_{i+1})$ , we have

$$(4.7) \quad (-1)^{n-i}(u - e_q^2)(\tau) \geq 0, \quad i = 0, \dots, n - 1,$$

where  $\tau_0 = a$  and  $\tau_n = b$ . Thus, it follows, as in [12], that

$$(4.8) \quad |u(\tau)| \leq |e_p^2(\tau)|, \quad \tau \in [t_1, 1].$$

Since  $u - e_q^2$  is not identically zero, and hence is not zero on a set of positive measure, we conclude that

$$\|u\|_{L_q(t_1,1)} < \|e_q^2\|_{L_q(t_1,1)}$$

for  $1 \leq q < \infty$ . Finally, we obtain

$$\|e_q^1\|_{L_q(t_1,1)} \leq \|u\|_{L_q(t_1,1)} < \|e_q^2\|_{L_q(t_1,1)}$$

as desired.

**5. A counterexample.** In §§ 2 and 3, we have seen that in approximating  $f'(1)$  from the data  $\{f(t_i)\}_{i=1}^N, 0 \leq t_1 < \dots < t_N \leq t_{N+1} = 1$ , it is advisable to choose the nodes  $\{t_i\}$  to be as close to  $t_{N+1} = 1$  as possible. Intuitively, this seems quite reasonable since one usually expects to obtain better information when the data values are taken closer to the point of interest. In this section, we will show that, surprisingly, in the approximation of

$$f''(1) - 2af'(1),$$

$a$  real, from the data  $\{f(t_i)\}_{i=1}^N$ , it is sometimes better to stay away from  $t_{N+1} = 1$ .

Let  $N_1, N_2$ , and  $N_3$  be the normalized B-splines of third-order and with the knot sequence  $\{0, \frac{1}{2}, t, 1, 1, 1\}$  where  $\frac{1}{2} < t < 1$ . Consider the error formula

$$(5.1) \quad M_a f = f''(1) - 2af'(1) - \sum_{i=1}^4 \gamma_i f(t_i),$$

where  $t_1 = 0, t_2 = \frac{1}{2}, t_3 = t, t_4 = 1$ , and  $a, \gamma_i$ 's are real. As in § 2, we consider

$$|M_a| := \sup \{|M_a f| : f \in H_2^3 \text{ and } \|f^{(3)}\|_2 \leq 1\}$$

and

$$(5.2) \quad \eta(a; t) = \inf \{|M_a| : \gamma_1, \dots, \gamma_4 \text{ real}\}.$$

Then by a simple calculation using the ideas in § 2, we have

$$(5.3) \quad \eta(a; t) = \inf \{\|N_3 + (1 - a + at)N_2 - \gamma N_1\|_2 : \gamma \text{ real}\}.$$

The following result gives a formula for  $\eta(a; t)$  in terms of the location of the variable knot  $t$ .

PROPOSITION 5.1. For each real  $a$ ,

$$(5.4) \quad \eta(a; t) = \left\{ (1/5)(1-t) + (2/15)(1-a+at)(1-t)^2 + (1/15)(1-a+at)^2(t-1/2) + \frac{[t(1-t)^2 + (1/2)(1-a+at)(t^2+t/2+1/4)]^2}{15t(2t^2-5t+1/2)} \right\}^{1/2}.$$

For example, if we take  $a = 5$ , then the minimum value of  $\eta(a; t)$  is unique and is attained at  $t \doteq .85$ . Additionally one can check that  $\eta'(5; 1) > 0$ , showing  $t = 1$  not to be a minimum. This shows that in the differentiation formula (5.1) the node  $t$  should not be chosen to be too close to 1.

*Proof of Proposition 5.1.* By the standard calculus technique of finding local minima (with respect to  $\gamma$ ), it is clear that

$$(5.5) \quad \eta^2(a; t) = \int_0^1 N_3^2 + 2\beta \int_0^1 N_3 N_2 + \beta^2 \int_0^1 N_2^2 - \frac{(\int_0^1 N_3 N_1 + \int_0^1 N_2 N_1)^2}{\int_0^1 N_1^2},$$

where  $\beta = 1 - a + at$ . By some tedious computations, we obtain:

$$(5.6) \quad \begin{aligned} \int_0^1 N_3^2 &= \frac{1-t}{5}, \\ \int_0^1 N_3 N_2 &= \frac{1-t^2}{15}, \\ \int_0^1 N_2^2 &= \frac{2t+1}{30}, \\ \int_0^1 N_3 N_1 &= \frac{(1-t)^2}{15}, \\ \int_0^1 N_2 N_1 &= \frac{t^2+t/2+1/4}{30t}, \quad \text{and} \\ \int_0^1 N_1^2 &= \frac{-2t+5t-1/2}{15t}. \end{aligned}$$

Putting (5.6) into (5.5), we obtain (5.4).

**6. Final remarks.** When  $p = 2$ , a strictly Hilbert space argument can be given to obtain the results in § 2. Indeed, if  $\eta$  is as given in (2.2), and  $s_f$  is the natural spline of order  $2k$  interpolating  $f$  at  $t_1, \dots, t_{n+k-1}$ , then a result in ([6] Lemma 2.3) gives

$$\eta = \sup \{ |f'(1) - s'_f(1)| : f \in H_2^k, \|f^{(k)}\|_2 \leq 1 \}.$$

But then  $g := f - s_f$  is in  $H_2^k$  and interpolates the zero data at the nodes  $t_1, \dots, t_{n+k-1}$ . Let  $S^*$  be the natural spline of order  $2k$  with knots at  $t_1, \dots, t_{n+k}$  determined by the data  $S^*(1) = 1$  and  $S^*(t_1) = \dots = S^*(t_{n+k-1}) = 0$ . Then

$$(6.1) \quad \begin{aligned} \eta &= \sup \{ |g'(1)| : g \in H_2^k, \|g^{(k)}\|_2 \leq 1, g(t_1) = \dots = g(t_{n+k-1}) = 0 \} \\ &= \left( \inf \left\{ \frac{\|h^{(k)}\|_2}{|h'(1)|} : h \in H_2^{(k)}, h(t_1) = \dots = h(t_{n+k-1}) = 0 \right\} \right)^{-1} \\ &= \|S^{*(k)}\|_2^{-1}. \end{aligned}$$

(See [5]). The quantity  $\|S^{*(k)}\|_2$  can be studied in the following manner. Let  $g \in H_2^k$  satisfy  $g(t_1) = \dots = g(t_{n+k-1}) = 0$  and  $g'(1) = 1$ . Following de Boor [2], we have

$$(6.2) \quad \|S^{*(k)}\|_2 = \inf \left\{ \|f^{(k)}\|_2 : f \in H_2^k, \int_0^1 f^{(k)} N_{i,k} = \int_0^1 g^{(k)} N_{i,k} \text{ for } i = 1, \dots, n \right\}.$$

By the Peano kernel theorem, it is clear that

$$(6.3) \quad \int_0^1 g^{(k)} N_{i,k} = (k-1)! (t_{i+k} - t_i) [t_i, \dots, t_{i+k}] g$$

$$= \begin{cases} 0 & \text{if } 1 \leq i \leq n-1, \\ \frac{(k-1)!}{(1-t_{n+1}) \cdots (1-t_{n+k-2})} & \text{if } i = n. \end{cases}$$

Hence, using the notation in § 2, we have

$$S^{*(k)} = C \left( N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right),$$

for an appropriate constant  $C$ , where  $c_i^*$ 's are chosen such that

$$\left\| N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right\|_2 = \inf \{ \|a\|_2 : a \in A \}.$$

To find  $C$ , we note that

$$\int_0^1 S^{*(k)} N_{n,k} = C \int_0^1 \left( N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right) N_{n,k} = C \left\| N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right\|_2^2.$$

Hence, using (6.2) and (6.3), we have

$$C = \left\| N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right\|_2^{-2} \frac{(k-1)!}{(1-t_{n+1}) \cdots (1-t_{n+k-2})};$$

and substituting into (6.1) yields

$$\eta = \left\| N_{n,k} - \sum_{i=1}^{n-1} c_i^* N_{i,k} \right\|_2^{-1} C = c(\mathbf{t}) \inf \{ \|a\|_2 : a \in A \},$$

where  $c(\mathbf{t})$  is defined in (2.9). This is the result in (2.8) for the case  $p = 2$ . Thus, we obtain the interesting result that  $\|S^{*(k)}\|_2$  is an increasing function of  $\mathbf{t}$ .

There are many more interesting problems which we have not addressed in this paper. In particular, the study of linear functions other than differentiation is of interest, as is the location of optimal nodes on both sides of the support of the functional. The study of the above and related problems will be delayed to a later date.

#### REFERENCES

- [1] C. DE BOOR, *On calculating with B-splines*, J. Approximation Theory, 6 (1972), pp. 50–62.
- [2] ———, *On "best" interpolation*, *Ibid.*, 15 (1976), pp. 28–42.
- [3] ———, *Total positivity of the spline collocation matrix*, Indiana Univ. J. Math., 25 (1976), pp. 541–551.
- [4] I. BOROSH, C. K. CHUI AND P. W. SMITH, *Best uniform approximation from a collection of subspaces*, Math. Z., 156 (1977), pp. 13–18.
- [5] C. K. CHUI, P. W. SMITH AND J. D. WARD, *Comparing digital filters which produce derivative approximations*, Proceeding of the 1978 Army Numerical Analysis and Computers conference, pp. 111–116.
- [6] M. COLOMB, *Interpolation operators as optimal recovery schemes for classes of analytic functions*, Optimal Estimation in Approximation Theory, C. A. Micchelli and T. J. Rivlin, eds., Plenum Press, New York, 1977, pp. 93–138.
- [7] S. KARLIN, *Total Positivity*, Vol. 1, Stanford University Press, Stanford, CA, 1968.
- [8] G. G. LORENTZ, *Approximation by incomplete polynomials (problems and results)*, Pade' and Rational Approximation, E. B. Saff and R. S. Varga, eds., Academic Press, New York, 1977, pp. 289–302.

- [9] ———, *Problems in Approximation Theory*, Univ. of Arkansas Lecture Notes in Mathematics, Vol. 1, Fayetteville, AR, 1977.
- [10] C. A. MICCHELLI, *Best  $L^1$  approximation by weak Chebyshev systems and uniqueness of interpolating perfect splines*, *J. Approximation Theory*, 19 (1977), pp. 1–14.
- [11] O. SHISHA, *Tchebysheff systems and best partial bases*, in manuscript.
- [12] P. W. SMITH, *An improvement theorem for Descartes systems*, *Proc. Amer. Math. Soc.*, 70 (1978), pp. 26–30.

## OPERATOR MEASURES, SELF-ADJOINT OPERATORS AND DYNAMICAL SYSTEMS\*

PAUL A. FUHRMANN†

**Abstract.** The paper studies multiplication operators in  $L^2$ -spaces of matrix measures as models for self-adjoint operators of finite multiplicity. The module theoretic aspects are emphasized and an analysis of intertwining maps, that is, module homomorphisms relative to the algebra of multiplication operators by bounded Borel functions, is given. Finally the machinery is applied to the study of dynamical systems with self-adjoint generators. Controllability aspects are studied and a version of the state space isomorphism theorem is derived.

**1. Introduction.** In [4] R. W. Brockett and the author studied the theory of symmetric dynamical systems with normal generators. The main points of the study were questions of spectral minimality, isomorphism theorems and realizability criterias. At least as far as the state space isomorphism proved there, the canonical spectral representation turned out to be a useful tool. The treatment had a somewhat ad-hoc flavor and the relations and similarities with the existing body of theory of linear time invariant dynamical systems remained somewhat obscure. In retrospect it seems the missing ingredient was the lack of stressing of the underlying algebraic ideas.

It was Kalman [17], [18] who first emphasized the use of modules as the natural framework within which to develop the linear theory. By now a significant part of the new results are obtained within that framework [10], [13], [23].

Lately it has been realized that much of the recent work on infinite dimensional systems has been developed essentially on the line of module theory, with the obvious adjustments needed to make it work in the analytic case [2], [11]. The systems under consideration used restricted shift operators in the discrete case and translation semigroups in the continuous one. Thus it was a highly nonself-adjoint theory.

It is the object of this paper to remedy this situation. In the process we develop the theory of spectral representations along nonclassical lines which point out more clearly the connections with structure theory of operators in finite dimensional vector spaces as developed in [10] and the structure theory of shift operators [15], [24].

Much as the left shift operator, restricted to an invariant subspace, served as a model for a general contraction, the idea of a model for an operator was used in an algebraic context in [10] and it appears in the theory of spectral representations of self-adjoint operators. Thus a spectral representation is a convenient model to work with. The relation between matrix measures their corresponding  $L^2$  spaces and spectral representations is studied in § 2. Section 3 outlines briefly the theory of multiplicity, the ordered and canonical spectral representations of a given self-adjoint operator.

A central role in the theory is played by operators intertwining two self-adjoint operators in Hilbert spaces. Since each self-adjoint operator in a Hilbert space induces a natural module structure on the space, the intertwining operators are essentially module homomorphisms where the relevant ring is the algebra of bounded measurable functions. Convenient representations for these homomorphisms are found by way of a lifting theorem which is the analogue of the Sz.-Nagy and Foias lifting theorem for contractions. This should be compared also with the purely algebraic result as it appears in [10]. Necessary and sufficient conditions for left and right invertibility of these intertwining operators are established. For the analogues in other contexts we refer to [9], [10], [25]. This takes up § 4. In the last section we apply all this machinery to the

---

\* Received by the editors December 4, 1978, and in revised form May 15, 1979.

† Department of Mathematics, Ben Gurion University of the Negev, Beer Sheva, Israel. This research was supported in part by the Israeli Academy of Sciences and the Israel Commission for Basic Research.

study of systems. We study controllability and place a theorem of Fattorini [7] in what seems to be its natural context. We introduce a stronger notion of controllability and obtain an isomorphism theorem along the lines of that of Helton [16].

For the simplification of the proofs we assume that all the self-adjoint operators are of finite multiplicity. However, there is no doubt that all the results generalize to the case of any self-adjoint operator in a separable Hilbert space.

**2. The spectral theorem and models for self-adjoint operators.** We take as our starting point the spectral theorem [1], [6], [12] which states that any self-adjoint operator  $A$  has an integral representation of the form

$$(2.1) \quad A = \int_{-\infty}^{\infty} \lambda E(d\lambda),$$

where  $E(\cdot)$  is the associated spectral measure, that is a (orthogonal) projection valued  $\sigma$ -additive set function defined on the Borel sets of the real line. In case  $A$  is bounded the integral (2.1) converges in the uniform operator topology whereas for unbounded operators questions of convergence are handled in the strong operator topology. The integral (2.1) allows us to construct a functional calculus. For each bounded Borel measurable function  $\varphi$  defined on  $\mathbb{R}$  we can define  $\varphi(A)$ . This is done by letting

$$(2.2) \quad \varphi(A) = \int_{-\infty}^{\infty} \varphi(\lambda) E(d\lambda).$$

The map  $\varphi \rightarrow \varphi(A)$  is a  $*$ -homomorphism of the algebra  $\mathcal{B}$  of all bounded Borel measurable functions on  $\mathbb{R}$  into the algebra  $B(H)$  of all bounded operators on the Hilbert  $H$ . In particular the underlying Hilbert space becomes a  $\mathcal{B}$ -module via the definition

$$(2.3) \quad \varphi \cdot x = \varphi(A)x \quad \text{for all } x \in H.$$

Our next object is to obtain a functional representation of the Hilbert space. Having a functional representation of the Hilbert space has the advantage of providing extra structure in terms of which certain problems can be resolved in a concrete way. To this end we assume our self-adjoint operator  $A$  has a finite set of generators. Here a set of vectors  $x_1, \dots, x_r \in H$  is called a *set of generators* if the set of all vectors of the  $\sum_{i=1}^r \varphi_i(A)x_i$  where  $\varphi_i \in \mathcal{B}$  is a dense subset of  $H$ . Let  $\mathcal{B}^r$  be the Cartesian product of  $r$  copies of  $\mathcal{B}$ . Clearly  $\mathcal{B}^r$  is a  $\mathcal{B}$ -module. We define the map  $\rho : \mathcal{B}^r \rightarrow H$  by

$$(2.4) \quad \rho(\varphi_1, \dots, \varphi_r) = \sum_{i=1}^r \varphi_i(A)x_i,$$

where  $x_1, \dots, x_r$  is the fixed set of generators for  $A$ . The map  $\rho$  is, by elementary properties of the functional calculus a  $\mathcal{B}$ -module homomorphism, and by our assumption that  $x_1, \dots, x_r$  is a set of generators it follows that  $\rho$  has range which is dense in  $H$ .

Computing the norm of  $\rho(\varphi_1, \dots, \varphi_r)$  we obtain

$$\begin{aligned} \|\sum \varphi_i(A)x_i\|^2 &= \sum_i \sum_j (\varphi_i(A)x_i, \varphi_j(A)x_j) = \sum \sum (\overline{\varphi_j(A)}\varphi_i(A)x_i, x_j) \\ &= \sum \sum \int \overline{\varphi_j(\lambda)}\varphi_i(\lambda)(E(d\lambda)x_j, x_j). \end{aligned}$$

Define now the (complex) measures  $\mu_{ij}$  by

$$(2.5) \quad \mu_{ij}(\sigma) = (E(\sigma)x_i, x_j)$$



for all Borel sets  $\sigma$  and let  $\mathbb{M}$  be the matrix whose  $i, j$  entry is  $\mu_{ij}$ . We call such an object a *matrix measure* [6]. We say a matrix measure is a positive matrix measure if for each Borel set  $\sigma$ ,  $\mathbb{M}(\sigma)$  is a nonnegative definite Hermitian matrix. It is easily checked that the matrix measure  $\mathbb{M}$  constructed in (2.5) is a positive matrix measure. Indeed let  $\sigma$  be a Borel subset of  $\mathbb{R}$  and let  $\alpha_1, \dots, \alpha_r$  be complex numbers; then with  $a = (\alpha_1, \dots, \alpha_r)$

$$\begin{aligned} (\mathbb{M}(\sigma)a, a) &= \sum_i \sum_j \mu_{ij} \alpha_i \bar{\alpha}_j = \sum_i \sum_j (E(\sigma)x_i, x_j) \alpha_i \bar{\alpha}_j \\ &= (E(\sigma) \sum_i \alpha_i x_i, \sum_j \alpha_j x_j) = \|E(\sigma) \sum_i \alpha_i x_i\|^2 \geq 0. \end{aligned}$$

In terms of the matrix measure introduced we have

$$(2.6) \quad \|\rho F\|^2 = \|\sum f_i(A)x_i\|^2 = \int (d\mathbb{M}F, F),$$

where  $F \in \mathcal{B}^r$  is the vector function whose components are  $f_1, \dots, f_r$ . Equality (2.6) indicates that if we define properly the  $L^2$  space of a matrix measure  $\mathbb{M}$  which we will denote naturally by  $L^2(\mathbb{M})$  then the map  $\rho : \mathcal{B}^r \rightarrow H$  will have a natural extension to a unitary map of  $L^2(\mathbb{M})$  onto  $H$ . Moreover, such a map satisfies

$$(2.7) \quad \rho(\varphi F) = \varphi(A)(\rho F) \quad \text{for all } \varphi \in \mathcal{B}.$$

Also for any vector  $x$  in the domain of  $A$  we have

$$(2.8) \quad [\rho^{-1}(Ax)](\lambda) = \lambda \cdot (\rho^{-1}x)(\lambda).$$

Thus in the functional representation  $A$  acts like multiplication by  $\lambda$ .

We note that  $\mathbb{M}$  has a convenient description in terms of the spectral measure  $E(\cdot)$  that is associated with  $A$ . If  $J : \mathbb{C}^r \rightarrow H$  is the map sending  $(\alpha_1, \dots, \alpha_r)$  to  $\sum \alpha_i x_i$  then for each Borel set  $\sigma$  we have

$$(2.9) \quad \mathbb{M}(\sigma) = J^* E(\sigma) J.$$

To define  $L^2(\mathbb{M})$  we proceed as in [6]. We denote by  $L^2_0(\mathbb{M})$  the set of all  $r$ -tuples  $(f_1, \dots, f_r)$  of Borel measurable functions for which

$$(2.10) \quad \|F\|^2 = \int_{-\infty}^{\infty} (d\mathbb{M}F, F) = \int_{-\infty}^{\infty} \sum \sum f_i(\lambda) \overline{f_j(\lambda)} d\mu_{ij} < \infty$$

and define  $L^2(\mathbb{M})$  as the set of all equivalence classes in  $L^2_0(\mathbb{M})$  modulo the set of null functions, a null function being one for which  $\|F\| = 0$ . With the inner product in  $L^2(\mathbb{M})$  defined by

$$(F, G) = \int (d\mathbb{M}F, G) = \int \sum \sum f_i(\lambda) \overline{g_j(\lambda)} d\mu_{ij},$$

$L^2(\mathbb{M})$  becomes a pre-Hilbert space and the only open question is that of completeness. There is one class of matrix measures for which  $L^2(\mathbb{M})$  is clearly complete, namely the class of positive diagonal measures, i.e., those for which  $i \neq j$  implies  $\mu_{ij} = 0$  and the diagonal elements are positive measures. If  $\mu_1, \dots, \mu_r$  are the diagonal elements of a diagonal matrix measure then in this case

$$\|F\|^2 = \int \sum |f_i(\lambda)|^2 d\mu_i = \sum \int |f_i(\lambda)|^2 d\mu_i = \sum \|f_i\|^2,$$

where  $\|f_i\|$  is the norm of  $f_i$  as an element of  $L^2(\mu_i)$ . Hence in this case  $L^2(\mathbb{M})$  is clearly equal to the direct sum  $L^2(\mu_1) \oplus \dots \oplus L^2(\mu_r)$  which is a complete space. We will use this

observation to show completeness of  $L^2(\mathbb{M})$  by exhibiting a unitary map that diagonalizes  $\mathbb{M}$ .

As a first step we simplify the problem by replacing matrix measures by density matrices and one scalar measure. We choose a positive measure  $\mu$  such that all  $\mu_{ij}$  are absolutely continuous with respect to  $\mu$ ,  $\mu_{ij} \ll \mu$ . One candidate is the sum of the total variations of all the  $\mu_{ij}$ . A better choice turns out later to be the trace of  $\mathbb{M}$ . If  $m_{ij} = d\mu_{ij}/d\mu$  is the Radon–Nikodym derivative of  $\mu_{ij}$  with respect to  $\mu$  then we introduce the density matrix

$$(2.11) \quad M(\lambda) = (m_{ij}(\lambda)).$$

The next lemma is quoted from [6].

LEMMA 2.1. *If  $M(\lambda)$  is density matrix of a matrix measure  $\mathbb{M}$  with respect to a scalar measure  $\mu$  then  $M(\lambda)$  is nonnegative definite  $\mu$ -a.e.*

Consider next the set of all positive matrix measures on  $\mathbb{R}$ . We say that  $\mathbb{M}$  divides  $\mathbb{N}$ , and write  $\mathbb{M}|\mathbb{N}$ , if there exists a Borel matrix function  $H$  such that

$$(2.12) \quad d\mathbb{M} = H^* d\mathbb{N}H.$$

Two matrix measures  $\mathbb{M}$  and  $\mathbb{N}$  are *equivalent* and we write  $\mathbb{M} \sim \mathbb{N}$  if  $\mathbb{M}|\mathbb{N}$  and  $\mathbb{N}|\mathbb{M}$ .

The division relation is clearly reflexive and transitive and hence induces a partial order in the set of all matrix measures. Relation (2.12) is a generalization of the concept of absolute continuity as applied to matrix measures. Heuristically the matrix function  $H$  has the interpretation of a “square root” of a generalized Radon–Nikodym derivative of  $\mathbb{M}$  with respect to  $\mathbb{N}$ . We point out that  $\mathbb{M}$  and  $\mathbb{N}$  do not have to be necessarily of the same size. In that case  $H$  will not be a square matrix. For scalar measures  $\mu$  and  $\nu$  we have of course  $\mu|\nu$  if and only if  $\mu \ll \nu$ .

The partial order in the set of positive matrix measures is reflected in the corresponding  $L^2(\mathbb{M})$  spaces. Given two matrix measures  $\mathbb{M}$  and  $\mathbb{N}$  then we say that a map  $U : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  is an *embedding* if it is an injective  $\mathcal{B}$ -homomorphism. If  $U$  is also an isometry we say  $U$  is an *isometric embedding*.

The next lemma provides a large class of isometric embeddings. The scalar case appears in [5], [20].

LEMMA 2.2 *Let  $\mathbb{M}$  and  $\mathbb{N}$  be positive matrix measures and assume that  $\mathbb{M}|\mathbb{N}$ . Then there exists an isometric embedding of  $L^2(\mathbb{M})$  into  $L^2(\mathbb{N})$ .*

*Proof.* Since  $\mathbb{M}|\mathbb{N}$  there exists a measurable matrix function  $H$  such that (2.12) holds. Define  $U_{\mathbb{M}}^{\mathbb{N}} : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  by

$$(2.13) \quad U_{\mathbb{M}}^{\mathbb{N}}F = HF \quad \text{for } F \in L^2(\mathbb{M});$$

then clearly

$$\|U_{\mathbb{M}}^{\mathbb{N}}F\|^2 = \int (d\mathbb{N}HF, HF) = \int (H^*d\mathbb{N}HF, F) = \int (d\mathbb{M}F, F) = \|F\|^2.$$

So  $U_{\mathbb{M}}^{\mathbb{N}}$  is an isometry and it is easily checked that it is a  $\mathcal{B}$ -homomorphism.

We note that the set of isometries  $U_{\mathbb{M}}^{\mathbb{N}}$  is a *coherent set of isometries* [5] in the sense that if  $\mathbb{M}|\mathbb{N}$  and  $\mathbb{N}|\mathbb{S}$  then we have

$$(2.14) \quad U_{\mathbb{M}}^{\mathbb{S}} = U_{\mathbb{N}}^{\mathbb{S}}U_{\mathbb{M}}^{\mathbb{N}}.$$

The equivalence of two matrix measures can be described also in terms of their density matrices with respect to a common scalar measure. To this end we define a notion of equivalence between measurable matrix functions. Let  $M$  and  $N$  be Borel measurable  $n \times m$  matrix functions defined on a subset  $\chi$  of  $\mathbb{R}$ , and let  $\sigma$  be a positive

measure on  $\mathbb{R}$ . We say that  $M$  and  $N$  are  $\sigma$ -equivalent if there exist  $\sigma$ -a.e. invertible measurable  $n \times n$  and  $m \times m$  matrix functions  $P$  and  $R$  such that

$$(2.15) \quad M(\lambda) = P(\lambda)N(\lambda)R(\lambda), \quad \sigma\text{-a.e.}$$

If  $M$  and  $N$  are square matrix functions we say that  $M$  and  $N$  are unitarily  $\sigma$ -equivalent if there exists a measurable  $\sigma$ -a.e. unitary matrix function  $P$  such that

$$(2.16) \quad M(\lambda) = P(\lambda)^*N(\lambda)P(\lambda), \quad \sigma\text{-a.e.}$$

It is clear that both relations are bona fide equivalence relations and unitary  $\sigma$ -equivalence implies  $\sigma$ -equivalence. Also if  $\nu$  is a positive measure and  $\nu \ll \sigma$  then  $\sigma$ -equivalence implies  $\nu$ -equivalence.

In terms of these notions we can state the next lemma, quoted from [6], which is the main technical result for the proof of completeness, in the following form.

LEMMA 2.3. *Let  $\mathbb{M}$  be a positive matrix measure and let  $M$  be its density with respect to a positive measure  $\mu$  that satisfies  $\mu_{ij} \ll \mu$ . Then there exists a diagonal matrix function  $D$  such that  $M$  and  $D$  are unitarily  $\mu$ -equivalent.*

Alternately stated there exists a measurable matrix function  $H$  such that

$$(2.17) \quad H(\lambda)^*H(\lambda) = I$$

and

$$(2.18) \quad M(\lambda) = H(\lambda)^*D(\lambda)H(\lambda)$$

hold  $\mu$ -a.e.

We note that  $\mu$ -a.e.  $M(\lambda)$  is a nonnegative definite matrix and hence can be diagonalized by a unitary matrix. The lemma's content is that the pointwise diagonalizations can be made in a globally measurable way. With this lemma we can prove the completeness of  $L^2(\mathbb{M})$  following [6].

THEOREM 2.4. *If  $\mathbb{M}$  is a positive measure on  $\mathbb{R}$  then  $L^2(\mathbb{M})$  is a Hilbert space.*

*Proof.* Let  $\mu$ ,  $H$  and  $D$  be as in the previous lemma and let  $d\mathbb{D} = D d\mu$ . The map  $U_{\mathbb{M}}^{\mathbb{D}} : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{D})$  given by (2.13) is an isometric embedding. However, it is invertible and we have

$$(2.19) \quad (U_{\mathbb{M}}^{\mathbb{D}})^{-1} = U_{\mathbb{D}}^{\mathbb{M}} = (U_{\mathbb{M}}^{\mathbb{D}})^*,$$

where  $U_{\mathbb{D}}^{\mathbb{M}}G = H^*G$ . Thus  $U_{\mathbb{M}}^{\mathbb{D}}$  is a unitary map and

$$(2.20) \quad L^2(\mathbb{D}) = L^2(\delta_1) \oplus \cdots \oplus L^2(\delta_r),$$

where  $\delta_i$  are the measures defined by  $\delta_i(\sigma) = \int d_i(\lambda) d\mu$ . Thus  $L^2(\mathbb{D})$  is complete and so is  $L^2(\mathbb{M})$ .

For  $\varphi \in \mathcal{B}$  we define the operator of multiplication by  $\varphi$  in  $L^2(\mathbb{M})$  by

$$(2.21) \quad M_{\varphi, \mathbb{M}}F = \varphi F \quad \text{for } F \in L^2(\mathbb{M}).$$

We single out the identity function  $\chi$ ,  $\chi(\lambda) = \lambda$ , in terms of which we can summarize the previous results and exhibit a functional representation for the Hilbert space  $H$  and the self-adjoint operator  $A$  acting in it.

THEOREM 2.5. *Any operator  $A$  in a Hilbert space  $H$  is unitarily equivalent to an operator  $M_{\chi, \mathbb{M}}$  in  $L^2(\mathbb{M})$  for some positive matrix measure  $\mathbb{M}$  on the real line if and only if  $A$  is a finitely generated self-adjoint operator.*

**3. Unitary invariants and multiplicity theory.** This section is devoted to the characterization of the unitary invariants of self-adjoint operators by means of the

ordered spectral representation. Again we restrict ourselves to the finitely generated case. By Theorem 2.5 we may assume that  $A$  is the self-adjoint operator  $\mathbb{M}_{\chi, \mathbb{M}}$  defined by (2.21) in the Hilbert space  $L^2(\mathbb{M})$ . Since the matrix measure  $\mathbb{M}$  was determined by a choice of generators the question of a best choice of generators immediately arises. By best we mean a choice that is canonical in some sense, modulo natural equivalences, and which exhibits the structure of the operator in as simple form as possible. This is a classical problem first resolved by Hellinger [14]. For various expositions we refer to [6], [12], [21], [5], [3], [20]. Our approach gets at the result by simple matrix manipulation. The price is the loss of generality involved by assuming finite multiplicity.

**THEOREM 3.1.** *Let  $A$  be a finitely generated self adjoint generator in a Hilbert space  $H$ . Then there exists a finite sequence of positive measures  $\mu_1 \gg \mu_2 \gg \dots \gg \mu_p$  such that  $A$  is unitarily equivalent to*

$$(3.1) \quad M_{\chi, \mu_1} \oplus \dots \oplus M_{\chi, \mu_p}$$

acting in

$$(3.2) \quad L^2(\mu_1) \oplus \dots \oplus L^2(\mu_p).$$

The sequence  $\mu_1, \dots, \mu_p$  is determined by  $A$  up to equivalence of measures.

The representation (3.1) of the operator  $A$  is referred to as the *ordered spectral representation*. The integer  $p$  is referred to as the *multiplicity* of  $A$ .

The proof of Theorem 3.1 is a direct consequence of the following lemma.

**LEMMA 3.2.** *Let  $\mathbb{L} = (\lambda_{ij})$  be a positive matrix measure and let  $\sigma$  be a positive measure such that  $\mathbb{L} \ll \sigma$ . Then there exists a diagonal matrix measure  $\mathbb{M}$  with diagonal entries  $\mu_1, \dots, \mu_p$  such that  $d\mu_i = m_i d\sigma$  and the following statements hold:*

- (i)  $\mu_1 \gg \mu_2 \gg \dots \gg \mu_p$ ; and
- (ii)  $\mathbb{L}$  and  $\mathbb{M}$  are unitarily  $\sigma$  equivalent.

Moreover if  $\mathbb{N}$  is another diagonal matrix measure with diagonal entries  $\nu_1, \dots, \nu_p$  such that  $d\nu_i = n_i d\rho$  and the statements

- (i')  $\nu_1 \gg \nu_2 \gg \dots \gg \nu_p$ ; and
- (ii')  $\mathbb{L}$  and  $\mathbb{N}$  are unitarily  $\rho$ -equivalent

hold then  $\mathbb{M}$  and  $\mathbb{N}$  are unitarily  $\tau$ -equivalent where  $\tau = \rho \wedge \sigma$  is the infimum of the measures  $\rho$  and  $\sigma$  [5], [12].

*Proof.* By Lemma 2.3 it suffices to show that given a diagonal matrix measure  $\mathbb{L}$  it can be reduced to canonical form. Thus without loss of generality we let  $\mathbb{L}$  be diagonal with diagonal elements  $\lambda_1, \dots, \lambda_p$  where, by assumption,  $\lambda_i \ll \sigma$ . Let  $d\lambda_i = l_i d\sigma$ , i.e.  $l_i$  is the Radon–Nikodym derivative of  $\lambda_i$  with respect to  $\sigma$ . For simplicity of notation we assume  $p = 2$ . Let  $\lambda_2 = \lambda'_2 + \lambda''_2$  be the Lebesgue decomposition of  $\lambda_2$  with respect to  $\lambda_1$ , assuming  $\lambda'_2 \ll \lambda_1$  and  $\lambda''_2 \perp \lambda_1$ . Let  $l_2 = l'_2 + l''_2$  with  $l'_2$  and  $l''_2$  the respective Radon–Nikodym derivatives of  $\lambda'_2$  and  $\lambda''_2$  with respect to  $\sigma$ . Let  $E_2 = \{\lambda \mid l''(\lambda) \neq 0\}$  and  $F_2 = \{\lambda \mid l''(\lambda) = 0\}$  and let  $\chi_{E_2}$  and  $\chi_{F_2}$  be the corresponding characteristic function of the two sets.

Define a  $2 \times 2$  matrix function  $H(\lambda)$  by

$$H(\lambda) = \begin{pmatrix} \chi_{F_2}(\lambda) & \chi_{E_2}(\lambda) \\ \chi_{E_2}(\lambda) & \chi_{F_2}(\lambda) \end{pmatrix}.$$

A simple calculation yields the equality

$$(3.3) \quad L(\lambda) = H(\lambda)^* L'(\lambda) H(\lambda),$$

where

$$(3.4) \quad L(\lambda) = \begin{pmatrix} l_1(\lambda) & 0 \\ 0 & l_2(\lambda) \end{pmatrix} \quad \text{and} \quad L'(\lambda) = \begin{pmatrix} l_1(\lambda) + l''_2(\lambda) & 0 \\ 0 & l'_2(\lambda) \end{pmatrix}$$

which proves the statement for  $p = 2$ . The necessary modifications needed to make the proof work for  $p > 2$  are obvious. So much for the existence of the canonical diagonalization.

To prove the uniqueness part we note the obvious fact that if  $\mathbb{L}$  and  $\mathbb{M}$  are unitarily  $\sigma$ -equivalent they are also unitarily  $\sigma'$ -equivalent for any  $\sigma' \ll \sigma$ . It follows that if we form the infimum  $\tau = \rho \wedge \sigma$  of the measures  $\rho$  and  $\sigma$  transitivity  $\mathbb{M}$  and  $\mathbb{N}$  are unitarily  $\tau$ -equivalent. Thus  $\tau$ -a.e. the diagonal matrices

$$\begin{pmatrix} m'_1(\lambda) & \cdots & 0 \\ 0 & \cdots & m'_p(\lambda) \end{pmatrix} \text{ and } \begin{pmatrix} n'_1(\lambda) & \cdots & 0 \\ 0 & \cdots & n'_p(\lambda) \end{pmatrix}$$

are unitarily equivalent. Here  $m'_i$  and  $n'_i$  are the Radon–Nikodym derivatives of  $\mu_i$  and  $\nu_i$  with respect to  $\tau$ . Since assumptions (i) and (i') imply  $m'_{i+1}(\lambda) = 0$  whenever  $m'_i(\lambda) = 0$  it follows that the zero sets of  $m'_i$  and  $n'_i$  are equal  $\tau$ -a.e. This is equivalent to  $\mu_i \approx \nu_i$ .

There is another representation associated with a self adjoint operator which is closely related to the ordered spectral representation.

**THEOREM 3.3.** *Let  $A$  be a finitely generated self-adjoint operator in a Hilbert space  $H$ . Then there exists a finite sequence of mutually singular positive measures  $\nu_1, \dots, \nu_p$  such that  $A$  is unitarily equivalent to*

$$(3.5) \quad M_{x, \mathbb{N}_1} \oplus \cdots \oplus M_{x, \mathbb{N}_p}$$

acting in

$$(3.6) \quad L^2(\mathbb{N}_1) \oplus \cdots \oplus L^2(\mathbb{N}_p),$$

where  $N_j = \nu_j I_j$ ,  $I_j$  being the  $j \times j$  identity matrix. The sequence of measures  $\nu_1, \dots, \nu_p$  is determined by  $A$  up to equivalence of measures.

The representation (3.5) is referred to as the *canonical spectral representation* of  $A$ .

The passage from Theorem 3.1 to Theorem 3.3 is straightforward using repeatedly the Lebesgue decomposition theorem for measures. We omit the details.

We remark that another alternative way of writing the canonical spectral representation is to define the matrix measure  $\mathbb{N}$  by

$$(3.7) \quad \mathbb{N} = \begin{pmatrix} \nu_1 + \cdots + \nu_p & & & \\ & \nu_2 + \cdots + \nu_p & & \\ & & \ddots & \\ & & & \nu_p \end{pmatrix};$$

then  $A$  is unitarily equivalent to the operator

$$(3.8) \quad M_{x, \mathbb{N}}$$

acting on  $L^2(\mathbb{N})$ .

**4. Operators intertwining self-adjoint operators.** Given two operators  $A_1$  and  $A_2$  we say operator  $X$  *intertwines*  $A_1$  and  $A_2$  if

$$(4.1) \quad XA_1 = A_2X.$$

If the intertwining operator  $X$  is boundedly invertible then  $A_1$  and  $A_2$  are similar. Given two self-adjoint operators  $A_1$  and  $A_2$  then their similarity implies unitary equivalence. However, even if  $X$  is not boundedly invertible the existence of intertwining operators having some additional properties (left or right invertibility) yields information regarding the structure of  $A_1$  and  $A_2$ . If  $A_1$  and  $A_2$  act in the Hilbert spaces  $H_1$  and  $H_2$  respectively then the characterization of intertwining operators is equivalent to the determination of a class of module homomorphisms, where the modules are  $H_1$  and

$H_2$  with the module structure induced by  $A_1$  and  $A_2$  through the classical functional calculus.

As a consequence of Theorem 2.5 the study of operators intertwining two (finitely generated) self adjoint operators reduces to those intertwining two operators of the form  $S_{x, \mathbb{M}}$ . In the set of all matrix measures we single out the set of all *scalar type measures* which are the matrix measures of the form  $\sigma I$ , i.e., diagonal matrix measures with all diagonal elements being equal to  $\sigma$ .

Given a matrix measure  $\mathbb{M}$ , a subspace  $K$  of  $L^2(\mathbb{M})$  is called an *invariant subspace* if

$$(4.2) \quad M_{\varphi, \mathbb{M}} K \subset K$$

for all  $\varphi \in \mathcal{B}$ , i.e., if it is invariant under all multiplication operators by bounded measurable functions.

The following theorem is generally known. One version of it appears in [15].

**THEOREM 4.1.** *A subspace  $K$  of  $L^2(\sigma I)$  is an invariant subspace if and only if  $K = PL^2(\sigma I)$  where  $P$  is a measurable  $\sigma$  a.e. projection valued matrix function.*

Clearly a subspace  $K$  is invariant if and only if its orthogonal complement  $K^\perp$  is invariant. If  $P^\perp$  is the projection valued function corresponding to  $K^\perp$  then we have  $P^\perp = I - P$ .

The next theorem characterizes all  $\mathcal{B}$  homomorphisms of  $L^2(\sigma I)$ .

**THEOREM 4.2.** *Let  $X : L^2(\sigma I) \rightarrow L^2(\sigma I)$  be a  $\mathcal{B}$  homomorphism. Then there exists a measurable  $\sigma$ -a.e. bounded matrix function  $\Xi$  such that for all  $F \in L^2(\sigma I)$*

$$(4.3) \quad (XF)(\lambda) = \Xi(\lambda)F(\lambda).$$

*Conversely any operator  $X$  defined by (4.3) is a  $\mathcal{B}$  homomorphism.*

If  $\sigma I$  is a scalar type measure then we will write  $U_{\mathbb{M}}^\sigma$  for the isometric embedding of  $L^2(\mathbb{M})$  into  $L^2(\sigma I)$ . Here we assume  $\mathbb{M} \ll \sigma I$  or equivalently  $d\mathbb{M} = H(\lambda)^* H(\lambda) d\sigma$ . If  $M(\lambda)$  is the Radon–Nikodym derivative of  $\mathbb{M}$  with respect to  $\sigma$  then  $M(\lambda) = H(\lambda)^* H(\lambda) \sigma$ -a.e. It will be of interest to have a concrete representation for  $(U_{\mathbb{M}}^\sigma)^*$ , the adjoint of the isometric embedding  $U_{\mathbb{M}}^\sigma$ . For this we need to know something about pseudoinverses.

If  $T : H_1 \rightarrow H_2$  is a bounded operator between two Hilbert spaces and has closed range then  $T|_{\{\text{Ker } T\}^\perp} \rightarrow \text{Range } T$  is an invertible operator. Extend the definition of that inverse to  $\{\text{Range } T\}^\perp$  by defining it to be zero there. The extended operator, uniquely determined by  $T$ , is called the pseudoinverse of  $T$  and denoted by  $T^\#$  [19]. The property of the pseudoinverse which we need is

$$(4.4) \quad TT^\# T = T.$$

If we deal with complex matrices the above definition makes sense with the usual inner product in  $\mathbb{C}^n$ .

**THEOREM 4.3.** *Let  $\mathbb{M}$  be a matrix measure and  $\mathbb{M} \ll \sigma I$  with*

$$(4.5) \quad d\mathbb{M} = M(\lambda) d\sigma = H(\lambda)^* H(\lambda) d\sigma.$$

*Let  $P$  be the projection valued function corresponding to the invariant subspace  $U_{\mathbb{M}}^\sigma L^2(\mathbb{M})$  of  $L^2(\sigma I)$ . Then we have*

$$(4.6) \quad (U_{\mathbb{M}}^\sigma)^* G = H^\# P G$$

*for all  $G \in L^2(\sigma I)$  where  $H^\#$  is the pseudoinverse of  $H$ .*

*Proof.* Let  $F \in L^2(\mathbb{M})$  and  $G \in L^2(\sigma I)$ , then

$$\begin{aligned} (F, (U_{\mathbb{M}}^{\sigma})^* G) &= (U_{\mathbb{M}}^{\sigma} F, G) = \int (H(\lambda)F(\lambda), G(\lambda)) \, d\sigma \\ &= \int (H(\lambda)F(\lambda), P(\lambda)G(\lambda)) \, d\sigma. \end{aligned}$$

Since  $PG \in U_{\mathbb{M}}^{\sigma} L^2(\mathbb{M})$  there exists an element  $G_0 \in L^2(\mathbb{M})$  such that  $HG_0 = PG$ . By (4.4) we have

$$PG = HG_0 = HH^{\#}HG_0 = HH^{\#}PG.$$

Using this equality we obtain

$$\begin{aligned} (F, (U_{\mathbb{M}}^{\sigma})^* G) &= \int (H(\lambda)F(\lambda), P(\lambda)G(\lambda)) \, d\sigma \\ &= \int (H(\lambda)F(\lambda), H(\lambda)H(\lambda)^{\#}P(\lambda)G(\lambda)) \, d\sigma \\ &= \int (H(\lambda)^*H(\lambda)F(\lambda), H(\lambda)^{\#}P(\lambda)G(\lambda)) \, d\sigma \\ &= \int (d\mathbb{M}F, H^{\#}PG) \end{aligned}$$

which proves (4.6).

Using this theorem we can obtain a representation for the adjoint of any isometric embedding.

**COROLLARY 4.4.** *Let  $\mathbb{M}$  and  $\mathbb{N}$  be matrix measures such that  $\mathbb{M}|\mathbb{N}$  and let  $\sigma I$  be a scalar type measure divisible by both  $\mathbb{M}$  and  $\mathbb{N}$ . Assume  $d\mathbb{M} = H^*H \, d\sigma$  and  $d\mathbb{N} = K^*K \, d\sigma$ ; then*

$$(4.7) \quad (U_{\mathbb{M}}^{\mathbb{N}})^* F = H^{\#} QKF$$

for all  $F \in L^2(\mathbb{N})$ , where  $Q$  is the projection valued function corresponding to the invariant subspace  $U_{\mathbb{M}}^{\sigma} L^2(\mathbb{M})$  of  $L^2(\sigma I)$ .

*Proof.* We have  $U_{\mathbb{M}}^{\sigma} = U_{\mathbb{N}}^{\sigma} U_{\mathbb{M}}^{\mathbb{N}}$  and hence  $(U_{\mathbb{M}}^{\sigma})^* = (U_{\mathbb{M}}^{\mathbb{N}})^* (U_{\mathbb{N}}^{\sigma})^*$ . Since  $U_{\mathbb{M}}^{\sigma}$  is isometric we have

$$(4.8) \quad (U_{\mathbb{M}}^{\mathbb{N}})^* = (U_{\mathbb{M}}^{\sigma})^* U_{\mathbb{N}}^{\sigma}.$$

Applying Theorem 4.3 to (4.8) yields (4.7).

The next two results are instances of lifting theorems. They describe complicated  $\mathcal{B}$ -homomorphisms between two spaces of type  $L^2(\mathbb{M})$  in terms of  $\mathcal{B}$ -homomorphisms of  $L^2(\sigma I)$  which have been described in Theorem 4.2. Theorem 4.6 below is modeled after the Sz.-Nagy–Foias lifting theorem [24]. For the algebraic analogue of this result we refer to [10].

**LEMMA 4.5.** *Let  $\mathbb{M}$  be a matrix measure and assume  $\mathbb{M}|\sigma I$ . Let  $\chi : L^2(\sigma I) \rightarrow L^2(\mathbb{M})$  be a  $\mathcal{B}$ -homomorphism. Then there exists a  $\mathcal{B}$ -homomorphism  $\bar{\chi} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  for which*

$$(4.9) \quad X = (U_{\mathbb{M}}^{\sigma})^* \bar{X}$$

and  $\|\bar{X}\| = \|X\|$ . This implies the existence of a measurable  $\sigma$ -a.e. bounded matrix function  $\Xi$ , with  $\|\Xi\|_{\infty} = \|X\|$  in terms of which we have the representation

$$(4.10) \quad XF = H^{\#} P \Xi F \quad \text{for } F \in L^2(\sigma I).$$

$P$  is the projection valued matrix function corresponding to  $U_M^\sigma L^2(\mathbb{M})$ .

Conversely any map  $X : L^2(\sigma I) \rightarrow L^2(\mathbb{M})$  defined by (4.9) where  $\bar{X}$  is a  $\mathcal{B}$ -homomorphism is also a  $\mathcal{B}$ -homomorphism and

$$(4.11) \quad \|X\| = \|\bar{X}\| = \|\Xi\|_\infty.$$

*Proof.* If  $\bar{X} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  is a  $\mathcal{B}$ -homomorphism then so is its composition with  $(U_M^\sigma)^*$  and obviously (4.11) holds.

Conversely let  $X : L^2(\sigma I) \rightarrow L^2(\mathbb{M})$  be a  $\mathcal{B}$ -homomorphism. Define  $\bar{X} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  by

$$(4.12) \quad \bar{X}F = U_M^\sigma X F \quad \text{for } F \in L^2(\sigma I).$$

Clearly  $\bar{X}$  as a product of  $\mathcal{B}$ -homomorphisms is also one and since  $U^\sigma$  is isometric  $\|\bar{X}\| = \|X\|$ . By Theorem 4.2 there exists a  $\sigma$ -a.e. bounded measurable matrix function  $\Xi$  for which  $\bar{X}F = \Xi F$  and hence (4.10) holds by an application of Theorem 4.3.

**THEOREM 4.6.** *Let  $\mathbb{M}$  and  $\mathbb{N}$  be matrix measures and  $X : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  a  $\mathcal{B}$ -homomorphism. Let  $\sigma I$  be a positive scalar type measure divisible by both  $\mathbb{M}$  and  $\mathbb{N}$  and let  $d\mathbb{M} = H^*H d\sigma$  and  $d\mathbb{N} = K^*K d\sigma$ . Let  $P$  and  $Q$  be the measurable projection valued functions corresponding to  $U_M^\sigma L^2(\mathbb{M})$  and  $U_N^\sigma L^2(\mathbb{N})$  respectively. Then there exists a  $\mathcal{B}$ -homomorphism  $\bar{X} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  satisfying  $\|\bar{X}\| = \|X\|$  for which*

$$(4.13) \quad X F = (U_N^\sigma)^* \bar{X} U_M^\sigma F \quad \text{for } F \in L^2(\mathbb{M}).$$

Moreover, there exists a measurable  $\sigma$ -a.e. bounded matrix function  $\Xi$  satisfying

$$(4.14) \quad \|\Xi\|_\infty = \|\bar{X}\| = \|X\|,$$

$$(4.15) \quad \Xi(\lambda) = \Xi(\lambda)P(\lambda) = Q(\lambda)\Xi(\lambda), \quad \sigma\text{-a.e.},$$

and for which

$$(4.16) \quad X F = K^\# Q \Xi H F \quad \text{for all } F \in L^2(\mathbb{M}).$$

Conversely every operator  $X$  defined by (4.16) for  $\Xi$  measurable and  $\sigma$ -a.e. bounded is a  $\mathcal{B}$ -homomorphism from  $L^2(\mathbb{M})$  into  $L^2(\mathbb{N})$ .

*Proof.* If  $X$  is given by (4.16) then it is clearly a  $\mathcal{B}$ -homomorphism and satisfies (4.14). Let us assume therefore that  $X : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  is a  $\mathcal{B}$ -homomorphism. Define  $Y : L^2(\sigma I) \rightarrow L^2(\mathbb{N})$  by

$$(4.17) \quad Y F = X (U_M^\sigma)^* F.$$

$Y$  is a  $\mathcal{B}$ -homomorphism as a product of such and  $Y \{U_M^\sigma L^2(\mathbb{M})\}^\perp = 0$  or equivalently stated  $Y P^\perp L^2(\sigma I) = 0$  which boils down to

$$(4.18) \quad Y P^\perp = 0.$$

If we apply now Lemma 4.5; then we obtain

$$(4.19) \quad Y = (U_N^\sigma)^* \bar{X}$$

for a  $\mathcal{B}$ -homomorphism  $\bar{X} : L^2(\sigma I) \rightarrow L^2(\sigma I)$ . Now  $\bar{X}F = \Xi F$  where  $\Xi$  is a measurable  $\sigma$ -a.e. bounded matrix function that satisfies  $\|\Xi\|_\infty \|\bar{X}\| = \|Y\|$ . Since by (4.18)  $\bar{X} P^\perp L^2(\sigma I) = U_N^\sigma Y P^\perp L^2(\sigma I) = 0$  we have

$$(4.20) \quad \Xi P^\perp = 0$$

which is equivalent to

$$(4.21) \quad \Xi = \Xi P.$$



Also since  $\bar{X} = U_{\mathbb{N}}^{\sigma}Y$  we have

$$\Xi L^2(\sigma I) \subset U_{\mathbb{N}}^{\sigma}L^2(\mathbb{N}) = QL^2(\sigma I),$$

which implies

$$(4.22) \quad Q^{\perp}\Xi = 0,$$

which is equivalent to

$$(4.23) \quad \Xi = Q\Xi$$

and (4.15) is proved. We note also that (4.15) implies the equality

$$(4.24) \quad \Xi P^{\perp} = Q^{\perp}\Xi.$$

Representation (4.16) follows now from (4.17), (4.19) and the formulas for  $U_{\mathbb{M}}^{\sigma}$  and  $(U_{\mathbb{N}}^{\sigma})^*$ .

We note for future reference that  $X^*: L^2(\mathbb{N}) \rightarrow L^2(\mathbb{M})$  is also a  $\mathcal{B}$ -homomorphism. In terms of the notation of the previous theorem we have the following corollary.

**COROLLARY 4.7.** *If  $X: L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  is the  $\mathcal{B}$ -homomorphism having the representation (4.16) with (4.15) satisfied then  $X^*: L^2(\mathbb{N}) \rightarrow L^2(\mathbb{M})$  is a  $\mathcal{B}$ -homomorphism having the representation*

$$(4.25) \quad X^*G = (U_{\mathbb{M}}^{\sigma})^*\bar{X}^*U_{\mathbb{N}}^{\sigma}G \quad \text{for } G \in L^2(\mathbb{N})$$

or more specifically

$$(4.26) \quad X^*G = H^{\#}P\Xi^*KG,$$

where

$$(4.27) \quad \Xi(\lambda)^* = \Xi(\lambda)^*Q(\lambda) = P(\lambda)\Xi(\lambda)^*$$

holds  $\sigma$ -a.e.

For the analysis of the deeper properties of intertwining operators we will introduce the several relevant notions of coprimeness. All definitions will be relative to a fixed positive scalar measure  $\sigma$ . A measurable projection valued  $n \times n$  matrix  $P$  will be called trivial with respect to  $\sigma$ , or  $\sigma$ -trivial, if  $P(\lambda) = I\sigma$ -a.e. Two measurable,  $n \times m$  and  $n \times l$  respectively, matrix functions  $A$  and  $B$  are called  $\sigma$ -left coprime if there exists no  $\sigma$ -nontrivial projection function  $P$  for which  $A = PA$  and  $B = PB$ . We denote the  $\sigma$ -left coprimeness of  $A$  and  $B$  by  $(A, B)_L^{\sigma} = I$ . Analogously we define  $\sigma$ -right coprimeness and denote it by  $(A, B)_R^{\sigma} = I$ . There is also a stronger notion of coprimeness. We say  $A$  and  $B$  are strongly  $\sigma$ -left coprime, and write  $[A, B]_L^{\sigma} = I$ , if there exists a  $\delta > 0$  such that for all  $\xi, \|\xi\| = 1$  we have

$$(4.28) \quad \|A(\lambda)^*\xi\| + \|B(\lambda)^*\xi\| \geq \delta, \quad \sigma\text{-a.e.}$$

Again the analogous notion of strong  $\sigma$ -right coprimeness is introduced in the same manner. The above definitions extend easily to the coprimeness of a finite number of matrix functions.

As expected the coprimeness relations are connected with the ideal structure in the algebra of bounded measurable functions.

**THEOREM 4.8.** (i) *Let  $A_1, \dots, A_p$  be bounded measurable  $n \times m_i$  matrix valued functions. Then there exist bounded measurable  $m_i \times n$  matrix valued functions  $B_i$  such that*

$$(4.29) \quad \sum_{i=1}^p A_i(\lambda)B_i(\lambda) = I, \quad \sigma\text{-a.e.}$$

if and only if

$$(4.30) \quad [A_1, \dots, A_p]_L^\sigma = I.$$

(ii) Let  $A_1, \dots, A_p$  be measurable  $m_i \times n$  matrix functions. Then there exist  $n \times m_i$  matrix functions  $B_i$  such that

$$(4.31) \quad \sum_{i=1}^p B_i(\lambda)A_i(\lambda) = I, \quad \sigma\text{-a.e.}$$

if and only if

$$(4.32) \quad [A_1, \dots, A_p]_R^\sigma = I.$$

*Proof.* Assume there exist  $B_i$  such that (4.29) holds. Taking adjoints and applying the resulting equality to a unit vector  $\xi$  we have

$$\xi = \sum B_i(\lambda)^*A_i(\lambda)^*\xi$$

and hence

$$\begin{aligned} 1 = \|\xi\| &\leq \sum \|B_i(\lambda)^*A_i(\lambda)^*\xi\| \\ &\leq \sum \|B_i(\lambda)^*\| \|A_i(\lambda)^*\xi\| \\ &\leq B \sum \|A_i(\lambda)^*\xi\|, \end{aligned}$$

where  $B = \max_i \|B_i(\lambda)^*\|$ . Equivalently we have

$$(4.33) \quad \sum \|A_i(\lambda)^*\xi\| \geq B^{-1},$$

that is  $[A_1, \dots, A_p]_L^\sigma = I$ .

Conversely assume  $A_1, \dots, A_p$  are strongly  $\sigma$ -left coprime. From (4.30) it follows that

$$(4.34) \quad \sum \|A_i(\lambda)^*\xi\|^2 \geq \delta^2$$

for some  $\delta > 0$  and all unit vectors  $\xi$ . Inequality (4.34) can be rewritten as  $\sum_i A_i(\lambda)A_i(\lambda)^* \geq \delta^2 I$ . Thus  $\sum_i A_i(\lambda)A_i(\lambda)^*$  is measurable and invertible in the algebra of all bounded measurable  $n \times n$  matrix functions. Define  $B_i$  by  $B_i(\lambda) = A_i(\lambda)^*(\sum_i A_j(\lambda)A_j(\lambda)^*)^{-1}$ . Then the  $B_i$  are bounded and measurable and (4.29) holds. Part (ii) follows by a simple duality argument.

The following corollary justifies the distinction between  $\sigma$ -left coprimeness and strong  $\sigma$ -left coprimeness.

**COROLLARY 4.9.** *If  $A_1, \dots, A_p$  are bounded measurable  $n \times m_i$  matrix valued functions then  $[A_1, \dots, A_p]_L^\sigma = I$  implies  $(A_1, \dots, A_p)_L^\sigma = I$ .*

*Proof.* Assume  $[A_1, \dots, A_p]_L^\sigma = I$ . Then there exist  $B_i$  such that  $\sum_i A_i B_i = I$ . From this it follows that  $A_1, \dots, A_p$  cannot have a common  $\sigma$ -nontrivial projection valued left factor. Thus  $\sigma$ -left coprimeness.

The various coprimeness relations provide the language in which to phrase the next result.

**THEOREM 4.10.** *Let  $X : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  be a  $\mathcal{B}$ -homomorphism having the representation (4.16) with relation (4.15) satisfied. Then*

(i)  *$X$  has dense range if and only if*

$$(4.35) \quad (\Xi, Q^\perp)_L^\sigma = I.$$

(ii)  $X$  is one-to-one if and only if

$$(4.36) \quad (\Xi, P^\perp)_R^\sigma = I.$$

(iii)  $X$  has a bounded right inverse if and only if

$$(4.37) \quad [\Xi, Q^\perp]_L^\sigma = I.$$

(iv)  $X$  has a bounded left inverse if and only if

$$(4.38) \quad [\Xi, P^\perp]_R^\sigma = I.$$

*Proof.* (i) The range of  $X$  is dense in  $L^2(\mathbb{N})$  if and only if the range of  $\bar{X}$  is dense in  $U_{\mathbb{N}}^\sigma L^2(\mathbb{N}) = QL^2(\sigma I)$ . This occurs if and only if the span of the two linear manifolds  $\{\Xi HF | F \in L^2(\mathbb{M})\}$  and  $Q^\perp L^2(\sigma I)$  is all of  $L^2(\sigma I)$ . Now  $\{\Xi HF | F \in L^2(\mathbb{M})\} = \Xi PL^2(\sigma I)$  and since  $\Xi P^\perp = Q^\perp \Xi$  it follows that  $\Xi P^\perp L^2(\sigma I) \subset Q^\perp L^2(\sigma I)$ . Hence  $X$  has dense range if and only if the span of  $\Xi L^2(\sigma I)$  and  $Q^\perp L^2(\sigma I)$  is  $L^2(\sigma I)$ . Since the span of two invariant subspaces is an invariant subspace we apply Theorem 4.1 on the characterization of invariant subspaces to obtain the result that

$$(4.39) \quad \Xi L^2(\sigma I) \vee Q^\perp L^2(\sigma I) = L^2(\sigma I)$$

if and only if (4.35) holds.

(ii) This follows from (i) by a duality argument.  $X$  is one-to-one if and only if  $X^* : L^2(\mathbb{N}) \rightarrow L^2(\mathbb{M})$  has dense range. Now  $X^*$  is given by (4.26) with relation (4.27) holding. By applying part (i)  $X^*$  has dense range if and only if

$$(4.40) \quad (\Xi^*, P^\perp)_L^\sigma = I$$

which is equivalent to (4.36).

(iii) Assume (4.37) holds. By Theorem 4.8 there exist matrix valued functions  $\theta$  and  $R$  such that

$$(4.41) \quad \Xi(\lambda)\theta(\lambda) \oplus Q^\perp(\lambda)R(\lambda) = I, \quad \sigma\text{-a.e.}$$

Define maps  $Y : L^2(\mathbb{N}) \rightarrow L^2(\mathbb{M})$  and  $\bar{Y} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  by

$$(4.42) \quad \bar{Y}F = \theta F \quad \text{for } F \in L^2(\sigma I)$$

and

$$(4.43) \quad YF = (U_{\mathbb{M}}^\sigma)^* \bar{Y} U_{\mathbb{N}}^\sigma F \quad \text{for } F \in L^2(\mathbb{N}).$$

Obviously  $Y$  and  $\bar{Y}$  are bounded linear operators. We claim  $XY = I$ . Let  $F \in L^2(\cdot)$ ; then

$$XYF = (U_{\mathbb{N}}^\sigma)^* \bar{X} U_{\mathbb{M}}^\sigma (U_{\mathbb{M}}^\sigma)^* \bar{Y} U_{\mathbb{N}}^\sigma F.$$

Since  $U_{\mathbb{M}}^\sigma$  is an isometry  $U_{\mathbb{M}}^\sigma (U_{\mathbb{M}}^\sigma)^*$  is the projection on the range of  $U_{\mathbb{M}}^\sigma$  which is just the multiplication by the projection valued function  $P$ . So

$$XYF = K^\# Q \Xi P \theta K F$$

and using the equality  $\Xi P = Q \Xi$  as well as  $Q^\perp = Q$  yields

$$XYF = K^\# Q \Xi \theta K F.$$

From (4.41) we have  $\Xi \theta = I - Q^\perp R$  and since  $Q Q^\perp = 0$  we have

$$XYF = K^\# Q K F = (U_{\mathbb{N}}^\sigma)^* U_{\mathbb{N}}^\sigma F = F.$$

To prove the necessity of the condition (4.37) for the existence of a bounded right inverse for  $X$  it suffices, by duality considerations, to prove the necessity of the

condition (4.38) for the existence of a bounded left inverse for  $X$ . Thus assume (4.38) is not satisfied. We will show the existence of a sequence of functions  $F_n$  in  $L^2(\mathbb{M})$  such that  $\lim \|F_n\| = 1$  and  $\lim \|XF_n\| = 0$ . This would imply the nonexistence of a bounded left inverse for  $X$ . Since (4.38) is not satisfied then for all  $n > 0$  there exists a unit vector  $\xi_n$  for which

$$(4.44) \quad \|\Xi(\lambda)\xi_n\| + \|P^\perp(\lambda)\xi_n\| < \frac{1}{n}$$

for all  $\lambda$  in a set  $\Lambda_n$  of positive  $\sigma$ -measure. Let  $\chi_{\Lambda_n}$  be the characteristic function of the set  $\Lambda_n$ ; then

$$\Psi_n(\lambda) = [\sigma(\Lambda_n)]^{-1/2} \chi_{\Lambda_n} \xi_n$$

is a function in  $L^2(\sigma I)$  of norm one. We decompose  $\Psi_n$  relative to the direct sum  $L^2(\sigma I) = PL^2(\sigma I) \oplus P^\perp L^2(\sigma I)$  to obtain  $\Psi_n = \Phi_n + \Gamma_n$  with

$$\Phi_n = [\sigma(\Lambda_n)]^{1/2} P \chi_{\Lambda_n} \xi_n$$

and

$$\Gamma_n = [\sigma(\Lambda_n)]^{-1/2} P^\perp \chi_{\Lambda_n} \xi_n.$$

Since  $\Phi_n \in PL^2(\sigma I) = U_{\mathbb{M}}^\sigma L^2(\mathbb{M})$  we have  $\Phi_n = U_{\mathbb{M}}^\sigma F_n$  for some  $F_n \in L^2(\mathbb{M})$  with  $\|F_n\| = \|\Phi_n\|$ . We note also that

$$\|\Gamma_n\| = [\sigma(\Lambda_n)]^{-1/2} \|P^\perp \chi_{\Lambda_n} \xi_n\| < \frac{1}{n}$$

and therefore

$$\lim \|F_n\|^2 = \lim [\|\Psi_n\|^2 - \|\Gamma_n\|^2] = 1.$$

We will show now that  $\lim \|XF_n\| = 0$ .

$$XF_n = (U_{\mathbb{N}}^\sigma)^* \bar{X} U_{\mathbb{M}}^\sigma F_n = (U_{\mathbb{N}}^\sigma)^* \bar{X} \Phi_n$$

and

$$\bar{X} \Phi_n = \bar{X}(\Psi_n - \Gamma_n) = [\sigma(\Lambda_n)]^{-1/2} \bar{\Xi} \chi_{\Lambda_n} \xi_n - [\sigma(\Lambda_n)]^{-1/2} \bar{\Xi} P^\perp \chi_{\Lambda_n} \xi_n.$$

The following estimate

$$\begin{aligned} \|XF_n\| &= \|(U_{\mathbb{N}}^\sigma)^* \bar{X} \Phi_n\| \leq \|\bar{X} \Phi_n\| \\ &\leq [\sigma(\Lambda_n)]^{-1/2} \|\chi_{\Lambda_n} \bar{\Xi} \xi_n\| + [\sigma(\Lambda_n)]^{-1/2} \|\bar{\Xi} \chi_{\Lambda_n} P^\perp \xi_n\| \\ &\leq [\sigma(\Lambda_n)]^{-1/2} \left\{ \int_{\Lambda_n} \|\bar{\Xi}(\lambda) \xi_n\|^2 d\sigma \right\}^{1/2} \\ &\quad + [\sigma(\Lambda_n)]^{-1/2} \|\bar{\Xi}\|_\infty \left\{ \int_{\Lambda_n} \|P^\perp(\lambda) \xi_n\|^2 d\sigma \right\}^{1/2} \\ &\leq (1 + \|\bar{\Xi}\|_\infty) \frac{1}{n} \end{aligned}$$

completes the proof.

**5. Dynamical systems with self-adjoint generators.** This section is devoted to the study of linear, time invariant systems with self-adjoint generators. Thus the object under consideration is a triple  $(A, B, C)$  where  $A$  is the infinitesimal generator of a

strongly continuous semigroup of operators  $T(t)$  in a Hilbert space  $H$ . This assumption is equivalent to the existence of some  $\omega$  such that  $\operatorname{Re} (Ax, x) \leq \omega \|x\|^2$ . We make the extra assumption that  $A$  is finitely generated and therefore, by § 2, we may without loss of generality assume that  $A$  is given in a spectral representation. Thus  $H = L^2(\mathbb{M})$  for some  $n \times n$  matrix measure  $\mathbb{M}$  and the semigroup  $T(t)$  acts by

$$(5.1) \quad (T(t)F)(\lambda) = e^{\lambda t}F(\lambda) \quad \text{for all } F \in L^2(\mathbb{M}).$$

The operators  $B : \mathbb{C}^m \rightarrow L^2(\mathbb{M})$  and  $C : L^2(\mathbb{M}) \rightarrow \mathbb{C}^p$  relate to the inputs and outputs of the system. The elements of  $L^2(\mathbb{M})$  are called *states* and  $L^2(\mathbb{M})$  is the *state space*.

Since  $B : \mathbb{C}^m \rightarrow L^2(\mathbb{M})$  then

$$(5.2) \quad (B\xi)(\lambda) = \mathcal{B}(\lambda)\xi \quad \text{for } \xi \in \mathbb{C}^m,$$

where  $B(\lambda)$  is some measurable  $n \times m$  matrix valued function. Similarly  $C^* : \mathbb{C}^p \rightarrow L^2(\mathbb{M})$  is given by

$$(5.3) \quad (C^*\eta)(\lambda) = C(\lambda)^*\eta \quad \text{for } \eta \in \mathbb{C}^p,$$

where  $C(\lambda)$  is some measurable  $p \times n$  matrix valued function.

As usual we say the system  $(A, B, C)$  is *controllable* if

$$(5.4) \quad \bigcap_{t \geq 0} \operatorname{Ker} B^*T(t) = \{0\}$$

and *observable* if

$$(5.5) \quad \bigcap_{t \geq 0} \operatorname{Ker} CT(t) = \{0\}.$$

Condition (5.4) is equivalent to the density, in  $L^2(\mathbb{M})$ , of the set of vectors  $\{T(t)B\xi \mid \xi \in \mathbb{C}^m, t \geq 0\} = \{e^{\lambda t}B(\lambda)\xi \mid \xi \in \mathbb{C}^m, t \geq 0\}$ . Similarly condition (5.5) is equivalent to the density of  $\{T(t)C^*\eta \mid \eta \in \mathbb{C}^p, t \geq 0\} = \{e^{\lambda t}C(\lambda)^*\eta \mid \eta \in \mathbb{C}^p, t \geq 0\}$ .

We define the *controllability* operator  $\mathcal{C}$  by

$$(5.6) \quad (\mathcal{C}u)(\lambda) = \int_{-\infty}^0 T(-t)Bu(t) dt$$

for all  $\mathbb{C}^m$ -valued functions  $u$  which are bounded and have compact support. Thus each such function  $u$  produces an element of  $L^2(\mathbb{M})$ , i.e., a state in the state space. Thus the controllability operator  $\mathcal{C}$  produces an input to state mapping. Now given a state  $F$  at time  $t = 0$  we can observe the output of the system when no additional inputs are applied. This gives us a  $\mathbb{C}^p$ -valued function on  $(0, \infty)$  defined by

$$(5.7) \quad (\mathcal{O}F)(t) = CT(t)F, \quad F \in L^2(\mathbb{M}).$$

We refer to  $\mathcal{O}$  as the *observability operator*. The product  $\mathcal{O}\mathcal{C}$  determines the input/output behavior of the system. It maps functions defined on  $(-\infty, 0)$  into functions defined on  $(0, \infty)$ . Clearly

$$(\mathcal{O}\mathcal{C}u)(t) = CT(t) \int_{-\infty}^0 T(-\tau)Bu(\tau) d\tau = \int_{-\infty}^0 CT(t-\tau)Bu(\tau) d\tau$$

or

$$(5.8) \quad (\mathcal{O}\mathcal{C}u)(t) = \int_{-\infty}^0 \gamma(t-\tau)u(\tau) d\tau,$$

where

$$(5.9) \quad \gamma(t) = CT(t)B, \quad t \geq 0.$$

Since the system is time invariant if we drop the assumption that  $u(t)$  is zero on  $(0, \infty)$  then the input/output relation is determined by

$$(5.10) \quad y(t) = \int_{-\infty}^t \gamma(t-\tau)u(\tau) d\tau.$$

The  $p \times m$  matrix valued function  $\gamma(t)$  is called the *weighting pattern* or the *impulse response* of the system. If  $(A, B, C)$  is a dynamical system for which (5.9) holds we say that  $(A, B, C)$  is a *realization* of the weighting pattern  $\gamma(t)$ , or that it realizes  $\gamma(t)$ .

Since we have a specific model for the system we can obtain more concrete information about the controllability and observability operators. From (5.6), (5.1) and (5.2) it follows that the controllability operator  $\mathcal{C}$  can be written as

$$(5.11) \quad (\mathcal{C}u)(\lambda) = \int_{-\infty}^0 e^{-\lambda t} B(x)u(t) dt = B(\lambda) \int_{-\infty}^0 e^{-\lambda t} u(t) dt.$$

But  $\int_{-\infty}^0 e^{-\lambda t} u(t) dt$  is just the Laplace transform  $\hat{u} = \mathcal{L}u$  of  $u$ . Thus we can define the operator  $\hat{\mathcal{C}}$  on the set of all Laplace transforms of permissible inputs by

$$(5.12) \quad \hat{\mathcal{C}}\hat{u} = \hat{\mathcal{C}}\hat{u}$$

or  $\hat{\mathcal{C}}\mathcal{L} = \mathcal{L}\mathcal{C}$  and

$$(5.13) \quad (\hat{\mathcal{C}}\hat{u})(\lambda) = B(\lambda)\hat{u}(\lambda).$$

Thus on its domain of definition  $\hat{\mathcal{C}}$  is a multiplication operator. Clearly if the domain of definition of  $\hat{\mathcal{C}}$  can be extended by continuity to a function space which is a  $\mathcal{B}$ -module then  $\hat{\mathcal{C}}$  becomes a  $\mathcal{B}$ -homomorphism.

An analogous situation holds for the observability operator, or rather its adjoint. Given a state  $F \in L^2(\mathbb{M})$ ,  $(\mathcal{O}F)(t)$  is a continuous  $\mathbb{C}^p$ -valued function on  $[0, \infty)$ . Let  $v$  be any  $\mathbb{C}^p$ -valued function for which the  $L^2(0, \infty)$  inner product  $(\mathcal{O}F, v)$  makes sense. Now from (5.3) we have

$$\begin{aligned} (CF, \eta) &= (F, C^*\eta) = \int (d\mathbb{M}F(\lambda), C(\lambda)^*\eta) \\ &= \int (C(\lambda) d\mathbb{M}F(\lambda), \eta) = \left( \int C(\lambda) d\mathbb{M}F(\lambda), \eta \right) \end{aligned}$$

or

$$(5.14) \quad CF = \int C(\lambda) d\mathbb{M}F(\lambda), \quad F \in L^2(\mathbb{M}),$$

and this in turn implies

$$(\mathcal{O}F)(t) = CT(t)F = \int C(\lambda) d\mathbb{M} e^{\lambda t} F(\lambda)$$

or

$$(5.15) \quad (\mathcal{O}F)(t) = \int e^{\lambda t} C(\lambda) d\mathbb{M}F(\lambda).$$

If we use the previously implied equality  $(\mathcal{O}F, v) = (F, \mathcal{O}^*v)$  we readily obtain the

representation

$$(5.16) \quad (\mathcal{O}^*v)(\lambda) = C(\lambda)^* \int_0^\infty e^{\lambda t} v(t) dt$$

or

$$(5.17) \quad (\hat{\mathcal{O}}^*\hat{v})(\lambda) = C(\lambda)^*\hat{v}(\lambda),$$

where now  $\hat{v}$  refers to the Laplace transform on  $[0, \infty)$ . Thus also  $\hat{\mathcal{O}}^*$  is potentially a  $\mathcal{B}$ -homomorphism if a right extension of its domain of definition can be found. Now since the controllability operator  $\mathcal{C}$  is determined by the matrix function  $B$  it is natural to characterize the controllability of the system  $(A, B, C)$  in terms of  $B$  and the matrix measure  $\mathbb{M}$ . Similarly for observability. To this end let  $\sigma$  be any positive measure for which  $\mathbb{M} \ll \sigma$ . Define operators  $B' : \mathbb{C}^m \rightarrow L^2(\sigma I)$  and  $C'^* : \mathbb{C}^p \rightarrow L^2(\sigma I)$  by

$$(5.18) \quad B'\xi = U_{\mathbb{M}}^\sigma B\xi \quad \text{and} \quad C'^*\eta = U_{\mathbb{M}}^\sigma C^*\eta.$$

The introduction of  $B'$  and  $C'$  removes the redundancies in the definition of  $B$  and  $C$ . We can use now the results of the previous section, particularly Theorem 4.10, to obtain the following generalization of a theorem of Fattorini [7], [4].

**THEOREM 5.1.** *Let  $(A, B, C)$  be the dynamical system in  $L^2(\mathbb{M})$  where  $A$  is the infinitesimal generator of the semigroup  $T(t)$  defined by (5.1),  $B : \mathbb{C}^m \rightarrow L^2(\mathbb{M})$  defined by (5.5) and  $C^* : \mathbb{C}^p \rightarrow L^2(\mathbb{M})$  is given by*

$$(5.19) \quad (C^*\eta)(\lambda) = C(\lambda)^*\eta \quad \text{for } \eta \in \mathbb{C}^p.$$

*Then the system  $(A, B, C)$  is controllable if and only if*

$$(5.20) \quad (B, P^{\perp})_L^\sigma = I$$

*and observable if and only if*

$$(5.21) \quad (C, P^{\perp})_R^\sigma = I.$$

To see how Fattorini's result can be derived from Theorem 5.1 we consider first the case of a scalar type matrix measure  $\mathbb{M} = \mu I$ . Obviously we can identify  $\sigma$  with  $\mu$ . The projection function  $P$  is identically equal to  $I$  and hence  $P^{\perp} = 0$ . Therefore  $(B, P^{\perp})_L^\sigma = I$  if and only if there exists no  $\mu$ -nontrivial projection valued function  $R$  such that  $B = RB$ . This is obviously the case if and only if  $B$  has full row rank  $\mu$ -a.e. Summarizing we obtain the following corollary.

**COROLLARY 5.2.** *Let  $\mathbb{M} = \mu I$  be an  $n \times n$  scalar type measure and let the dynamical system  $(A, B, C)$  be as in Theorem 5.1. Then  $(A, B, C)$  is controllable if and only if  $B$  has full row rank  $\mu$ -a.e. and observable if and only if  $C$  has full column rank  $\mu$ -a.e. This means*

$$(5.22) \quad \text{rank } B(\lambda) = n, \quad \mu\text{-a.e.}$$

and

$$(5.23) \quad \text{rank } \mathcal{C}(\lambda) = n, \quad \mu\text{-a.e.}$$

respectively.

Next suppose  $A$  is given in the canonical spectral representation (3.8) with  $\mathbb{N}$  being given by (3.7). Take  $\sigma = \nu_1 + \dots + \nu_p$ ; then if  $n_i$  is the Radon–Nikodym derivative of  $\nu_i$  with respect to  $\sigma$ , and  $E_i = \{\lambda \mid n_i(\lambda) \neq 0\}$  then  $\sigma$ -a.e. the projection function  $P$  for which

$U_{\mathbb{N}}^{\sigma}L^2(\mathbb{N}) = PL^2(\sigma I)$  is given by

$$P = \begin{pmatrix} \chi_{E_1} + \cdots + \chi_{E_p} & & \\ & \ddots & \\ & & \chi_{E_p} \end{pmatrix}$$

and hence

$$P^{-1} = \begin{pmatrix} 0 & & \\ & \chi_{E_1} & \\ & & \chi_{E_1} + \cdots + \chi_{E_{p-1}} \end{pmatrix}.$$

Therefore the condition  $(B, P^{-1})_{L}^{\sigma} = I$  is equivalent, by applying the previous corollary, to

$$(5.24) \quad \text{rank}(b_{ij}(\lambda)) = k, \quad i = 1, \dots, k, \quad j = 1, \dots, m,$$

$\nu_k$ -a.e. for all  $k = 1, \dots, p$ . Condition (5.24) is the essence of Fattorini’s result.

To get to the isomorphism theorem we have to introduce a stronger notion of controllability, in fact a whole class of controllability criterias. We saw that a priori the controllability operator  $\mathcal{C}$  defined by (5.4) had a domain of definition consisting of all  $\mathbb{C}^m$ -valued bounded measurable functions of compact support. Let now  $\sigma$  be a positive measure on  $(-\infty, 0]$ . We say the system  $(A, B, C)$  is  $\sigma$ -exactly controllable if  $\mathcal{C}$  can be extended by continuity to a bounded operator of  $L^2(\sigma I)$  onto  $L^2(\mathbb{M})$ . In an analogous way we define the notion of  $\sigma$ -exact observability.

If the system  $(A, B, C)$  is  $\sigma$ -exactly controllable then  $\mathcal{C}$  is a  $\mathcal{B}$ -homomorphism. In particular we can apply Lemma 4.5 to show the existence of a  $\mathcal{B}$ -homomorphism  $\tilde{\mathcal{C}} : L^2(\sigma I) \rightarrow L^2(\sigma I)$  for which  $\mathcal{C} = (U_{\mathbb{M}}^{\sigma})^* \tilde{\mathcal{C}}$ . If  $\rho$  is now any measure for which  $\sigma \ll \rho$  then  $\tilde{\mathcal{C}}$  can be lifted to a  $\mathcal{B}$ -homomorphism  $\tilde{\tilde{\mathcal{C}}} : L^2(\rho I) \rightarrow L^2(\rho I)$  for which

$$(5.25) \quad \tilde{\tilde{\mathcal{C}}} U_{\sigma}^{\rho} = U_{\sigma}^{\rho} \tilde{\mathcal{C}}.$$

It follows that  $(U_{\mathbb{M}}^{\rho})^* \tilde{\tilde{\mathcal{C}}}$  is a bounded extension of the controllability operator  $\mathcal{C}$  to a  $\mathcal{B}$ -homomorphism of  $L^2(\rho I)$  onto  $L^2(\mathbb{M})$ . Summarizing we have obtained the following.

**THEOREM 5.3.** *Let  $(A, B, C)$  be the dynamical system described in Theorem 5.1. Then if  $(A, B, C)$  is  $\sigma$ -exactly controllable and  $\sigma \ll \rho$  then  $(A, B, C)$  is  $\rho$ -exactly controllable.*

We would like to apply Theorem 4.10 to give a characterization of  $\sigma$ -exact controllability in terms of the matrix function  $B$  and, essentially, the measures  $\sigma I$  and  $\mathbb{M}$ . For this we would need the existence of an isometric embedding of  $L^2(\mathbb{M})$  into  $L^2(\sigma I)$ , and with this in mind we prove the following theorem which may be of interest in itself.

**THEOREM 5.4.** *Let  $A$  and  $A_1$  be two self-adjoint operators acting in the Hilbert spaces  $H$  and  $H_1$  respectively. Let  $X : H \rightarrow H_1$  intertwine  $A$  and  $A_1$ , a.e.  $XA = A_1X$ . Then*

- (i) *If  $X$  has range density in  $H$ , there exists a coisometry  $V$  such that  $VA = A_1V$ .*
- (ii) *If  $X$  is one-to-one there exists an isometry  $W$  such that  $WA = A_1W$ .*
- (iii) *If  $X$  is one-to-one and has range dense in  $H_1$  (in particular if  $X$  is boundedly invertible) then there exists a unitary  $U$  such that  $UA = A_1U$ .*

*Proof.* From  $XA = A_1X$  it follows by taking adjoints that  $AX^* = X^*A$ , and hence  $AX^*X = X^*A_1X = X^*XA$  or  $A(X^*X) = (X^*X)A$  and analogously  $A_1(XX^*) = (XX^*)A_1$ . By a standard approximation argument it follows that

$$(5.26) \quad A(X^*X)^{1/2} = (X^*X)^{1/2}A$$



and

$$(5.27) \quad A_1(XX^*)^{1/2} = (XX^*)^{1/2}A_1.$$

Now assume  $X$  has range dense in  $H_1$ . Since  $\{0\} = \{\text{Range } X\}^\perp = \text{Ker } X^* = \text{Ker } (XX^*)^{1/2} = \{\text{Range } (XX^*)^{1/2}\}^\perp$  it follows that also  $(XX^*)^{1/2}$  has range dense in  $H_1$ . From the equality  $\|X^*y\| = \|(XX^*)^{1/2}y\|$  it follows that if we define  $V$  by

$$VX^*y = (XX^*)^{1/2}y,$$

then  $V$  can be extended by continuity for an isometry from  $\overline{\text{Range } X^*}$  onto  $H_1$ . Extend  $V$  to all of  $H$  by defining  $V|_{\text{Ker } X} = 0$  and  $V$  becomes a coisometry satisfying  $VX^* = (XX^*)^{1/2}$ . By our assumption  $(XX^*)^{1/2}$  has dense range; hence  $(XX^*)^{-1/2}$  is a closed densely defined operator. Thus  $VX^*(XX^*)^{-1/2}y = y$  for all  $y$  in range  $(XX^*)^{1/2}$ . Since  $V$  is isometric on  $\text{Range } X^*$  we have  $X^*(XX^*)^{-1/2}$  is isometric on its domain of definition, hence extendable by continuity to an isometry on  $H_1$  which has to coincide with  $V^*$ . So we have

$$(5.28) \quad V = (XX^*)^{-1/2}X.$$

Since from (5.18) it follows that  $A_1(XX^*)^{-1/2} = (XX^*)^{-1/2}A_1$  we have

$$VA = (XX^*)^{-1/2}XA = (XX^*)^{-1/2}A_1X = A_1(XX^*)^{-1/2}X = A_1V$$

which proves (i). Part (ii) follows by duality considerations. Finally if  $X$  is one-to-one and has dense range then both  $X^*(XX^*)^{-1/2}$  and  $X(X^*X)^{-1/2}$  are isometric. Now from the equality  $X(X^*X) = (XX^*)X$  it follows that  $X(X^*X)^{1/2} = (XX^*)^{1/2}X$  and hence that  $(XX^*)^{-1/2}X = X(X^*X)^{-1/2}$ . This means that  $V$  given by (5.28) is also isometric and therefore unitary.

A weaker statement of (iii), that for normal operators, can be found in [22, p. 316]. As a corollary we prove the following theorem which is a converse of Lemma 2.2.

**THEOREM 5.5.** *Let  $\mathbb{M}$  and  $\mathbb{N}$  be two positive matrix measures and let  $X : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$  be a  $\mathcal{B}$ -homomorphism. Then*

- (i) *If  $X$  has dense range then  $\mathbb{N}|\mathbb{M}$ .*
- (ii) *If  $X$  is one-to-one then  $\mathbb{M}|\mathbb{N}$ .*

*Proof.* (ii) Applying the previous theorem we deduce the existence of an isometric, and easily checked  $\mathcal{B}$ -homomorphism,  $U : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{N})$ . Let  $E$  be any Borel set with compact closure,  $\chi_E$  its characteristic function  $\xi \in \mathbb{C}^n$ . Then  $\chi_E\xi$  belongs to any  $L^2(\mathbb{M})$  space. Since  $U$  is a  $\mathcal{B}$ -homomorphism  $(U(\chi_E\xi))(\lambda) = \chi_E(U\xi)(\lambda) = \chi_EJ(\lambda)\xi$  for some measurable matrix function  $J$ . Since  $U$  is isometric we have

$$\int (d\mathbb{N}_{\chi_EJ(\lambda)\xi}, J(\lambda)\xi) = \int (d\mathbb{M}_{\chi_E\xi}, \chi_E\xi)$$

or

$$(5.29) \quad \int_E (J(\lambda)^* d\mathbb{N}J(\lambda)\xi, \xi) = \int_E (d\mathbb{M}\xi, \xi).$$

Since (5.20) holds for arbitrary  $\xi \in \mathbb{C}^n$  and Borel sets  $E$  then

$$(5.30) \quad d\mathbb{M} = J^* d\mathbb{N}J$$

as  $\mathbb{M}|\mathbb{N}$ . Part (i) follows by duality.

We have now the trivial corollary of the above theorem which follows from the definition of  $\sigma$ -exact controllability. An analogous result holds for  $\sigma$ -exact observability.

**COROLLARY 5.6.** *Let  $(A, B, C)$  be the dynamical system described in Theorem 5.1. Then if  $(A, B, C)$  is  $\sigma$ -exactly controllable then  $\mathbb{M}|\sigma I$ .*

Applying now Theorem 4.10 together with the above corollary we can characterize  $\sigma$ -exact controllability in terms of strong  $\sigma$ -coprimeness relations.

**THEOREM 5.7.** *Let  $(A, B, C)$  be the dynamical system in  $L^2(\mathbb{M})$  where  $A$  is the infinitesimal generator of the semigroup  $T(t)$  defined by (5.1),  $B : \mathbb{C}^m \rightarrow L^2(\mathbb{M})$  defined by*

$$(5.31) \quad (B\xi)(\lambda) = B(\lambda)\xi$$

and  $C^* : \mathbb{C}^p \rightarrow L^2(\mathbb{M})$  is given by

$$(5.32) \quad (C^*\eta)(\lambda) = C(\lambda)^*\eta.$$

(i) *If  $\mathbb{M}|\sigma I, B$  is measurable and bounded  $\sigma$ -a.e. and  $P$  the projection valued function corresponding to  $PL^2(\sigma I) = U_{\mathbb{M}}^{\sigma}L^2(\mathbb{M})$  and*

$$(5.33) \quad [B, P^{\perp}]_L^{\sigma} = I$$

*holds then  $(A, B, C)$  is a  $\sigma$ -exactly controllable system. Conversely if  $(A, B, C)$  is a  $\sigma$ -exactly controllable system then  $\mathbb{M}|\sigma I$  and there exists a measurable  $\sigma$ -a.e. bounded function  $B(\lambda)$  such that (5.31) and (5.33) hold.*

(ii) *If  $\mathbb{M}|\sigma I, C$  is measurable and bounded  $\sigma$ -a.e. and  $P$  the projection valued function corresponding to  $PL^2(\sigma I) = U_{\mathbb{M}}^{\sigma}L^2(\mathbb{M})$  and*

$$(5.34) \quad [C, P^{\perp}]_R^{\sigma} = I$$

*holds then  $(A, B, C)$  is a  $\sigma$ -exactly observable system. Conversely if  $(A, B, C)$  is a  $\sigma$ -exactly observable system then  $\mathbb{M}|\sigma I$  and there exists a measurable  $\sigma$ -a.e. bounded function  $C(\lambda)$  such that (5.32) and (5.34) hold.*

In conclusion we prove an isomorphism theorem of a type different from the one proved in (4). There are no symmetry requirements on the systems involved but we impose a stronger controllability assumption. This is in line with Helton's work [16]. Two systems  $(A, B, C)$  and  $(A_1, B_1, C_1)$  with state spaces  $H$  and  $H_1$  respectively are called similar if there exists an invertible linear map  $X : H \rightarrow H_1$  such that

$$(5.35) \quad B_1 = XB, \quad XT(t) = T_1(t)X, \quad C_1X = C$$

hold. Here  $T(t)$  and  $T_1(t)$  are the semigroups generated by  $A$  and  $A_1$  respectively. If  $A$  and  $A_1$  are bounded then the condition  $XT(t) = T_1(t)X$  can be replaced by  $XA = A_1X$ .

**THEOREM 5.8.** *Let  $(A, B, C)$  and  $(A_1, B_1, C_1)$  be two dynamical systems of the type described in Theorem 5.7 which realize the same weighting pattern. If both systems are observable and  $\sigma$ -exactly controllable then the two systems are similar.*

*Proof.* Let  $\mathcal{C}, \mathcal{C}_1$  and  $\mathcal{O}, \mathcal{O}_1$  be the respective controllability and observability operators of the two systems. Since they both realize the same weighting pattern we have  $\mathcal{O}\mathcal{C} = \mathcal{O}_1\mathcal{C}_1$ . By the assumption of observability we have  $\text{Ker } \mathcal{O} = \{0\}$  and  $\text{Ker } \mathcal{O}_1 = \{0\}$ . This implies that  $\text{Ker } \mathcal{C} = \text{Ker } \mathcal{C}_1$  and hence also that  $\text{Ker } \hat{\mathcal{C}} = \text{Ker } \hat{\mathcal{C}}_1$ .

Define a map  $X : L^2(\mathbb{M}) \rightarrow L^2(\mathbb{M}_1)$  by

$$(5.36) \quad X\hat{\mathcal{C}} = \hat{\mathcal{C}}_1.$$

If  $\hat{\mathcal{C}}^{\#}$  is the pseudoinverse of  $\hat{\mathcal{C}}$  defined and bounded on all of  $L^2(\mathbb{M})$  as  $\hat{\mathcal{C}}$  is onto  $L^2(\mathbb{M})$  then  $X = \hat{\mathcal{C}}_1\hat{\mathcal{C}}^{\#}$  and clearly  $X$  is a  $\mathcal{B}$ -isomorphism of  $L^2(\mathbb{M})$  onto  $L^2(\mathbb{M}_1)$ .

From the definition of  $X$  we have  $XB(\lambda)\hat{u}(\lambda) = B_1(\lambda)\hat{u}(\lambda)$  for all  $\hat{u} \in L^2(\sigma I)$  which implies the equality  $XB = B_1$ . Since  $X$  is a  $\mathcal{B}$ -homomorphism and the semigroups  $T(t)$  and  $T_1(t)$  and by multiplication by  $e^{\lambda t}$  in the respective spaces we obtain  $XT(t) =$

$T_1(t)X$ . Finally we note that from  $\mathcal{O}\mathcal{C} = \mathcal{O}_1\mathcal{C}_1$  and (5.36) it follows that  $\mathcal{O}_1\mathcal{C}_1 = \mathcal{O}_1X\mathcal{C} = \mathcal{O}\mathcal{C}$  which implies, as  $\mathcal{C}$  is onto, that  $\mathcal{O}_1X = \mathcal{O}$ .

Now for  $F \in L^2(\mathbb{M})$ ,  $\mathcal{O}F = CT(t)F$  and hence  $C_1T_1(t)X = CT(t)$ . Using the equality  $XT(t) = T_1(t)X$  and evaluating at  $t = 0$  we have  $C_1X = C$  and thus (5.35) holds and the similarity of the two systems is proved.

## REFERENCES

- [1] I. N. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, F. Ungar, New York, 1961.
- [2] J. S. BARAS, *Algebraic structure of infinite dimensional linear systems in Hilbert space*, Mathematical Systems Theory, Udine 1975, Springer 1976, pp. 193–203.
- [3] R. BEALS, *Topics in Operator Theory*, University of Chicago Press, Chicago, 1972.
- [4] R. W. BROCKETT AND P. A. FUHRMANN, *Normal symmetric dynamical systems*, SIAM J. Control, 14 (1976), pp. 107–119.
- [5] A. BROWN, *A version of multiplicity theory*, Topics in Operator Theory, C. Pearcy, ed., American Mathematical Society, Providence RI, 1974.
- [6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Vols. 1, 2, Interscience, New York, 1957, 1963.
- [7] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
- [8] P. A. FUHRMANN, *On the corona theorem and its applications to spectral problems in Hilbert space*, Trans. Amer. Math. Soc., 132 (1968), pp. 55–66.
- [9] ———, *A functional calculus in Hilbert space based on operator valued analytic functions*, Israel J. Math., 6 (1968), pp. 267–278.
- [10] ———, *Algebraic system theory: An analyst's point of view*, J. Franklin Inst., 301 (1976), pp. 521–540.
- [11] ———, *Algebraic ideas in infinite dimensional system theory*, Mathematical Systems Theory, Udine 1975, Springer, 1976, pp. 237–251.
- [12] P. R. HALMOS, *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, Chelsea, New York, 1951.
- [13] M. L. J. HAUTUS AND M. HEYMAN, *Linear feedback—An algebraic approach*, SIAM J. Control Optim., 16 (1978), pp. 83–105.
- [14] E. HELLINGER, *Die Orthogonalinvarianten Quadratischer Formen*, Inaugural—dissertation, Göttingen, 1907.
- [15] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.
- [16] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Functional Analysis, 16 (1974), pp. 15–38.
- [17] R. E. KALMAN, *Lectures on Controllability and Observability*, CIME summer 1968, Cremonese, Roma, 1969.
- [18] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [19] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [20] E. NELSON, *Topics in Dynamics I: Flows*, Mathematical Notes, Princeton University Press, Princeton, NJ, 1969.
- [21] A. I. PLESSNER, *Spectral Theory of Linear Operators*, vol. 1, 2, F. Ungar, New York, 1969.
- [22] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [23] E. D. SONTAG, *On linear systems and noncommutative rings*, Math. Systems Theory, 9 (1976), 327–344.
- [24] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
- [25] ———, *On the structure of intertwining operators*, Acta Sci. Math. (Szeged), 35 (1973), pp. 225–253.

## ASYMPTOTIC APPROXIMATIONS AND EXTENSION OF TIME-SCALES\*

JAN A. SANDERS†

**Abstract.** It is shown how to obtain  $O(\varepsilon)$  (or any higher order)-approximations to the solutions of the differential equation

$$\begin{aligned} \dot{\phi} &= 1 + \sum_{p=1}^P \varepsilon^p X^p(\phi, x), & \phi \in S^1, \\ \dot{x} &= \sum_{p=1}^P \varepsilon^p Y^p(\phi, x), & x \in D \subset \mathbb{R}^n, \\ & & \varepsilon \in (0, \varepsilon_0] \end{aligned}$$

in such a way that they are valid on the interval  $0 \leq \varepsilon^{\tilde{N}} t \leq L$ , with  $\tilde{N} \in \mathbb{N}$  arbitrary and  $L$  an  $\varepsilon$ -independent constant under the condition that the averaged equation has an attracting nondegenerate limit-cycle.

The proof uses higher order averaging techniques and the Sanchez–Palencia contraction argument, together with Gronwall-estimates. In fact, for the  $x$ -component the approximation is uniformly valid on  $[0, \infty)$ .

An application to the Van der Pol-oscillator is given, extending the usual interval of length  $1/\varepsilon$  to the interval  $0 \leq \varepsilon^2 t \leq L$ , with  $O(\varepsilon)$ -accuracy. It turns out that this approximation is hardly more complicated than the usual one.

**Introduction.** The aim of this article is to show how to obtain  $O(\varepsilon)$  (or higher order)-approximations to the solutions of the differential equation:

$$\begin{aligned} \dot{\phi} &= 1 + \sum_{p=1}^P \varepsilon^p X^p(\phi, x), & \phi \in S^1, \\ & & \varepsilon \in (0, \varepsilon_0], \\ \dot{x} &= \sum_{p=1}^P \varepsilon^p Y^p(\phi, x), & x \in D_0 \subset \mathbb{R}^n \end{aligned}$$

in such a way that they are valid on the interval  $0 \leq \varepsilon^{\tilde{N}} t \leq L$ , with  $\tilde{N} \in \mathbb{N}$  arbitrary and  $L$  an  $\varepsilon$ -independent constant if the higher-order-averaged equation has an attracting nondegenerate limit-cycle, (that is, all eigenvalues of the “normal form” of the system are negative and can be strictly ordered; if  $\lambda_1, \dots, \lambda_n$  are the eigenvalues, then one has  $\lambda_n < \dots < \lambda_1 < 0$ .) It follows from the proof of Theorems 3 and 4 that the approximations of the  $x$ -component of the solution have uniform validity on  $[0, \infty)$ . An application of this theory can be found in Sanders (1978b), where an  $O(\varepsilon)$ -approximation to the solutions of the Van der Pol-equation has been derived, valid on  $0 \leq \varepsilon^2 t \leq L$ .

It can also be used to give asymptotic estimates for Hopf-bifurcation problems.

Averaging consists of two parts: First a transformation is found which “connects” the original equation (4.1) with its “almost normal form,” i.e., an equation which is the sum of its normal form ( $\phi$ -independent) and arbitrary small remainder terms. Secondly, one estimates the difference between the solutions of the “almost normal form” and of the normal form equation.

While the first step renders uniformly valid (i.e., on  $0 \leq t \leq \infty$ ) estimates, the second step, in general, introduces exponential growth terms in our error-estimate.

If, however, we assume contraction of the flow in the  $x$ -component of the normal form equation, then we can use this contraction to compensate for the error-propagation which leads to these exponential growth terms (Theorem 3).

This idea is due to Sanchez–Palencia (1975).

\* Received by the editors November 9, 1978, and in final revised form October 16, 1979.

† Wiskundig Seminarium, Vrije Universiteit, De Boelelaan 1081, 1007 MC Amsterdam, the Netherlands.

It is then a simple matter to impose sufficient conditions on a stationary  $x$ -value (i.e., a limit-cycle in  $(\phi, x)$ -coordinates) to assure contraction in a certain neighborhood of magnitude  $O_S(1)$  (Theorem 4, using Lemma 2 and 3).

The organization of this article is as follows: In §§ 1–3, we sketch the theory and some of the proofs for the simple problem

$$\dot{x} = \varepsilon g(t, x)$$

with  $g$  periodic in  $t$ , using first-order averaging. This is done to clarify the theory and proofs given in §§ 4 and 5. These are, however, independent of the earlier sections in the strict mathematical sense.

There is some overlap between the two parts, but, since this makes it possible to read them independently, we hope that it will prove to be only a minor annoyance for the reader to meet the same definition or argument twice.

Before we start, there is one remark which has to be made. One of the most natural ideas, and probably the first most people get when considering the time-scale extension problem, is to use nonlinear variation of constants. Since, after the usual transformations, the system seems to change on a longer time-scale, it does seem to be only a straightforward exercise in estimation theory. This, however, is not the case: one needs to have estimates on both the transformation and the inverse of its derivative. This makes application to any but the most trivial example impossible. An illustration of these thoughts can be found in Persek and Hoppensteadt (1978). One should also note in this context that the proof in Verhulst (1975), where the nonlinear variation of constants has been used, can be considerably simplified by using the theory to be developed here.

**1. The basic perturbation theorem.** The asymptotic theory of initial value problems in ordinary differential equations consists largely of formal results.

In this paper, we will be concerned with the explicit statement and proof of the validity of one of the many formal methods, averaging (of vector-fields) in the special case where the averaged vector field has a nondegenerate attracting limit-cycle.

The theory developed can be applied directly to such problems as the Van der Pol-oscillator and the Hopf bifurcation.

Before stating the well-known basic perturbation theorem, we introduce some old and new notation.

Let  $M$  be a manifold, and  $X_\varepsilon : M \rightarrow TM$ ,  $\varepsilon \in (0, \varepsilon_0]$ , a one-parameter family of vectorfields.

$I \subset \mathbb{R}$  is some interval.

We call the vector field morphism  $\phi_\varepsilon$ , defined by

$$(1.1) \quad \begin{array}{ccc} & TI & TM \\ \phi_\varepsilon : & \uparrow & \uparrow \\ & \iota & x_\varepsilon \\ & I & M \end{array}$$

a solution of  $X_\varepsilon$ , where  $\iota$  is the natural section  $\iota : t \mapsto (t, 1)$ . As usual  $\phi_\varepsilon$  consists of a map  $\phi_\varepsilon : I \rightarrow M$ , the solution curve, such that

$$(1.2) \quad \begin{array}{ccc} & TI & TM \\ & \uparrow & \uparrow \\ & \iota & x_\varepsilon \\ & I & M \end{array} \quad \begin{array}{ccc} & \xrightarrow{\phi_\varepsilon} & \\ & & \\ \xrightarrow{\phi_\varepsilon} & & \end{array}$$

commutes. Or, we solved the equation

$$\dot{\phi}_\varepsilon(t) = X_\varepsilon(\phi_\varepsilon(t)).$$

If for some  $\tau \in I$ , the additional requirement  $\phi_\varepsilon(\tau) = m \in M$  has been fulfilled, we have solved an initial value problem.

We say that for two curves in two manifolds  $M$  and  $N$ , the following diagram  $\varepsilon$ -commutes iff we have the estimate

$$(1.3) \quad \begin{array}{ccc} M & \xrightarrow{H} & N \\ & \searrow \phi_\varepsilon & \nearrow \psi_\varepsilon \\ & I & \end{array} \quad \begin{array}{c} * \\ \bullet \end{array}$$

$$\|H(\phi_\varepsilon(t)) - \psi_\varepsilon(t)\| = O(\varepsilon) \quad \forall t \in I,$$

where the norm is induced by  $\varepsilon$ -independent local charts.

If we require all manifolds to be compact and all maps differentiable, then we may chase diagrams at will. That is, if we have two  $\varepsilon$ -commuting diagrams,

$$(1.4a) \quad \begin{array}{ccc} M & \xrightarrow{H'H} & P \\ & \searrow \phi_\varepsilon & \nearrow \chi_\varepsilon \\ & I & \end{array} \quad \begin{array}{c} * \\ \bullet \end{array} \quad \begin{array}{ccc} N & \xrightarrow{H'} & P \\ & \searrow \psi_\varepsilon & \nearrow \chi_\varepsilon \\ & I & \end{array} \quad \begin{array}{c} * \\ \bullet \end{array}$$

then we may conclude that

$$(1.4b) \quad \begin{array}{ccc} M & \xrightarrow{H} & N \\ & \searrow \phi_\varepsilon & \nearrow \psi_\varepsilon \\ & I & \end{array} \quad \begin{array}{c} * \\ \bullet \end{array}$$

is  $\varepsilon$ -commutative.

**THEOREM 1.** *Let  $X_\varepsilon$  be of the following form*

$$(1.5)(\varepsilon) \quad \dot{x} = f(x) + \varepsilon R(t, x; \varepsilon), \quad \phi_\varepsilon(0) = x_0^\varepsilon \in D \subset \mathbb{R}^n.$$

*There exists some  $L > 0$  such that  $\phi_0$ , the solution of  $X_0$  with initial value  $x_0^0$ , exists on  $I = [0, L]$  and stays away from the boundary  $\partial D$  of  $D$ . Take  $L$  maximal but  $O(1)$  for  $\varepsilon > 0$ . If  $\|X_0^\varepsilon - X_0^0\| = O(\varepsilon)$ ,  $f \in C^1(D)$  and  $\sup_{t \in [0, L]} \sup_{x \in D} \|R(t, x; \varepsilon)\| \leq C$ , then there exists also a solution  $\phi_\varepsilon$  of  $X_\varepsilon$  with  $\phi_\varepsilon(0) = X_0^\varepsilon$ , and we have the following  $\varepsilon$ -commuting diagram*

$$(1.6) \quad \begin{array}{ccc} D & \xrightarrow{\text{id}_D} & D \\ & \searrow \phi_\varepsilon & \nearrow \phi_0 \\ & I & \end{array} \quad \begin{array}{c} * \\ \bullet \end{array}$$

( $\text{id}_D = \text{identity on } D$ ).

*Proof.* We shall need the following lemma, the proof of which has been given as an exercise in Coddington and Levinson (1955).

LEMMA 1 (Gronwall). *Let  $\psi(t) \geq 0, \forall t \in [0, L]$ , and suppose*

$$(1.7) \quad u(t) \leq \phi'(t) + \int_0^t \psi(\tau)u(\tau) d\tau.$$

Then

$$(1.8) \quad u(t) \leq \phi(0) \exp\left(\int_0^t \psi(\tau) d\tau\right) + \int_0^t \phi'(\tau) \exp\left(\int_\tau^t \psi(\tau') d\tau'\right) d\tau. \quad \square$$

We write the differential equations (1.5)( $\varepsilon$ ) and (1.5)(0) in integral form:

$$(1.9) \quad \begin{aligned} \phi_\varepsilon(t) &= X_0^\varepsilon + \int_0^t f(\phi_\varepsilon(\tau)) d\tau + \varepsilon \int_0^t R(\tau, \phi_\varepsilon(\tau); \varepsilon) d\tau, \\ \phi_0(t) &= x_0^0 + \int_0^t f(\phi_0(\tau)) d\tau \end{aligned}$$

( $\phi_\varepsilon$  does exist locally; if we can show that it  $\varepsilon$ -commutes with  $\phi_0$ , then it cannot reach  $\partial D$ : it has to exist on  $[0, L]$ .)

$$(1.10) \quad \begin{aligned} \|\phi_\varepsilon(t) - \phi_0(t)\| &\leq \|X_0^\varepsilon - X_0^0\| + \sup_{\xi \in D} \|\nabla f(\xi)\| \int_0^t \|\phi_\varepsilon(\tau) - \phi_0(\tau)\| d\tau \\ &\quad + \varepsilon t \sup_{\tau \in [0, t]} \sup_{\xi \in D} \|R(\tau, \xi; \varepsilon)\| \\ &\leq \|X_0^\varepsilon - X_0^0\| + C' \int_0^t \|\phi_\varepsilon(\tau) - \phi_0(\tau)\| d\tau + C\varepsilon t. \end{aligned}$$

Applying Gronwall's lemma, this results in

$$\|\phi_\varepsilon(t) - \phi_0(t)\| \leq \|X_0^\varepsilon - X_0^0\| e^{C't} + C\varepsilon \int_0^t e^{C'(t-\tau)} d\tau.$$

This last expression is  $O(\varepsilon)$  on  $[0, L]$ .

The occurrence of exponential terms in the final estimate implies that one has to think of something new in order to give an extended theorem, i.e., on  $[0, L/\varepsilon]$  or  $[0, \infty]$  for example.

If we assume one extra condition, then there is in fact a way out: the Sanchez-Palencia trick.

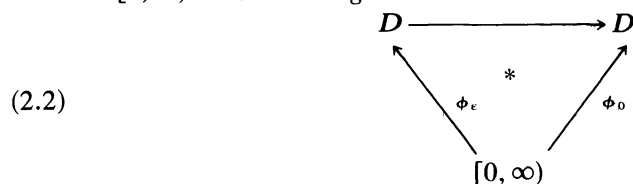
We will give a sketch of the argument here, and bother with the technical details later.

**2. A contraction theorem.**

THEOREM 2. *Consider the differential equation (1.5)( $\varepsilon$ ). Suppose that any two solutions  $\phi_0^1$  and  $\phi_0^2$  of (1.5)(0) allow the following contraction estimate:*

$$(2.1) \quad \|\phi_0^1(t+L) - \phi_0^2(t+L)\| \leq k(L)\|\phi_0^1(t) - \phi_0^2(t)\|$$

with  $k(L) < 1$  and (clearly)  $\varepsilon$ -independent. Suppose that for some fixed initial value,  $\phi_0$  exists on  $[0, \infty)$ . Then the diagram



is  $\varepsilon$ -commutative under the assumptions of Theorem 1, i.e.,  $\phi_0$  is a uniform  $O(\varepsilon)$ -approximation to  $\phi_\varepsilon$ , and  $\phi_\varepsilon$  exists on  $[0, \infty)$ .

*Sketch of the proof.* Let, for some  $C$  to be defined later,  $\tau$  be such that

$$(2.3) \quad \sup_{t \in [0, \tau]} \|\phi_\varepsilon(t) - \phi_0(t)\| \leq C\varepsilon.$$

Define  $\phi_0^2$  by  $\phi_0^2(\tau) = \phi_\varepsilon(\tau)$  and let  $\phi_0^1 = \phi_0$ . Then it follows from Theorem 1 that there exists for every  $L$  some  $C'$  such that:

$$(2.4) \quad \begin{aligned} \sup_{t \in [\tau, \tau+L]} \|\phi_\varepsilon(t) - \phi_0(t)\| &\leq \sup_{t \in [\tau, \tau+L]} \{\|\phi_\varepsilon(t) - \phi_0^2(t)\| + \|\phi_0^1(t) - \phi_0^2(t)\|\} \\ &\leq C'\varepsilon + \|\phi_0^1(t_+) - \phi_0^2(t_+)\|_{t_+ \in [\tau, \tau+L]} \\ &= C'\varepsilon + \|\phi_0^1(t_- + L) - \phi_0^2(t_- + L)\|_{t_- \in [\tau-L, \tau]} \\ &\leq C'\varepsilon + k(L)\|\phi_0^1(t_-) - \phi_0^2(t_-)\| \\ &\leq C'\varepsilon + k(L) \sup_{t \in [\tau-L, \tau]} \|\phi_0^1(t) - \phi_0^2(t)\| \\ &\leq C'\varepsilon + k(L) \sup_{t \in [\tau-L, \tau]} \{\|\phi_\varepsilon(t) - \phi_0^1(t)\| + \|\phi_\varepsilon(t) - \phi_0^2(t)\|\} \\ &\leq C'\varepsilon + k(L)(C\varepsilon + C'\varepsilon) = \{(1+k)C' + kC\}\varepsilon. \end{aligned}$$

If we take  $C = (1+k)/(1-k)C'$ , then we have extended the original estimate with an  $\varepsilon$ -independent interval without changing the constant in the estimate. If we assume  $\tau$  to be maximal, then this assumption is contradicted by our extension, and it follows that the estimate is valid on  $[0, \infty)$ .  $\square$

The contraction estimate (2.1) might be difficult to obtain in practice, but there is an important class of problems where it is always valid: if the unperturbed ( $\varepsilon = 0$ ) problem has an attracting stationary point then the estimate is valid in a neighborhood of this point. The exact results are formulated in § 5.

**3. First-order averaging.** The reader may have been thinking: If I can get my problem in the form (1.5)( $\varepsilon$ ) then I believe all estimates from there, but how do I get there in the first place?

There is, however, a class of nontrivial problems that can be brought into this form: Consider the differential equation

$$(3.1) \quad \dot{x} = \varepsilon g(t, x), \quad x \in D \subset \mathbb{R}^n$$

with  $g$  periodic (with period 1, say) in the time-variable. We write this as follows:

$$(3.2) \quad \begin{aligned} \dot{x} &= \varepsilon g(\tau, x), \\ \dot{\tau} &= 1, \end{aligned}$$

and consider a transformation

$$(3.3) \quad \Phi^\varepsilon : S^1 \times D^* \rightarrow S^1 \times D$$

defined by

$$(3.4) \quad \Phi^\varepsilon : \begin{pmatrix} x \\ \tau \end{pmatrix} \rightarrow \begin{pmatrix} x + \varepsilon u(\tau, x) \\ \tau \end{pmatrix}.$$

We want to take  $\Phi^\varepsilon$  in such a way that the vector field

$$(3.5) \quad \dot{x} = \varepsilon \bar{g}(x) + \varepsilon^2 R(t, x; \varepsilon), \quad \dot{\tau} = 1,$$

with  $\bar{g}$  and  $R$  to be defined next, is carried into (3.2); i.e., if we denote

$$(3.6) \quad \nu_\varepsilon = \varepsilon g \frac{\partial}{\partial x} + \frac{\partial}{\partial \tau}, \quad \bar{\nu}_\varepsilon = \varepsilon \bar{g} \frac{\partial}{\partial x} + \frac{\partial}{\partial \tau},$$



then the diagram

$$(3.7) \quad \begin{array}{ccc} T(S^1 \times D^*) & \xrightarrow{T\Phi_\varepsilon} & T(S^1 \times D) \\ \uparrow \bar{\nu}_\varepsilon & & \uparrow \nu_\varepsilon \\ S^1 \times D^* & \xrightarrow{\Phi_\varepsilon} & S^1 \times D \end{array}$$

must commute, or

$$(3.8) \quad \begin{pmatrix} 1 + \varepsilon \nabla u & \varepsilon (\partial u / \partial t) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon \bar{g} + \varepsilon^2 R \\ 1 \end{pmatrix} = \begin{pmatrix} \varepsilon g(\tau, x + \varepsilon u) \\ 1 \end{pmatrix}.$$

This requires

$$(3.9) \quad \frac{\partial u}{\partial t} = g(t, x) - \bar{g}(x).$$

If we let

$$(3.10) \quad \bar{g}(x) = \int_0^1 g(t, x) dt,$$

then we can solve for  $u$ . This is where the ‘‘averaging’’ comes in. Since  $u$  is defined on a compact, it is uniformly bounded.  $D^*$  is closed and contained in  $D$  such that  $\Phi_\varepsilon(S^1 \times D^*) \subset S^1 \times D$ .

This is where we need the initial condition  $x$  to be well inside  $D$ , otherwise we cannot be sure that there is a corresponding point  $\bar{x} \in D^*$ , such that  $\Phi_\varepsilon(0, x) = (0, \bar{x})$ . If, however, this the case, then solutions of (3.5) are mapped to solutions of (3.2).

We have the following commutative diagram

$$\begin{array}{ccc} T(S^1 \times D^*) & \xrightarrow{T\phi_\varepsilon} & T(S^1 \times D) \\ \uparrow \bar{\nu}_\varepsilon & \swarrow \bar{\psi} & \searrow \psi \\ & T\mathbb{R} & \\ & \uparrow i & \\ S^1 \times D^* & \xrightarrow{\phi_\varepsilon} & S^1 \times D \\ & \swarrow \bar{\psi} & \searrow \psi \\ & \mathbb{R} & \end{array}$$

where  $\psi$  and  $\bar{\psi}$  are solutions of (3.2) and (3.5) with  $\Phi_\varepsilon(0, \bar{\psi}(0)) = (0, \psi(0))$ . Since  $u$  is uniformly bounded, together with its derivatives,  $\phi_\varepsilon$  can be approximated by the canonical injection  $i : S^1 \times D^* \rightarrow S^1 \times D$ , and the diagram

$$(3.12) \quad \begin{array}{ccc} S^1 \times D^* & \xrightarrow{i} & S^1 \times D \\ & \swarrow \bar{\psi} & \searrow \psi \\ & [0, \infty) & \end{array}$$

is  $\varepsilon$ -commutative on  $[0, \infty)$ .

Since (3.5) is of type (1.5)( $\varepsilon$ ), be it on another time-scale, we are now in a position to apply Theorems 1 and 2. If  $\bar{g}$  gives a flow with contraction, then we define  $\bar{\psi}$  by putting  $\bar{\psi}(0) = \psi(0)$  and requiring that

$$(3.13) \quad \dot{\bar{\psi}} = \bar{g}(\bar{\psi}(t)).$$

Since  $\|\psi(0) - \bar{\psi}(0)\| = O(\varepsilon)$ , this gives the following  $\varepsilon$ -commutative diagram on  $[0, \infty)$ :

$$(3.14) \quad \begin{array}{ccccc} S^1 \times D^* & \xrightarrow{\text{id}} & S^1 \times D^* & \xrightarrow{i} & S^1 \times D \\ & \searrow \bar{\psi} & \uparrow \bar{\psi} & \nearrow \bar{\psi} & \\ & & [0, \infty) & & \end{array}$$

or

$$\|\bar{\psi}(t) - \psi(t)\| = O(\varepsilon) \quad \text{on } [0, \infty).$$

With this strong result, due to Sanchez-Palencia under somewhat different conditions, we conclude the first part of this paper. In the following sections, we formulate the theory of stable limit-cycles.

**4. Averaging.** Consider the equations

$$(4.1) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=1}^P \varepsilon^p X^p(\phi, x), & \phi \in S^1, \\ \dot{x} &= \sum_{p=1}^P \varepsilon^p Y^p(\phi, x), & x \in D_0 \subset \mathbb{R}^n, \end{aligned}$$

with  $X^p$  and  $Y^p$ ,  $p = 1, \dots, P$ , (sufficiently) smooth on  $\bar{D}_0$ , and  $D_0$  open and relatively compact with sufficiently smooth boundary. Then there exists for each  $N \in \mathbb{N} \cup \{0\}$  a domain  $D_N \subset D_0$  and a transformation

$$\Phi_\varepsilon^N : S^1 \times D_N \rightarrow S^1 \times D_0,$$

such that the distance between the boundaries of  $D_0$  and  $D_N$  is  $O_S(1)$  for  $\varepsilon \downarrow 0$  and  $N$  arbitrary and (4.1) and (4.2) given by

$$(4.2) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \varepsilon^p \bar{X}^{(p)}(x) + \varepsilon^{N+1} R_1^N(\phi, x; \varepsilon), & \phi \in S^1, \\ \dot{x} &= \sum_{p=K}^N \varepsilon^p \bar{Y}^{(p)}(x) + \varepsilon^{N+1} R_2^N(\phi, x; \varepsilon), & x \in D_N \subset D_0, \quad M, K \in \mathbb{N} \end{aligned}$$

are  $\Phi_\varepsilon^N$ -related. (Two vector fields  $\xi: \mathcal{M} \rightarrow T\mathcal{M}$  and  $\eta: \mathcal{N} \rightarrow T\mathcal{N}$  are  $f$ -related iff there is a map  $f: \mathcal{M} \rightarrow \mathcal{N}$  such that

$$\begin{array}{ccc} T\mathcal{M} & \xrightarrow{Tf} & T\mathcal{N} \\ \uparrow \xi & & \uparrow \eta \\ \mathcal{M} & \xrightarrow{f} & \mathcal{N} \end{array}$$

commutes).

The right-hand side of (4.2) is smooth (and therefore uniformly bounded in all variables) and  $\Phi_\epsilon^N$  is of the form:

$$\Phi_\epsilon^N : \begin{pmatrix} \phi \\ x \end{pmatrix} \rightarrow \begin{pmatrix} \phi \\ x \end{pmatrix} + \sum_{p=1}^N \epsilon^p u^{(p)}(\phi, x),$$

where  $u^{(p)} : \mathcal{S}^1 \times N \rightarrow \mathbb{R}^{n+1}$  is uniformly bounded for  $p = 1, \dots, N$  and has mean value zero, i.e.:

$$\int_{\mathcal{S}^1} u^{(p)}(\phi, x) d\phi = 0.$$

We will not prove this result here. The easiest way to do this seems to be induction on  $N$ . The idea of the proof is a straightforward application of averaging theory, but it takes a while to write all things out and to take care of all the little technicalities. A proof can be found in the author’s thesis (Sanders (1978a)).

Consider the equations

$$(4.3) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \epsilon^p \bar{X}^{(p)}(x) + \epsilon^{N+1} R_1^N(\phi, x; \epsilon), & \phi \in \mathcal{S}^1, \\ & & M, K \in \mathbb{N}, \\ \dot{x} &= \sum_{p=K}^N \epsilon^p \bar{Y}^{(p)}(x) + \epsilon^{N+1} R_2^N(\phi, x; \epsilon), & x \in D_N, \end{aligned}$$

and

$$(4.4) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \bar{X}^{(p)}(x), & \phi \in \mathcal{S}^1, \\ & & M, K \in \mathbb{N}, \\ \dot{x} &= \sum_{p=K}^N \epsilon^p \bar{Y}^{(p)}(x), & x \in D_N. \end{aligned}$$

Let  $(\tilde{\phi}, \tilde{x})$  be the solution of (4.3) with initial condition  $(\tilde{\phi}_0, \tilde{x}_0)$  and  $(\bar{\phi}, \bar{x})$  the solution of (4.4) with initial condition  $(\bar{\phi}_0, \bar{x}_0)$ . Let  $L$  be a constant, independent of  $\epsilon$ , such that we have existence of solution for both equations on  $0 \leq \epsilon^K t \leq L$ . (Such an  $L$  always exists). Then we have on this interval the following estimate:

$$\begin{aligned} \|\tilde{x}(t) - \bar{x}(t)\| &\leq C(\|\tilde{x}_0 - \bar{x}_0\| + \epsilon^{N-K+1}), \\ |\tilde{\phi}(t) - \bar{\phi}(t)| &\leq |\tilde{\phi}_0 - \bar{\phi}_0| + C(\epsilon^{N-K+1} + \epsilon^{M+N-2K+1} + \epsilon^{M-K} \|\tilde{x}_0 - \bar{x}_0\|). \end{aligned}$$

The proof of this statement is a straightforward application of Gronwall’s lemma. The reader should try to prove it if he does not understand why only  $K$  determines the time-scale. He should also notice that if we can extend the time scale of validity for the first estimate to  $[0, \infty)$ , say, then one can always extend the time-scale for the second estimate to  $0 \leq \epsilon^Q t \leq L$ ,  $Q \in \mathbb{N}$  arbitrary, provided one can estimate  $\|\tilde{x}_0 - \bar{x}_0\|$  up to arbitrary high powers of  $\epsilon$ .

Combining the two results given above, it is possible to prove the validity of approximation of the solutions of (4.1) to the solutions of (4.1) up to arbitrary high order on the interval  $0 \leq \epsilon^K t \leq L$ . Of course, one has to compute the right initial conditions  $\tilde{x}_0$  and  $\tilde{\phi}_0$  first, before solving (4.4), from inverting  $\Phi_\epsilon^N$  up to the desired order in  $\epsilon$ .

**5. Attraction and contraction.** We first state two lemmas without proofs, the first of which is standard and the second almost.

LEMMA 2 (Attraction). Consider the equation

$$(5.1) \quad \dot{x} = F(x), \quad x \in D \subset \mathbb{R}^n$$

$F$  is (sufficiently) smooth;  $F(0) = 0$  and all eigenvalues of  $dF(0)$  are different and have negative real parts. Denote  $dF(0)$  by  $A$  and let  $P$  be the matrix which diagonalizes  $A$ . Let  $C = \|P\| \|P^{-1}\| \geq 1$ . Write the equation as follows:

$$(5.2) \quad \dot{x} = Ax + G(x), \quad x \in D \subset \mathbb{R}^n$$

with

$$(5.3) \quad \|G(x)\| \leq C_1 \|x\|^2.$$

Let  $-\lambda, \lambda > 0$ , be the biggest eigenvalue of  $A$  and let  $\delta < \lambda/(CC_1)$ . (This is always possible, by definition of  $C_1$ ). Then one has the following estimate for any solution of (5.1) with initial value  $x_0$  such that  $\|x_0\| \leq \delta/(2C)$ :

$$(5.4) \quad \|x(t)\| < \delta e^{-\lambda t}.$$

Using this lemma one can prove the following.

LEMMA 3 (Contraction). Under the same assumptions as in Lemma 2, let  $C_2$  be such that

$$(5.5) \quad \|dG(x)\| \leq C_2 \|x\|, \quad \|x\| \leq \delta.$$

Let  $x_1$  and  $x_2$  be two solutions with initial conditions such that  $\|x_i^0\| < \delta/2C, i = 1, 2$ . Then

$$(5.6) \quad \|x_1(t) - x_2(t)\| \leq C e^{C_2/C_1} \|x_1^0 - x_2^0\| e^{-\lambda t} \quad \forall t \in [0, \infty).$$

We will now prove the validity of certain approximations under the assumption of contraction of the approximating flow in the  $x$ -component on an invariant domain. The idea of the present proof originates from Sanchez-Palencia (1975) and has already been used in Eckhaus (1975) and Verhulst (1975). It does not, however, assume the existence of the solution to be approximated. This made it necessary to slightly alter the proof-technique. Those readers familiar with the aforementioned papers might find it helpful to convince themselves that it is essentially the same proof.

With this theorem, one can prove the next theorem on the validity of approximations going to a limit cycle, while the theorems mentioned before only applied to attracting singular points.

THEOREM 3. Consider the equations

$$(5.7) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \varepsilon^p \bar{X}^{(p)}(x) + \varepsilon^{N+1} R_1^N(\phi, x; \varepsilon), & \phi \in S^1, \\ \dot{x} &= \sum_{p=K}^N \varepsilon^p \bar{Y}^{(p)}(x) + \varepsilon^{N+1} R_2^N(\phi, x; \varepsilon), & x \in D_N, \end{aligned}$$

and

$$(5.8) \quad \begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \varepsilon^p \bar{X}^{(p)}(x), & \phi \in S^1, \\ \dot{x} &= \sum_{p=K}^N \varepsilon^p \bar{Y}^{(p)}(x), & x \in D_N. \end{aligned}$$

Suppose any solution of (5.8), starting in  $S^1 \times D_\infty$ , stays there for all time (invariance of domain). Suppose one has, for any two solutions  $(\phi_1, x_1)$  and  $(\phi_2, x_2)$  of (5.8) with

initial values  $(\phi_1^0, x_1^0)$  and  $(\phi_2^0, x_2^0) \in D_N$ , respectively, the following estimate for all  $t_0 \in [0, \infty)$  and all  $t \in [t_0, \infty)$  (it suffices to have this for all  $t \in [t_0, t_0 + (L/\varepsilon)K]$ ),

$$\|x_1(t) - x_2(t)\| \leq C_3 e^{-C_4 \varepsilon^K (t-t_0)} \|x_1^0 - x_2^0\|.$$

Let  $(\tilde{\phi}, \tilde{x})$  and  $(\bar{\phi}, \bar{x})$  be solutions of (5.7) and (5.8) with initial conditions  $(\tilde{\phi}_0, \tilde{x}_0)$  and  $(\bar{\phi}_0, \bar{x}_0)$ , respectively.

$C_3 \geq 1$ , and we can always take  $C_3 > 1$ , such that  $C_3 - 1 = O_S(1)$  as  $\varepsilon \downarrow 0$ . Take the distance of the boundaries of  $D_0$  and  $D_\infty$ ,  $d$  such that  $d > (2\bar{M}/C_4) \log C_3$ , where  $\bar{M}$  is the sup-norm of the right-hand side of (5.7).

If  $\|\tilde{x}_0 - \bar{x}_0\| = O(\varepsilon)$ , and the distance of  $\tilde{x}_0$  to  $\partial D_N$  is bigger than  $d$ , one has the estimate

$$\|\tilde{x}(t) - \bar{x}(t)\| \leq C_5 (\|\hat{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) \left( \frac{C_3 + 1}{C_3 - 1} \right) \quad \forall t \in [0, \infty)$$

and

$$|\tilde{\phi}(t) - \bar{\phi}(t)| \leq |\tilde{\phi}_0 - \bar{\phi}_0| + C_6 t \{ \varepsilon^M (\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) + \varepsilon^{N+1} \} \quad \forall t \in [0, \infty).$$

*Remark.*  $\bar{x}$  does not leave  $D_\infty$ . As long as  $\tilde{x}$  is approximated by  $\bar{x}$ , it stays near  $D_\infty$  and therefore in  $D_N$ . Since  $\partial S^1$  is empty,  $(\tilde{\phi}, \tilde{x})$  exists as long as  $\tilde{x}$  is approximated by  $\bar{x}$ . It follows from the theorem that it exists for all time.

*Proof.* Let  $d$  be the distance of  $\tilde{x}_0$  to the boundary of  $D_N$ . By assumption,  $d > (2\bar{M}/C_4) \log C_3$ , and we may choose  $L$  such that

$$\frac{d}{\bar{M}} > L > \frac{2}{C_4} \log C_3.$$

The contraction estimate then implies

$$\left\| x_1 \left( t_0 + \frac{L}{\varepsilon K} \right) - x_2 \left( t_0 + \frac{L}{\varepsilon K} \right) \right\| \leq C_3 e^{-C_4 L} \|x_1^0 - x_2^0\| < \frac{1}{C_3} \|x_1^0 - x_2^0\|.$$

Let  $F$  be defined by

$$F(t) = \sup_{\tau \in [0, t]} \|\tilde{x}(\tau) - \bar{x}(\tau)\|.$$

We know from § 4 that  $F$  exists, is continuous and obeys the inequality

$$F(t) \leq C_5 (\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1})$$

on the interval  $0 \leq \varepsilon^K t \leq L$ .

Define  $F_\infty$  by

$$F_\infty = C_5 (\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) \frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}}.$$

It follows immediately that

$$F(t) < F_\infty \quad \text{on } 0 \leq \varepsilon^K t \leq L.$$

Suppose there exists a  $\tilde{t} \in (0, \infty)$  such that

- (i)  $F$  exists on  $[0, \tilde{t}]$ ,
- (ii)  $F(t) < F_\infty$  on  $[0, \tilde{t}]$ ,
- (iii)  $F(\tilde{t}) = F_\infty$ .

Since  $\|\tilde{x}_0 - \bar{x}_0\| = O(\varepsilon)$ ,  $\tilde{x}(t)$  stays close to  $\bar{x}(t)$  on  $[0, \tilde{t}]$ . But  $\bar{x}(t) \in D_\infty$  by assumption and, therefore,  $\tilde{x}(t) \in D_N$  for all  $t \in [0, \tilde{t}]$ . From the existence of  $\tilde{x}$  follows the existence

of  $F$ . If we show that the existence of  $\tilde{t}$  leads to contradiction, it will follow immediately that  $F(t) < F_\infty$  for all  $t \in [0, \infty)$ .

The distance of  $\tilde{x}(\tilde{t})$  to  $\partial D_N$  is  $d + O(\varepsilon)$ ; since the interval  $[d/\bar{M}, (2/C_4) \log C_3]$  has length of  $O_S(1)$ , we can take a slightly smaller  $L$  such that it fits the contraction requirements and we have existence of  $\tilde{x}$  on the interval  $[\tilde{t}, \tilde{t} + L/\varepsilon^K]$ .

Take  $\hat{t} < \tilde{t}$  such that  $\tilde{t} \in (\hat{t}, \hat{t} + L/\varepsilon^K)$  and define

$$I_+ = \left[ \hat{t}, \hat{t} + \frac{L}{\varepsilon^K} \right], \quad I_- = \left[ \hat{t} - \frac{L}{\varepsilon^K}, \hat{t} \right].$$

Since  $\tilde{x}(t) \in D_N$ , we may define  $(\hat{\phi}, \hat{x})$  as the solution of (5.8) with initial condition  $(\hat{\phi}(\hat{t}), \hat{x}(\hat{t})) = (\tilde{\phi}(\hat{t}), \tilde{x}(\hat{t}))$ .

Let  $\|\cdot\|_\pm$  be defined by

$$\|x\|_\pm = \sup_{t \in I_\pm} \|x(t)\|, \quad x \in C(I_\pm, \mathbb{R}^n).$$

It follows from § 4 that

$$\|\tilde{x} - \hat{x}\|_\pm \leq C_5 \varepsilon^{N-K+1}$$

since  $\tilde{x}(t)$  equals  $\hat{x}(t)$  for  $t = \hat{t}$ .

The  $\|\cdot\|_\pm$  are norms and we can use the triangle inequality as follows:

$$\|\tilde{x} - \bar{x}\|_\pm \leq \|\tilde{x} - \hat{x}\|_\pm + \|\hat{x} - \bar{x}\|_\pm.$$

$I_+$  is compact and everything is continuous. Therefore, there exists  $t_0 \in I_-$  such that

$$\|\bar{x} - \hat{x}\|_+ = \left\| \bar{x}\left(t_0 + \frac{L}{\varepsilon^K}\right) - \hat{x}\left(t_0 + \frac{L}{\varepsilon^K}\right) \right\|,$$

and, by the contraction hypothesis, this implies

$$\|\bar{x} - \hat{x}\|_\pm \leq C_3 e^{-C_4 L} \|\bar{x}(t_0) - \hat{x}(t_0)\| \leq C_3 e^{-C_4 L} \|\bar{x} - \hat{x}\|_-$$

since both  $\bar{x}(t_0)$  and  $\hat{x}(t_0)$  are in  $D_N$ , no matter what  $t_0$  is, as long as  $t_0 \in I_-$ .

Using these estimates and the fact that  $F(\hat{t}) < F_\infty$ , it follows that

$$\begin{aligned} \|\tilde{x} - \bar{x}\|_+ &\leq \|\tilde{x} - \hat{x}\|_+ + \|\hat{x} - \bar{x}\|_+ \\ &\leq \|\tilde{x} - \hat{x}\|_+ + C_3 e^{-C_4 L} \|\hat{x} - \bar{x}\|_- \\ &\leq \|\tilde{x} - \hat{x}\|_+ + C_3 e^{-C_4 L} \|\tilde{x} - \hat{x}\|_- + C_3 e^{-C_4 L} \|\tilde{x} - \bar{x}\|_- \\ &\leq C_5 \varepsilon^{N-K+1} (1 + C_3 e^{-C_4 L}) + C_3 e^{-C_4 L} F(\hat{t}) \\ &< C_5 \varepsilon^{N-K+1} (1 + C_3 e^{-C_4 L}) + C_3 e^{-C_4 L} F_\infty \\ &\leq C_5 \varepsilon^{N-K+1} (1 + C_3 e^{-C_4 L}) + C_5 (\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) \frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}} C_3 e^{-C_4 L} \\ &= C_5 \frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}} \{ \varepsilon^{N-K+1} (1 - C_3 e^{-C_4 L}) + C_3 (\|\tilde{x}_0 + \bar{x}_0\| + \varepsilon^{N-K+1}) \varepsilon^{-C_4 L} \} \\ &= C_5 \frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}} \{ \|\tilde{x}_0 - \bar{x}_0\| e^{-C_4 L} C_3 + \varepsilon^{N-K+1} \} \\ &< C_5 \frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}} \{ \|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1} \} = F_\infty. \end{aligned}$$

This implies

$$\begin{aligned} F(\tilde{t}) &= \sup_{t \in [0, \tilde{t}]} \|\tilde{x}(t) - \bar{x}(t)\| \leq \sup_{t \in [0, \tilde{t}] \cup I_+} \|\tilde{x}(t) - \bar{x}(t)\| \\ &= \max \left( \sup_{t \in [0, \tilde{t}]} \|\tilde{x}(t) - \bar{x}(t)\|, \|\tilde{x} - \bar{x}\|_+ \right) \\ &= \max (F(\hat{t}), \|\tilde{x} - \bar{x}\|_+) < F_\infty. \end{aligned}$$

This contradicts our assumption  $F(\tilde{t}) = F_\infty$ , and therefore,

$$F(t) < F_\infty \quad \forall t \in [0, \infty).$$

The desired estimate follows by estimating  $F_\infty$  as follows:

$$C_3 e^{-C_4 L} < \frac{1}{C_3},$$

and therefore,

$$\frac{1 + C_3 e^{-C_4 L}}{1 - C_3 e^{-C_4 L}} < \frac{1 + (1/C_3)}{1 - (1/C_3)} = \frac{C_3 + 1}{C_3 - 1}.$$

The proof for the  $\phi$ -component is straightforward and uses the remark at the end of § 4.

We are now able to state and prove

**THEOREM 4** (The “limit-cycle theorem”). We use the notation of Theorem 3. Suppose there is a  $x_\infty$  with  $\bar{Y}^{(K)}(x_\infty) = 0$  and  $d\bar{Y}^{(K)}(x_\infty)$  nondegenerate with all eigenvalues different and real parts (strictly) negative. Let  $D$  be the attraction domain of  $x_\infty$  and (5.8), and suppose there is a domain  $D'$  strictly contained in  $D$  (with  $O_S(1)$  distance of the boundaries). If  $x_0$  and  $\bar{x}_0$  are in  $D'$ , and

$$\|\tilde{x}_0 - \bar{x}_0\| = O(\varepsilon);$$

then

$$\begin{aligned} \|\tilde{x}(t) - \bar{x}(t)\| &\leq C_5(\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) \frac{C_3+1}{C_3-1} \quad \forall t \in [0, \infty), \\ |\tilde{\phi}(t) - \bar{\phi}(t)| &\leq |\tilde{\phi}_0 - \bar{\phi}_0| + C_6 t \{ \varepsilon^M (\|\tilde{x}_0 - \bar{x}_0\| + \varepsilon^{N-K+1}) + \varepsilon^{N+1} \}. \end{aligned}$$

*Proof.* Since  $D$  is the domain of attraction  $x_\infty$ ,  $\tilde{x}$  can reach every  $\varepsilon$ -independent neighborhood of  $x_\infty$  in  $0 \leq \varepsilon^K t \leq L$  we can use the result obtained in § 4 to estimate the difference between  $(\tilde{\phi}, \tilde{x})$  and  $(\bar{\phi}, \bar{x})$ . We then use the attraction and contraction lemma from there on. Using the contraction, we apply Theorem 3 to get the estimates given above.

**COROLLARY.** Consider the equations

$$\begin{aligned} \dot{\phi} &= 1 + \sum_{p=M}^N \varepsilon^p \bar{X}^{(p)}(x_\infty), \\ \dot{x} &= \sum_{p=K}^N \varepsilon^p \bar{Y}^{(p)}(x). \end{aligned} \tag{5.9}$$

We know from the attraction lemma that inside a certain neighborhood of  $x_\infty$ , we have

$$\|x^*(t) - x_\infty\| \leq \delta e^{-\lambda(t-t_0)}$$

if  $(\phi^*, x^*)$  is the solution of (5.9) with initial condition  $(\bar{\phi}, \bar{x}_0)$ . Therefore,

$$\begin{aligned} |\phi^*(t) - \bar{\phi}(t)| &\leq |\phi^*(t_0) - \bar{\phi}(t_0)| + \varepsilon^M \int_{t_0}^t \bar{K} \|\bar{x}(t) - x_\infty\| dt \\ &\leq C(\varepsilon^{N-K+1} + \varepsilon^{M+N-2K+1}) + \bar{K}C\varepsilon^{M-K} \\ &\leq C\varepsilon^{N-K+1} + C\varepsilon^{M-K}. \end{aligned}$$

This indicates, that if  $M - K > 0$ , the approximations are of a very simple type since we can explicitly solve  $\phi^*$  from (5.9):

$$\phi^*(t) = \bar{\phi}(t_0) + \left(1 + \sum_{p=M}^N \varepsilon^p \bar{X}^{(p)}(x_\infty)\right)(t - t_0).$$

In the case of the Van der Pol-oscillator, the approximation looks as follows:

Let  $y$  be the solution of

$$\ddot{y} + y = \varepsilon(1 - y^2)\dot{y}, \quad y(0) = y_0, \quad y'(0) = z_0, \quad y_0^2 + z_0^2 \neq 0,$$

and let

$$y^*(t, \varepsilon t, \varepsilon^2 t) = 2 \left(1 + \left(\frac{4}{y_0^2 + z_0^2} - 1\right) e^{-\varepsilon t}\right)^{-1/2} \sin \left(\operatorname{arctg} \left(\frac{y_0}{z_0}\right) + \left(1 - \frac{1}{16} \varepsilon^2\right) t\right);$$

then

$$y(t) = y^*(t, \varepsilon t, \varepsilon^2 t) + O(\varepsilon) \quad \text{on } 0 \leq \varepsilon^2 t \leq L.$$

See Sanders (1978b) for details.

**Acknowledgment.** I am gratefully indebted to both Dr. F. Verhulst and Prof. W. Eckhaus of the Mathematical Institute, State University of Utrecht, who taught me the ‘‘Sanchez-Palencia’’ trick, and encouraged me to write down the details of the limit-cycle case, respectively.

#### REFERENCES

- A. CODDINGTON AND N. LEVINSON, (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- W. ECKHAUS (1975a), *New approach to the asymptotic theory of nonlinear oscillations and wave-propagation*, J. Math. Anal. Appl., 49 (1975), pp. 575–611.
- (1975b), private communications.
- J. P. KEENER (1977), *On the validity of the two-timing method for large times*, this Journal, 8 (1977), pp. 1067–1091.
- S. C. PERSEK AND F. C. HOPPENSTEADT (1978), *Iterated averaging methods for systems of ordinary differential equations with a small parameter*, Comm. Pure Appl. Math., 31 (1978), pp. 133–156.
- E. SANCHEZ-PALENCIA (1975), *Méthode de Centrage et Comportement des Trajectoires dans l'Espace des Phases*, C.R. Acad. Sci. Paris Ser. A, 280, pp. 105–107.
- J. A. SANDERS (1978a), *On the theory of nonlinear resonance*, Thesis, Univ. of Utrecht, the Netherlands, 1978.
- (1978b), *Asymptotic approximations on arbitrary time-scales for systems with a limit-cycle*, Proceedings of ICNO 1978 Conference, Prague, pp. 601–602.
- F. VERHULST (1975), *On the theory of averaging*, Long Time Predictions in Dynamics, V. Szebehely and B. D. Tapley, eds., Reidel, Dordrecht, 1976, pp. 119–140.



## THE USE OF POLYNOMIALS IN EVALUATING CERTAIN PROBABILISTIC INTEGRALS\*

S. SIMONS†

**Abstract.** This paper is about the  $k$ -fold integral

$$I_k = \int d\mu(x_1) \int_{x_2 \leq x_1 + u} d\mu(x_2) \cdots \int_{x_k \leq x_{k-1} + u} d\mu(x_k),$$

where  $\mu$  is a probability measure on the real line and  $u > 0$ . This integral is important in statistical problems on ranking and selection. When  $k$  is large, it is not feasible to work out  $I_k$  directly by quadrature. In this paper, various approximations for  $I_k$  as polynomials in  $I_2$ , or in  $I_2$  and  $I_3$ , are investigated. There is a full analysis of the errors in these approximations, both in the general case and in the particular case where  $\mu$  is the normal distribution. For instance, in the case of the normal distribution and  $k \leq 9$  we can evaluate  $I_k$  with an error  $\leq 10^{-10}$  with no quadrature at all if  $u \geq 5$  and with an error of  $\leq 10^{-6}$  if  $u \geq 3$  with only one (one-variable) quadrature operation.

**Introduction.** Let  $\mu$  be any probability measure on the real line,  $R$ . It is of considerable importance in statistical problems on ranking and selection to be able to evaluate the  $k$ -fold integral

$$I_k = \int d\mu(x_1) \int_{x_2 \leq x_1 + u} d\mu(x_2) \cdots \int_{x_k \leq x_{k-1} + u} d\mu(x_k),$$

where  $u$  is a parameter  $> 0$ . These integrals arise in connection with problems that involve a complete ranking of all populations under study and/or a clustering of all populations into a random number of disjoint subsets. Though these problems have not yet been fully explored from a statistical point of view, we shall show in this paper that these integrals have a very rich analytic structure. More specifically, we shall show that  $I_k$  can be approximated by a polynomial in  $I_2$ , and the approximation becomes progressively better as  $u \rightarrow \infty$ .

This paper arose out of some joint work with Professor Milton Sobel in which  $I_k$  was computed for  $\mu$  the normal distribution by some very accurate adaptive quadrature subroutines for  $k = 2, 3, 4, 5$  and  $u = 0, .1, \dots, 5.0$ . The methods contained in this paper stem from an analysis of some inequalities proved by Professor Sobel. (See [2], in particular § 7.)

For our purposes, it is convenient to define  $I_k$  inductively. To this end, we define a sequence  $\{f_k\}_{k \geq 0}$  of real functions on  $R$  as follows:  $f_0 = 1$  and, for all  $k \geq 1$ ,

$$f_k(x) = \int_{(-\infty, x+u]} f_{k-1},$$

(where all integrals are with respect to  $\mu$  unless otherwise stated). Then, for all  $k \geq 1$ ,

$$I_k = \lim_{x \rightarrow -\infty} f_k(x) = \int f_{k-1}.$$

Clearly, for all  $k \geq 1$ ,  $f_k$  is a positive, increasing function on  $R$  and  $\lim_{x \rightarrow -\infty} f_k(x) = 0$ . Furthermore, for all  $x \in R$   $f_k(x) \leq I_k$ ; hence  $I_{k+1} \leq I_k$ . For convenience we define  $I_0 = 1$ .

\* Received by the editors August 23, 1978, and in final revised form October 18, 1979.

† Department of Mathematics, University of California, Santa Barbara, California, 93106.

So we have

$$1 = I_0 = I_1 \geq I_2 \geq I_3 \geq \dots$$

The general plan of this paper is as follows. In § 1 we define a sequence of numbers  $\{S_k\}_{k \geq 0}$ . The notation  $S_k$  was chosen for them because their usefulness depends on their being *small*.  $I_k$  can be written explicitly as a polynomial in  $S_0, \dots, S_k$  and in § 2 we give these polynomials explicitly for  $k = 2, 3, 4, 5$  as well as some other polynomial identities that can be deduced by various substitutions. Unfortunately, the unstructured nature of the coefficients of these polynomials makes them inconvenient for automatic computation and in § 3 we consider how to arrange the analysis more conveniently. Using generating functions we are able to prove in Theorem 6 a surprisingly good error estimate in terms of the numbers  $S_k$ . In § 4 we give an upper bound for  $S_k$  using the minimum-distance theorem of Hilbert space theory. In § 5 we consider the special case where  $\mu$  is the normal distribution. In § 6 we tabulate some results for the normal distribution that serve to show the accuracy of our methods. In § 7 we discuss some further inequalities that generalize some results from [2]. Finally, in § 8, we prove that

$$1 \geq I_2 \geq I_3^{1/2} \geq I_4^{1/3} \geq \dots,$$

and we investigate some of the properties of

$$\lim_{k \rightarrow \infty} I_k^{1/(k-1)}.$$

**1. The definition of  $S_k$ .** Our analysis involves certain quantities  $S_k$  (that depend both on  $\mu$  and on  $u$ ). We define  $S_0 = 1$  and, for all  $k \geq 1$ ,

$$(1) \quad S_k = \sum_{j=1}^k (-1)^{j-1} S_{k-j} I_j.$$

Since  $S_0 = 1$  and  $I_0 = 1$ , we can rewrite (1) as

$$(2) \quad I_k = \sum_{j=1}^k (-1)^{j-1} S_j I_{k-j}.$$

It is clearly possible to express  $I_k$  as a polynomial in  $S_0, S_1, \dots, S_k$  by a sequence of substitutions using (1) and (2). We return to this topic in § 2.

Our discussion of the properties of  $S_k$  is greatly facilitated by the definition of a sequence  $\{g_k\}_{k \geq 0}$  of real functions on  $R$  as follows: for all  $k \geq 0$ ,

$$(3) \quad g_k = \sum_{j=0}^k (-1)^j S_{k-j} f_j.$$

Our main results about  $g_k$  and  $S_k$  are contained in the following lemma.

LEMMA 1.

- (a) For all  $k \geq 1$ ,  $S_k = \int g_{k-1}$ .
- (b) For all  $k \geq 1$ ,  $g_k(x) = \int_{(x+u, \infty)} g_{k-1} \geq 0$ .
- (c) For all  $k \geq 1$ ,

$$S_k = \int d\mu(x_1) \int_{x_2 > x_1 + u} d\mu(x_2) \cdots \int_{x_k > x_{k-1} + u} d\mu(x_k).$$

- (d)  $1 = S_0 = S_1 \geq S_2 \geq \dots \geq 0$ .

*Proofs.* It follows from (3) with  $k$  replaced by  $k - 1$  that

$$(4) \quad g_{k-1} = \sum_{j=1}^k (-1)^{j-1} S_{k-j} f_{j-1}.$$

We obtain (a) by integrating (4) over  $R$  and using (1).

It follows from (a) that

$$\begin{aligned} \int_{(x+u, \infty)} g_{k-1} &= S_k - \int_{(-\infty, x+u]} g_{k-1}, \\ \text{using (4),} \quad &= S_k - \sum_{j=1}^k (-1)^{j-1} S_{k-j} f_j, \\ &= \sum_{j=0}^k (-1)^j S_{k-j} f_j, \end{aligned}$$

since  $f_0(x) = 1$ . Thus (b) follows from (3).

(c) is immediate from (a) and (b). (d) is immediate from (c).

**2. Some specimen polynomials.** In this section we give concrete examples of the polynomials mentioned at the beginning of § 1 and some of the identities that can be deduced from them.

The expressions for  $k = 2, \dots, 5$  obtained by successive substitutions using (1) and (2), taking account of the fact that  $S_0 = S_1 = 1$ , are:

$$\begin{aligned} I_2 &= 1 - S_2, \\ I_3 &= 1 - 2S_2 + S_3, \\ I_4 &= 1 - 3S_2 + S_2^2 + 2S_3 - S_4, \\ I_5 &= 1 - 4S_2 + 3S_2^2 - 2S_2S_3 + 3S_3 - 2S_4 + S_5. \end{aligned} \tag{5}$$

We can modify (5) slightly by the substitution  $S_2 = 1 - I_2$  and obtain expressions for  $I_k$  in terms of  $I_2, S_3, \dots, S_k$ :

$$\begin{aligned} I_3 &= [-1 + 2I_2] + S_3, \\ I_4 &= [-1 + I_2 + I_2^2] + 2S_3 - S_4, \\ I_5 &= [-2I_2 + 3I_2^2] + S_3 + 2I_2S_3 - 2S_4 + S_5. \end{aligned} \tag{6}$$

Now, whatever  $\mu$  is, the  $S_k$  are fairly small. For instance, we can deduce easily from Lemma 1(c) that, for fixed  $k$ ,  $\lim_{u \rightarrow \infty} S_k = 0$ , and we see in Theorem 8 that, for all  $k \geq 2$ ,  $S_k \leq 1/k!$ . Thus the expressions in (6) show the approximate polynomial dependence of  $I_k$  on  $I_2$  for  $k = 3, 4, 5$  and large  $u$ . The significance of this is that  $I_2$  is generally easier to evaluate by quadrature than  $I_3, I_4, \dots$ , since it involves fewer integrations.

We can modify (6) slightly by the substitution  $S_3 = 1 - 2I_2 + I_3$  and obtain expressions for  $I_k$  in terms of  $I_2, I_3, S_4, \dots, S_k$ . We see in § 6 how useful these results are for the normal distribution, where we can compute both  $I_2$  and  $I_3$  with great numerical accuracy. We obtain:

$$\begin{aligned} I_4 &= [1 - 3I_2 + I_2^2 + 2I_3] - S_4, \\ I_5 &= [1 - 2I_2 - I_2^2 + 2I_2I_3 + I_3] - 2S_4 + S_5. \end{aligned} \tag{7}$$

The somewhat unstructured nature of the coefficients in (6) and (7) makes these expressions inconvenient for automatic computation. In the next section we shall show how to arrange the analysis more conveniently. We introduce the concept of the *r-approximation* to  $I_k$ . The three expressions in square brackets in (6) are the 2-approximations to  $I_3, I_4$  and  $I_5$  and the two in (7) are the 3-approximations to  $I_4$  and  $I_5$ .

**3. The  $r$ -approximation to  $I_k$ .** Throughout this section,  $r$  is a fixed integer  $\geq 1$ . We have already observed in (2) that, for all  $k \geq 1$ ,

$$(8) \quad I_k = \sum_{j=1}^k (-1)^{j-1} S_j I_{k-j}.$$

We define a sequence  $\{A_k\}_{k \geq 0}$  of numbers as follows:  $A_0 = 1$  and, for all  $k \geq 1$ ,

$$(9) \quad A_k = \sum_{j=1}^{\min(r,k)} (-1)^{j-1} S_j A_{k-j}.$$

Clearly  $A_k = I_k$  for all  $k \leq r$ . The rationale behind this definition is that we are “neglecting”  $S_{r+1}, S_{r+2}, \dots$  in (8).  $A_k$  is the  $r$ -approximation to  $I_k$ . Our main goal in this section is to show that  $A_k$  is in fact a good approximation to  $I_k$ .

It is also convenient at this point to consider the *truncation error* in taking only  $r$  terms of (8). Specifically, if  $k > r$  we define

$$(10) \quad T_k = \sum_{j=r+1}^k (-1)^{j-1} S_j I_{k-j}.$$

It then follows from (8) that

$$(11) \quad I_k = \sum_{j=1}^r (-1)^{j-1} S_j I_{k-j} + T_k.$$

We extend the definition of  $T_k$  by writing

$$(12) \quad T_k = 0, \quad 0 \leq k \leq r.$$

The easiest way of handling  $I_k, A_k$  and  $T_k$  seems to be to use generating functions.

LEMMA 2.

(a) For all  $k \geq 0, I_k$  is the coefficient of  $z^k$  in the power series expansion of

$$\frac{1}{\sum_{j=0}^{\infty} (-1)^j S_j z^j}.$$

(b) For all  $k \geq 0, A_k$  is the coefficient of  $z^k$  in the power series expansion of

$$\frac{1}{\sum_{j=0}^r (-1)^j S_j z^j}.$$

(c) For all  $k \geq 0, T_k$  is the coefficient of  $z^k$  in the power series expansion of

$$\left( \sum_{j=r+1}^{\infty} (-1)^{j-1} S_j z^j \right) \left( \sum_{j=0}^{\infty} I_j z^j \right).$$

*Proofs.*

(a) It follows from (1) that, for all  $k \geq 1$ ,

$$\sum_{j=0}^k (-1)^j S_j I_{k-j} = 0.$$

Further,  $S_0 = I_0 = 1$ . Hence,

$$\left( \sum_{k=0}^{\infty} (-1)^k S_k z^k \right) \left( \sum_{k=0}^{\infty} I_k z^k \right) = 1.$$

This gives the desired result.

(b) We write  $S'_j = S_j (j \leq r)$  and  $S'_j = 0 (j > r)$ . It follows from (9) that, for all  $k \geq 1$ ,

$$\sum_{j=0}^k (-1)^j S'_j A_{k-j} = 0.$$

Further,  $S'_0 = A_0 = 1$ . Hence

$$\left( \sum_{k=0}^{\infty} (-1)^k S'_k z^k \right) \left( \sum_{k=0}^{\infty} A_k z^k \right) = 1.$$

This gives the desired result.

(c) We write  $\hat{S}_j = 0 (j \leq r)$  and  $\hat{S}_j = S_j (j > r)$ . It follows from (10) and (12) that, for all  $k \geq 0$

$$T_k = \sum_{j=0}^k (-1)^{j-1} \hat{S}_j I_{k-j}.$$

This gives the desired result.

In the next lemma we find an expression for the difference between  $I_k$  and its  $r$ -approximation.

LEMMA 3. For all  $k > r$ ,

$$I_k - A_k = \sum_{j=r+1}^k T_j A_{k-j}.$$

*Proof.* We observe from Lemma 2 that

$$\frac{1}{\sum_{k=0}^{\infty} I_k z^k} = \sum_{k=0}^{\infty} (-1)^k S_k z^k,$$

$$\frac{1}{\sum_{k=0}^{\infty} A_k z^k} = \sum_{k=0}^r (-1)^k S_k z^k,$$

and

$$\frac{\sum_{k=0}^{\infty} T_k z^k}{\sum_{k=0}^{\infty} I_k z^k} = - \sum_{k=r+1}^{\infty} (-1)^k S_k z^k.$$

Thus

$$\frac{1}{\sum_{k=0}^{\infty} A_k z^k} = \frac{1}{\sum_{k=0}^{\infty} I_k z^k} + \frac{\sum_{k=0}^{\infty} T_k z^k}{\sum_{k=0}^{\infty} I_k z^k}.$$

This can be rearranged to yield

$$\sum_{k=0}^{\infty} (I_k - A_k) z^k = \left( \sum_{k=0}^{\infty} A_k z^k \right) \left( \sum_{k=0}^{\infty} T_k z^k \right),$$

which gives the required result.

In Lemma 4 we describe a curious property of the partial sums in (8). We use this property in Lemma 5 to give bounds on  $T_k$ , which we then combine with the results of Lemma 3 to give our main approximation result in Theorem 6.

LEMMA 4. Let  $k \geq r$ . Then  $I_k$  is "bracketed" by the partial sums in (8), i.e., if  $r$  is even, then

$$I_k \geq \sum_{j=1}^r (-1)^{j-1} S_j I_{k-j};$$

while if  $r$  is odd, then

$$I_k \cong \sum_{j=1}^r (-1)^{j-1} S_j I_{k-j}.$$

*Proof.* Since the results are true (with equality) if  $k = r$ , we shall suppose that  $k > r$ . We note from (3), with  $k$  replaced by  $r$ , and  $j$  by  $i$ , that

$$\sum_{i=0}^r (-1)^i S_{r-i} f_i = g_r \cong 0.$$

Writing  $i = r - j$  we obtain

$$\sum_{j=0}^r (-1)^{r-j} S_j f_{r-j} \cong 0.$$

We now integrate this  $k - 1 - r (\cong 0)$  times over  $(-\infty, x + u]$  and then once over  $R$  and obtain

$$(13) \quad \sum_{j=0}^r (-1)^{r-j} S_j I_{k-j} \cong 0,$$

from which

$$\sum_{j=0}^r (-1)^j S_j I_{k-j}$$

has the same sign as  $(-1)^r$ . This gives the required result.

LEMMA 5. *Let  $k > r$ . Then*

$$0 \cong (-1)^r T_k \cong I_{k-r-1} S_{r+1} \cong S_{r+1}.$$

*Proof.* We note from (11) that

$$T_k = \sum_{j=0}^r (-1)^j S_j I_{k-j}$$

and so it is immediate from (13) that  $(-1)^r T_k \cong 0$ . Since  $k > r$ , we have  $k \cong r + 1$  so we can replace  $r$  by  $r + 1$  in (13) and get

$$\sum_{j=0}^{r+1} (-1)^{r+1-j} S_j I_{k-j} \cong 0.$$

Using (11) this can be written

$$(-1)^{r+1} T_k + S_{r+1} I_{k-r-1} \cong 0$$

and so  $(-1)^r T_k \cong S_{r+1} I_{k-r-1}$ , as required.

One would expect from (9) that  $A_k$  would get progressively worse as an approximation to  $I_k$  as  $k$  increases. Furthermore, since  $I_k \in [0, 1]$ , there is little point in considering cases in which  $A_k \notin [0, 1]$ . Taking these two observations together leads us to the following ‘‘reasonableness’’ condition:

$$(14) \quad A_j \in [0, 1] \quad \text{for all } j = 0, 1, \dots, k - 1.$$

We now come to our main approximation result. In order to explain it, we consider the case of (say)  $r = 3$ . We write  $D_k = I_k - A_k$ . It follows from Lemma 4 and (9) that, for

all  $k \geq 4$ ,

$$|I_k - I_{k-1} + S_2 I_{k-2} - S_3 I_{k-3}| \leq S_4,$$

and

$$A_k - A_{k-1} + S_2 A_{k-2} - S_3 A_{k-3} = 0,$$

from which

$$|D_k| \leq |D_{k-1}| + S_2 |D_{k-2}| + S_3 |D_{k-3}| + S_4.$$

Since  $D_0 = D_1 = D_2 = D_3 = 0$ , we can find inequalities for  $D_k$  recursively. For instance, we obtain that

$$|D_9| \leq (6 + 10S_2 + 4S_2^2 + 6S_3 + 2S_2S_3)S_4.$$

As opposed to this, Theorem 6 gives the simpler and sharper inequality

$$0 \leq D_9 \leq 6S_4.$$

**THEOREM 6.** *Let  $k > r$  and (14) be satisfied. If  $r$  is even, then*

$$A_k \leq I_k \leq A_k + (K - r)S_{r+1},$$

while if  $r$  is odd, then

$$A_k - (k - r)S_{r+1} \leq I_k \leq A_k.$$

*Proof.* It follows from Lemma 3 that

$$(-1)^r(I_k - A_k) = \sum_{j=r+1}^k (-1)^r T_j A_{k-j};$$

hence, from Lemma 5 and (14),

$$0 \leq (-1)^r(I_k - A_k) \leq \sum_{j=r+1}^k S_{r+1} = (k - r)S_{r+1}.$$

This gives the required result.

Of course, the usefulness of Theorem 6 as a device for computing  $I_k$  depends essentially on the size of  $S_{r+1}$ . We will consider this topic in the next section.

**4. An upper bound for  $S_k$ .** If  $x = (x_1, \dots, x_k) \in R$  we write  $\|x\| = (x_1^2 + \dots + x_k^2)^{1/2}$ .

**LEMMA 7.** *Let  $k \geq 2$  and*

$$W = \{x : x \in R^k, x_1 + u \leq x_2, x_2 + u \leq x_3, \dots, x_{k-1} + u \leq x_k\}.$$

Then

$$\inf_{x \in W} \|x\|^2 = \frac{k^3 - k}{12} u^2.$$

*Proof.* Since  $W$  is a closed convex set (and  $(R^k, \|\cdot\|)$  is a Hilbert space), there exists  $y \in W$  minimizing  $\|y\|^2$ . Now, for all  $\lambda \in R$ ,

$$y - \lambda(1, 1, \dots, 1) \in W;$$

hence,

$$\sum_{j=1}^k y_j^2 \leq \sum_{j=1}^k (y_j - \lambda)^2.$$

Since this holds for all  $\lambda$ , it follows that

$$(15) \quad \sum_{j=1}^k y_j = 0.$$

Thus there exists  $q, 1 \leq q < k$ , such that

$$y_q \leq 0 < y_{q+1}.$$

Clearly  $y_{q+1} = y_q + u$ , for otherwise we could decrease  $y_{q+1}$  slightly and find an element  $z$  of  $W$  such that  $\|z\|^2 < \|y\|^2$ . Using a similar argument we can prove that, for all  $j = 2, \dots, k$ ,

$$y_{j-1} + u = y_j.$$

It follows from (15) that, if  $k$  is odd,

$$y = \frac{u}{2}(-k-1, \dots, -4, -2, 0, 2, 4, \dots, k-1),$$

while, if  $k$  is even,

$$y = \frac{u}{2}(-k-1, \dots, -3, -1, 1, 3, \dots, k-1).$$

In either case, it follows by direct computation that

$$\|y\|^2 = \frac{k^3 - k}{12} u^2.$$

This completes the proof of the lemma.

In the result that follows we write  $\mu_k$  for the product measure  $\mu \times \dots \times \mu$  on  $R^k$ .

**THEOREM 8.** *Let  $k \geq 2$ . Then*

$$S_k \leq \frac{1}{k!} \mu_k \left( \left\{ x : x \in R^k, \|x\|^2 \geq \frac{k^3 - k}{12} u^2 \right\} \right).$$

*Proof.* For each permutation  $\sigma$  of  $\{1, \dots, k\}$  let

$$W_\sigma = \{(x_{\sigma_1}, \dots, x_{\sigma_k}) : x \in W\},$$

where  $W$  is the set introduced in Lemma 7. Clearly

$$(16) \quad \mu_k(W_\sigma) = \mu_k(W).$$

Since  $u > 0$ , if  $x \in W$  then

$$x_1 < x_2 < \dots < x_k;$$

consequently,

$$(17) \quad W_\sigma \cap W_\tau = \emptyset$$

if  $\sigma$  and  $\tau$  are different permutations of  $\{1, \dots, k\}$ . It follows from Lemma 1(c) that

$$(18) \quad S_k \leq \mu_k(W).$$

(The inequality “ $\leq$ ” arises because  $S_k$  is defined by integrals over *open* intervals, while  $W$  is defined by the corresponding *closed* intervals. Of course, if  $\mu$  is continuous then



we have equality.) Thus, from (16), (17) and (18)

$$k!S_k \leq \sum_{\sigma} \mu_k(W) = \sum_{\sigma} \mu_k(W_{\sigma}) = \mu_k(U_{\sigma}W_{\sigma}),$$

and the result now follows from Lemma 7.

**5. An upper bound for  $S_k$  for the normal distribution.** In this section we suppose that  $d\mu(x) = (1/\sqrt{2\pi}) e^{-x^2/2} dx$ .

LEMMA 9. For all  $k \geq 1$  and  $B \geq 0$ ,

$$\mu_k(\{x : x \in R^k, \|x\|^2 \geq B\}) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_{B/2}^{\infty} x^{k/2-1} e^{-x} dx.$$

*Proof.*

$$\begin{aligned} \mu_k(\{x : x \in R^k, \|x\|^2 \geq B\}) &= \int_{\|x\|^2 \geq B} \prod_{j=1}^k \left( \frac{1}{\sqrt{2\pi}} e^{-x_j^2/2} dx_k \right) \\ &= \frac{1}{(\sqrt{2\pi})^k} \int_{\|x\|^2 \geq B} e^{-\|x\|^2/2} dx_1 \cdots dx_k \\ &= \frac{A}{(\sqrt{2\pi})^k} \int_{r^2 \geq B} r^{k-1} e^{-r^2/2} dr, \end{aligned}$$

where  $A$  is the  $(k - 1)$ -dimensional Lebesgue measure of the surface of the unit sphere in  $R^k$ . If we now make the substitution  $x = r^2/2$  we obtain

$$(19) \quad \mu_k(\{x : x \in R^k, \|x\|^2 \geq B\}) = \frac{A}{2(\sqrt{\pi})^k} \int_{B/2}^{\infty} x^{k/2-1} e^{-x} dx.$$

By putting  $B = 0$  in (19) we obtain

$$(20) \quad 1 = \mu_k(R^k) = \frac{A}{2(\sqrt{\pi})^k} \int_0^{\infty} x^{k/2-1} e^{-x} dx.$$

The required result follows by dividing (19) by (20).

THEOREM 10. Let  $k \geq 2$ . Then

$$S_k \leq \frac{1}{k!} \left( \sum_{j=1}^m \frac{C^{h-j}}{\Gamma(h-j+1)} \right) e^{-C},$$

where

$$m = \left\lfloor \frac{k+1}{2} \right\rfloor, \quad C = \frac{k^3 - k}{24} u^2 \quad \text{and} \quad h = \frac{k}{2}.$$

*Proof.* By virtue of Theorem 8 and Lemma 9 it suffices to prove that

$$(21) \quad \int_C^{\infty} x^{h-1} e^{-x} dx \leq \Gamma(h) \left( \sum_{j=1}^m \frac{C^{h-j}}{\Gamma(h-j+1)} \right) e^{-C}.$$

For convenience we will separate the proof into two cases.

Case 1. ( $k$  is even). We observe that

$$\frac{d}{dx} \left[ \left( 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^{h-1}}{(h-1)!} \right) e^{-x} \right] = -\frac{x^{h-1}}{(h-1)!} e^{-x};$$

hence

$$\frac{1}{(h-1)!} \int_C^\infty x^{h-1} e^{-x} dx = \left(1 + C + \frac{C^2}{2!} + \dots + \frac{C^{h-1}}{(h-1)!}\right) e^{-C}.$$

Thus in this case (21) is true with equality.

Case 2. ( $k$  is odd). We observe that

$$\frac{d}{dx} \left[ \left( \frac{x^{1/2}}{\Gamma(\frac{3}{2})} + \frac{x^{3/2}}{\Gamma(\frac{5}{2})} + \dots + \frac{x^{h-1}}{\Gamma(h)} \right) e^{-x} \right] = - \left( \frac{x^{h-1}}{\Gamma(h)} - \frac{x^{-1/2}}{\Gamma(\frac{1}{2})} \right) e^{-x};$$

hence

$$\frac{1}{\Gamma(h)} \int_C^\infty x^{h-1} e^{-x} dx = \frac{1}{\Gamma(\frac{1}{2})} \int_C^\infty x^{-1/2} e^{-x} dx + \left( \frac{C^{1/2}}{\Gamma(\frac{3}{2})} + \dots + \frac{C^{h-1}}{\Gamma(h)} \right) e^{-C}.$$

Since  $x^{-1/2}$  is decreasing for  $x > 0$ ,

$$\int_C^\infty x^{-1/2} e^{-x} dx \leq C^{-1/2} \int_C^\infty e^{-x} dx = C^{-1/2} e^{-C};$$

thus,

$$\frac{1}{\Gamma(h)} \int_C^\infty x^{h-1} e^{-x} dx \leq \frac{1}{\Gamma(\frac{1}{2})} C^{-1/2} e^{-C} + \left( \frac{C^{1/2}}{\Gamma(\frac{3}{2})} + \dots + \frac{C^{h-1}}{\Gamma(h)} \right) e^{-C},$$

from which (21) is an immediate consequence.

**6. Some numerical results for the normal distribution.** We first give tables of the 2- and 3- approximations to  $I_k$ , computed from (9). The algorithm for the 2-approximations  $\{A_{2,k}\}_{k \geq 0}$  is:

$$A_{2,0} = A_{2,1} = 1,$$

$$A_{2,2} = I_2,$$

and for all  $k \geq 3$ ,

$$A_{2,k} = (A_{2,k-1} + I_2 A_{2,k-2}) - A_{2,k-2}.$$

The algorithm for the 3-approximations  $\{A_{3,k}\}_{k \geq 0}$  is:

$$A_{3,0} = A_{3,1} = 1,$$

$$A_{3,2} = I_2,$$

$$A_{3,3} = I_3,$$

and for all  $k \geq 4$ ,

$$A_{3,k} = (A_{3,k-1} + I_2 A_{3,k-2} + (1 + I_3) A_{3,k-3}) - (A_{3,k-2} + 2I_2 A_{3,k-3}).$$

The arithmetic has been arranged somewhat differently from that in (9) to avoid the risk of subtractive cancellation inherent in the numerical evaluation of expressions like  $1 - 2I_2 + I_3$ .

$I_2$  is worked out explicitly from the formula

$$I_2 = \Phi\left(\frac{u}{\sqrt{2}}\right),$$

and  $I_3$  is worked out by quadrature using the adaptive subroutine QUANC8 and the formula

$$I_3 = \int_{-\infty}^{\infty} \Phi(u+x)\Phi(u-x) d\Phi(x),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

The subroutine QUANC8 is described in Chapter 5 of [1], and the evaluation of  $I_3$  in § 9 of [2]. Although the results displayed are rounded to six digits, the machine is manipulating values of  $I_2$  and  $I_3$  that are probably correct to at least 10. (See Table 1.)

TABLE 1  
2- and 3-approximations to  $I_k$

$u$	$A_{2,3}$	$A_{3,3}$	$A_{2,4}$	$A_{3,4}$	$A_{2,5}$	$A_{3,5}$	$A_{2,6}$	$A_{3,6}$
0	.000000	.166667	-.250000	.083333	-.250000	.083333	-.125000	.069444
1	.520500	.536152	.338230	.369533	.213440	.252890	.132349	.172686
2	.842701	.843012	.770237	.770860	.703959	.704844	.643380	.644478
3	.966105	.966106	.949445	.949447	.933072	.933075	.916981	.916985
4	.995322	.995322	.992989	.992989	.990661	.990661	.988338	.988338
5	.999593	.999593	.999390	.999390	.999186	.999186	.998983	.998983
6	.999978	.999978	.999967	.999967	.999956	.999956	.999945	.999945
7	.999999	.999999	.999999	.999999	.999999	.999999	.999998	.999998
$u$	$A_{2,7}$	$A_{3,7}$	$A_{2,8}$	$A_{3,8}$	$A_{2,9}$	$A_{3,9}$	$A_{2,10}$	$A_{3,10}$
0	.000000	.041667	.062500	.020833	.062500	.011574	.031250	.008102
1	.081177	.117840	.049446	.080396	.029984	.054847	.018129	.037416
2	.588014	.589283	.537412	.538814	.491165	.492668	.448898	.450474
3	.901168	.901173	.885628	.885634	.870355	.870362	.855346	.855354
4	.986021	.986021	.983710	.983710	.981404	.981404	.979103	.979103
5	.998780	.998780	.998576	.998576	.998373	.998373	.998170	.998170
6	.999934	.999934	.999923	.999923	.999912	.999912	.999901	.999901
7	.999998	.999998	.999997	.999997	.999997	.999997	.999997	.999997

We have, in fact, worked out much more extensive tables with  $u$  running from 0 to 10 in steps of .1. It is interesting to observe that to construct such a table it required about the same amount of computing effort as was required in [2] to work out only one or two values of  $I_4$  by quadrature directly. So the methods given in this paper are preferable if they give a sufficiently accurate result.

This brings us to an analysis of the errors in our approximations. We next give a table from which upper bounds for  $S_k$  can be read off. (See Table 2.) The upper bound is computed directly from the formula in Theorem 10 and we have tabulated  $-\log_{10}$  (upper bound). Thus, for instance, it follows from the table that if  $u = 3$  then  $S_4 \leq 10^{-9.7}$ .

It will be observed that  $S_k$  decreases exceedingly rapidly as  $k$  and  $u$  increase. These results clearly justify the claim made in the introduction that  $S_k$  is small, at least for  $k \geq 4$  and  $u \geq 2$ .

We note from Theorem 6 that for all the values tabulated other than  $u = 0$ ,

$$A_{2,k} \leq I_k \leq A_{2,k} + (k-2)S_3, \quad k \geq 3,$$

TABLE 2  
Upper bounds for  $S_k$

$u$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
1	0.4	0.9	1.9	3.2	4.9	7.3	10.4	14.4	19.3
2	0.7	2.1	4.6	8.9	15.2	24.1	36.0	51.4	70.7
3	1.2	4.1	9.7	19.2	33.5	53.6	80.6	115.3	158.8
4	2.0	7.0	17.1	34.0	59.6	95.6	143.7	205.6	283.2
5	3.0	10.8	26.7	53.3	93.4	149.8	225.2	322.2	443.7
6	4.2	15.5	38.5	76.9	134.9	216.3	325.0	465.0	640.1
7	5.6	21.1	52.4	105.0	184.0	295.0	443.2	633.9	872.5
8	7.2	27.6	68.6	137.4	240.8	385.9	579.6	828.9	1140.7
9	9.0	34.9	87.0	174.1	305.2	489.0	734.4	1050.1	1444.9
10	11.1	43.1	107.5	215.3	377.2	604.3	907.4	1297.3	1784.9

and

$$A_{3,k} - (k - 3)S_4 \leq I_k \leq A_{3,k}, \quad k \geq 4.$$

We use these estimates to give a table from which upper bounds for  $|A_{2,k} - I_k|$  and  $|A_{3,k} - I_k|$  can be read off. (See Table 3.) Our notation is that

$$E_{r,k} = -\log_{10} (\text{upper bound for } |A_{r,k} - I_k|).$$

We summarize our conclusions in Table 4. The entry “2” means that the 2-approximation can be used and the entry “3” means that the 3-approximation can be used. We note that the 3-approximation involves only one quadrature operation and so it will be

TABLE 3  
Upper bounds for the error in the 2- and 3-approximations to  $I_k$

$u$	$E_{2,3}$	$E_{3,3}$	$E_{2,4}$	$E_{3,4}$	$E_{2,5}$	$E_{3,5}$	$E_{2,6}$	$E_{3,6}$
1	0.9		0.6	1.9	0.5	1.6	0.3	1.4
2	2.1		1.8	4.6	1.6	4.3	1.5	4.2
3	4.1		3.8	9.7	3.6	9.4	3.5	9.3
4	7.0		6.7	17.1	6.5	16.8	6.4	16.6
5	10.8		10.5	26.7	10.3	26.4	10.2	26.2
6	15.5		15.2	38.5	15.0	38.2	14.9	38.0
7	21.1		20.8	52.4	20.6	52.1	20.5	52.0
8	27.6		27.3	68.6	27.1	68.3	27.0	68.1
9	34.9		34.6	87.0	34.4	86.7	34.3	86.5
10	43.1		42.8	107.5	42.6	107.2	42.5	107.0
$u$	$E_{2,7}$	$E_{3,7}$	$E_{2,8}$	$E_{3,8}$	$E_{2,9}$	$E_{3,9}$	$E_{2,10}$	$E_{3,10}$
1	0.2	1.3	0.2	1.2	0.1	1.1	0.0	1.0
2	1.4	4.0	1.3	3.9	1.2	3.9	1.2	3.8
3	3.4	9.1	3.3	9.0	3.2	9.0	3.2	8.9
4	6.3	16.5	6.2	16.4	6.2	16.3	6.1	16.2
5	10.1	26.1	10.0	26.0	10.0	25.9	9.9	25.8
6	14.8	37.9	14.7	37.8	14.7	37.7	14.6	37.6
7	20.4	51.8	20.3	51.7	20.3	51.7	20.2	51.6
8	26.9	68.0	26.8	67.9	26.7	67.8	26.7	67.8
9	34.2	86.4	34.1	86.3	34.1	86.2	34.0	86.1
10	42.4	106.9	42.3	106.8	42.3	106.7	42.2	106.7

TABLE 4

$u$	Maximum error tolerated					
	$10^{-1}$	$10^{-3}$	$10^{-6}$	$10^{-10}$	$10^{-20}$	$10^{-40}$
$\equiv 1$	3					
$\equiv 2$	2	3				
$\equiv 3$	2	2	3			
$\equiv 4$	2	2	2	3		
$\equiv 5$	2	2	2	2	3	
$\equiv 7$	2	2	2	2	2	3
$\equiv 10$	2	2	2	2	2	2

fairly fast. The 2-approximation does not involve any quadrature at all and so it will be very fast indeed. (These results are valid for  $k \leq 9$ .)

**7. Some further inequalities.** In this section we prove some further inequalities for the situation considered in the first four sections—where  $\mu$  is a general measure. At this point we do not know if these inequalities can be used efficiently for computation. At any rate, the analysis seems more complicated than that in § 3.

LEMMA 11. *Let  $m \geq 1$ ,  $n \geq 2$  and  $r \geq 0$ . If  $r$  is even then*

$$\sum_{j=0}^r (-1)^j S_j (I_{m+n-1+r-j} - I_m I_{n+r-j}) \leq 0,$$

while if  $r$  is odd, then

$$\sum_{j=0}^r (-1)^j S_j (I_{m+n-1+r-j} - I_m I_{n+r-j}) \geq 0.$$

*Proof.* It follows from (3) with  $k$  replaced by  $r$  that

$$g_r = (-1)^r \sum_{j=0}^r (-1)^j S_j f_{r-j} \geq 0;$$

hence, by integrating  $n - 1$  times, the function

$$f = (-1)^r \sum_{j=0}^r (-1)^j S_j f_{r-j+n-1}$$

is increasing. Consequently, for all  $x \in R$ ,

$$\int_{(-\infty, x+u]} f \leq \mu((-\infty, x+u]) \int f = f_1(x) \int f;$$

i.e.,

$$(-1)^r \sum_{j=0}^r (-1)^j S_j f_{r-j+n} \leq (-1)^r f_1 \sum_{j=0}^r (-1)^j S_j I_{r-j+n},$$

which can be rewritten

$$(-1)^r \sum_{j=0}^r (-1)^j S_j (f_{r-j+n} - I_{r-j+n} f_1) \leq 0.$$

The required result now follows by integrating  $m - 1$  times.

Lemma 11 gives a large number of results. Here we will analyze only a few of them.

We first observe that it is immediate from the definition of  $I_k$  that for all  $m, n \geq 1$ ,  $I_{m+n} \leq I_m I_n$ . (cf [2, 7.12]). It was also observed in [2, 7.5] that, for the normal distribution, for all  $k \geq 3$ ,  $I_k \leq I_2 I_{k-1}$ . Both of these results are improved substantially by Corollary 12(a).

We next consider 7.6, 7.10 and 7.11 in [2]. The first of these has already been generalized in Lemma 4 (with  $r = 2$ ). The other two can be obtained from Corollary 12(b) by putting  $n = 2$  and  $n = 3$ , respectively.

**COROLLARY 12.**

(a) If  $m, n \geq 1$  then  $I_{m+n-1} \leq I_m I_n$ .

(b) If  $2 \leq n < n + 1 \leq k$  then  $I_k \geq I_{k-1} - (I_n - I_{n+1}) I_{k-n}$ .

*Proofs.* (a) The result is immediate if  $n \geq 1$ ; if  $n \geq 2$  it follows from Lemma 11 with  $r = 0$ .

(b) It follows from Lemma 11 with  $r = 1$  that for all  $m \geq 1$  and  $n \geq 2$ ,  $(I_{m+n} - I_m I_{n+1}) - (I_{m+n-1} - I_m I_n) \geq 0$ . The result follows by putting  $m = k - n$ .

**8. On  $I_k^{1/(k-1)}$ .**

**LEMMA 13.** Let  $k \geq 1$ . Then for all  $a < b$ ,

$$(22) \quad f_k(a) f_{k-1}(b) \leq f_{k-1}(a) f_k(b).$$

*Proof.* If  $k = 1$ , then (22) reduces, for all  $a < b$ , to  $f_1(a) \leq f_1(b)$ , which is clearly true from the definition of  $f_1$ . We prove the result by induction. We suppose that it is true for  $k = j \geq 1$ , i.e., for all  $x < y$ ,

$$f_j(x) f_{j-1}(y) \leq f_{j-1}(x) f_j(y).$$

If  $a < b$  then

$$(23) \quad \int_{(-\infty, a+u]} \mu(dx) \int_{(a+u, b+u]} f_j(x) f_{j-1}(y) \mu(dy) \\ \leq \int_{(-\infty, a+u]} \mu(dx) \int_{(a+u, b+u]} f_{j-1}(x) f_j(y) \mu(dy)$$

since  $x > y$  for all  $(x, y)$  in the domain of integration. Now (23) can be rewritten

$$f_{j+1}(a) [f_j(b) - f_j(a)] \leq f_j(a) [f_{j+1}(b) - f_{j+1}(a)],$$

from which

$$f_{j+1}(a) f_j(b) \leq f_j(a) f_{j+1}(b).$$

This completes the proof of the inductive step.

**COROLLARY 14.** For all  $k \geq 1$ ,  $I_{k-1} I_{k+1} \leq I_k^2$ .

*Proof.* Letting  $b \rightarrow \infty$  in (22) we obtain

$$I_{k-1} f_k \leq I_k f_{k-1}.$$

The required result follows by integrating.

**THEOREM 15.**

(a) If  $I_k > 0$  for all  $k$  then

$$1 \geq I_2 \geq \frac{I_3}{I_2} \geq \frac{I_4}{I_3} \geq \dots \geq 0.$$

(b) If  $I_k > 0$  for all  $k$  then

$$1 \geq \frac{I_3}{I_2} \geq \frac{I_4^2}{I_3} \geq \dots \geq 0.$$

(c)  $1 \geq I_2 \geq I_3^{1/2} \geq I_4^{1/3} \geq \dots \geq 0.$

*Proofs.* It follows from Corollary 14 that, for all  $k \geq 1,$

$$I_{k-1}I_{k+1} \leq I_k^2 \quad \text{and} \quad I_{k-1}^{k-1}I_{k+1}^{k-1} \leq I_k^{2k-2};$$

hence,

$$\frac{I_k}{I_{k-1}} \geq \frac{I_{k+1}}{I_k} \quad \text{and} \quad \frac{I_k^{k-2}}{I_{k-1}^{k-1}} \geq \frac{I_{k+1}^{k-1}}{I_k^k}$$

and (a) and (b) are immediate consequences.

We separate the proof of (c) into two cases.

*Case 1.* [ $I_k > 0$  for all  $k.$ ] Here it is immediate from (b) that, for all  $k \geq 1,$

$$1 \geq \frac{I_k^{k-2}}{I_{k-1}^{k-1}};$$

hence

$$I_{k-1}^{k-1} \geq I_k^{k-2},$$

from which

$$I_{k-1}^{1/(k-2)} \geq I_k^{1/(k-1)},$$

and (c) follows.

*Case 2.* [There exists  $k$  such that  $I_k = 0.$ ] Here we write  $K$  for the smallest value of  $k$  for which  $I_k = 0.$  Clearly  $K \geq 2.$  By suitably modifying the proofs already given we obtain

$$1 \geq I_2 \geq \frac{I_3}{I_2} \geq \dots \geq \frac{I_K}{I_{K-1}} = 0,$$

$$1 \geq \frac{I_3}{I_2} \geq \dots \geq \frac{I_K^{K-2}}{I_{K-1}^{K-1}} = 0$$

and

$$1 \geq I_2 \geq I_3^{1/2} \geq \dots \geq I_K^{1/(K-1)} = 0.$$

However,  $I_{K+1} = I_{K+2} = \dots = 0,$  and so (c) is true.

**THEOREM 16.**

(a) If  $I_k > 0$  for all  $k \geq 2,$  then

$$\lim_{k \rightarrow \infty} \frac{I_k}{I_{k-1}} \quad \text{and} \quad \lim_{k \rightarrow \infty} \frac{I_k^{k-2}}{I_{k-1}^{k-1}}$$

both exist and lie in  $[0, 1].$

(b)  $\lim_{k \rightarrow \infty} I_k^{1/(k-1)}$  exists and lies in  $[0, 1].$

(c) If  $I_k > 0$  for all  $k \geq 2$  then

$$\lim_{k \rightarrow \infty} I_k^{1/(k-1)} = \lim_{k \rightarrow \infty} \frac{I_k}{I_{k-1}}.$$

*Proofs.* (a) and (b) are immediate from Theorem 15. (c) follows from a classical result about the ratio test and the root test. (See, e.g., [3, 3.37, p. 59]).

**DEFINITION.** We write  $\rho = \lim_{k \rightarrow \infty} I_k^{1/(k-1)} = \lim_{k \rightarrow \infty} I_k/I_{k-1}$ .

*Comments.* It would, indeed, be interesting to have a *direct* method for computing  $\rho$  (say) when  $\mu$  is the normal distribution. We have numerical evidence that suggests that  $I_k^{1/(k-1)}$  converges to  $\rho$  very rapidly in this case and so it might well be that  $\rho^{k-1}$  would be a good approximation to  $I_k$ . Whether or not this would compete with the numerical methods already discussed is a matter of conjecture. Another interesting question is: what information does the value of  $\rho$  give about  $\mu$ ? Our final theorem gives a result in this direction.

**THEOREM 17.** *The conditions (a)–(e) are equivalent.*

- (a) *There exists  $w \in R$  such that  $\mu([w, w + u]) = 1$ .*
- (b)  $f_1 = 1[\mu]$ .
- (c) *For all  $k \geq 1$ ,  $I_k = 1$ .*
- (d)  $\rho = 1$ .
- (e)  $I_2 = 1$ .

*Proofs.* We first prove that (b)  $\Rightarrow$  (c)  $\Rightarrow$  (d)  $\Rightarrow$  (e)  $\Rightarrow$  (b). If (b) is true then, by integrating, for all  $k \geq 1$ ,

$$f_{k+1} = f_k,$$

from which, for all  $k \geq 1$ ,

$$I_{k+1} = I_k,$$

and (c) is an immediate consequence. The implications (c)  $\Rightarrow$  (d) and (d)  $\Rightarrow$  (e) are trivial. If (e) is true then  $\int f_1 = 1 = \int 1$  and (b) follows since  $f_1 \leq 1$ .

We now prove that (a)  $\Rightarrow$  (b)  $\Rightarrow$  (a). If (a) is true then, for all  $x \geq w$ ,

$$f_1(x) = \mu((-\infty, x + u]) = 1;$$

hence,

$$R\{f_1 \neq 1\} \subset (-\infty, w),$$

from which (b) is immediate. Conversely, if (b) is true then there exists  $z \in R$  such that  $f_1(z) = 1$ . It follows from a Dedekind section argument that there exists  $w \in R$  such that  $f_1(w) = 1$  and, for all  $x < w$ ,  $f_1(x) < 1$ . It is immediate from this that

$$(24) \quad \mu((-\infty, w + u]) = f_1(w) = 1$$

and

$$(-\infty, w) \subset R\{f_1 \neq 1\};$$

thus, from (b)

$$(25) \quad \mu((-\infty, w)) = 0.$$

(a) is now immediate from (24) and (25).

**Acknowledgment.** I am very grateful to Professor Sobel for his help and inspiration during the preparation of this paper.

REFERENCES

[1] GEORGE E. FORSYTHE, MICHAEL A. MALCOLM AND CLEVE B. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, NJ, 1977.  
 [2] M. SOBEL AND S. SIMONS, *Clustering problems in the framework of ranking and selection*, In preparation.  
 [3] W. RUDIN, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1964.



## NOTES ON THE DYSON CONJECTURE\*

GEORGE E. ANDREWS†

**Abstract.** A multiple integral formulation of Dyson's conjecture is presented; also Good's proof of the Dyson conjecture is modified to treat directly the multiple series representation of the constant term.

**1. Introduction.** In [1], I raised the following:

CONJECTURE. *In the expansion of*

$$(1.1) \quad \prod_{1 \leq i \neq j \leq n} \left\{ \left( 1 - \frac{x_i}{x_j} \varepsilon_{ij} \right) \left( 1 - \frac{x_i}{x_j} \varepsilon_{ij} q \right) \cdots \left( 1 - \frac{x_i}{x_j} \varepsilon_{ij} q^{a_i-1} \right) \right\}$$

(where  $\varepsilon_{ij} = 1$  if  $i < j$  and  $\varepsilon_{ij} = q$  if  $i > j$ ), the term independent of the  $x_i$ 's is

$$(1.2) \quad \frac{(q)_{a_1+a_2+\cdots+a_n}}{(q)_{a_1}(q)_{a_2} \cdots (q)_{a_n}},$$

where

$$(q)_A = (1-q)(1-q^2) \cdots (1-q^A)$$

F. J. Dyson [4] raised this question for the case  $q = 1$ . The original Dyson conjecture was settled independently by Gunson [6] and Wilson [11]. I. J. Good [5] also gave a beautiful proof which we shall discuss further in § 3.

I. G. Macdonald has pointed out that if all the  $a_i \rightarrow \infty$  in the above conjecture then the result is already established by considering the constant term of the corresponding "Macdonald identity" (our choice of words) for the Lie algebra  $A(n-1)$  [7]. Thus our conjecture appears to correspond to some finite version of the Macdonald identity for  $A(n-1)$ . Macdonald has also made a number of Dyson-like conjectures for other Lie algebras, so that such problems become even more intriguing.

Until recently, however, the only general theorem on the subject was the original Dyson conjecture. Within the past year, E. Bombieri and A. Selberg showed that the integral

$$(1.3) \quad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left( -k \sum_{i=1}^N x_i^2 \right) \prod_{i < j} (x_i - x_j)^{2k} dx_1 \cdots dx_N \\ = (2\pi)^{N/2} (2k)^{-N/2-kN(N-1)/2} (k!)^{-N} \prod_{j=1}^N (kj)!,$$

conjectured by M. L. Mehta [8, p. 42], is valid. They deduced this result from another integral formula due to Selberg [9]:

$$(1.4) \quad \int_0^1 \cdots \int_0^1 (x_1 x_2 \cdots x_N)^{\alpha-1} \left\{ \prod_{j=1}^N (1-x_j) \right\}^{\beta-1} \prod_{i < j} (x_i - x_j)^{2k} dx_1 \cdots dx_N \\ = \prod_{j=1}^N \frac{\Gamma(1+jk)\Gamma(\alpha+(j-1)k)\Gamma(\beta+(j-1)k)}{\Gamma(1+k)\Gamma(\alpha+\beta+(N+j-2)k)}.$$

Now the case  $N = 1$  of (1.4) is the famous Euler integral for the  $\beta$ -function. The case

\* Received by the editors October 19, 1979.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802. This work was supported in part by the National Science Foundation under Grant MCS-7722992.

$N = 2$  is easily reduced to

$$(1.5) \quad \frac{\Gamma(\beta)^2 \Gamma(\alpha) \Gamma(2k + \alpha)}{\Gamma(2k + \alpha + \beta) \Gamma(\alpha + \beta)} \sum_{j=0}^{2k} \prod_{h=0}^{j-1} \frac{(-2k + h)(-2k - \alpha - \beta + 1 + h)(\alpha + h)}{(h + 1)(\alpha + \beta + h)(-2k - \alpha + 1 + h)}$$

$$= \frac{\Gamma(\alpha) \Gamma(\beta) \Gamma(1 + 2k) \Gamma(\alpha + k) \Gamma(\beta + k)}{\Gamma(\alpha + \beta + k) \Gamma(1 + k) \Gamma(\alpha + \beta + 2k)},$$

which is Dixon’s theorem [3]. Since the Dyson conjecture when  $n = 3$  also reduces to Dixon’s theorem, we might hope that there is an analogue of (1.4) that is equivalent to the general Dyson conjecture.

In § 2, we prove,

$$(1.6) \quad \int_{P_1}^{(1+,0+,1-,0-)} \cdots \int_{P_{n-1}}^{(1+,0+,1-,0-)} \prod_{j=1}^{n-1} u_j^{-z-1} (1 - u_j)^{z+a_j}$$

$$\cdot \prod_{1 \leq i \neq j \leq n-1} \left(1 - \frac{u_i}{u_j}\right)^{a_i} du_1 \cdots du_{n-1}$$

$$= \frac{(2\pi i)^{2n-2} \Gamma(a_1 + \cdots + a_{n-1} + z + 1)}{\Gamma(z)^{n-1} \Gamma(1 - z)^{n-1} a_1! \cdots a_{n-1}! \Gamma(z + 1)},$$

by showing that (1.6) is essentially equivalent to the Dyson conjecture. The integrals appearing in (1.6) are the Pochhammer contour integrals: the path of integration in the  $i$ th integral starts from the point  $P_i$  on the real  $u_i$ -axis between 0 and 1, encircles the point 1 in the counterclockwise direction, encircles 0 in the counterclockwise direction, then 1 clockwise, 0 clockwise, and returns to  $P_i$  [10, p. 256].

In § 3 we give a second proof of the Dyson conjecture by reducing Good’s proof to the point where it applies only to the multiple series of binomial coefficients that can be easily derived as the constant term in (1.1) when  $q = 1$ .

Since the upshot of our work is to produce one proof and one equivalent formulation of a theorem already proved by Good [5] with astounding simplicity, some justification for our efforts should be given.

(1). Selberg’s integral (1.4) and (1.6) provide possible extensions of the  $\beta$ -integral to multiple integrals. Since the  $\beta$ -integral can be viewed as the cornerstone of the extensive and important theory of hypergeometric functions, it is natural to suppose that the multiple integral analogues of the  $\beta$ -integral will play an important role in the development of multiple hypergeometric series.

(2). The treatments and formulations of the Dyson conjecture presented here would seem to be more amenable to the development of  $g$ -analogues than any of the known proofs. Indeed, R. Askey has already formulated a conjecture for the  $q$ -analogue of (1.6). Due to the interest in the  $q$ -Dyson conjecture, it seems useful to present various possible methods for attacking it.

(3). The multiple integral (1.6) provides an added facet to the Dyson conjecture. In fact, R. Askey points out that the case  $z = 0$  of (1.6) (after multiplication by  $(-4\pi \sin \pi z)^{1-n}$ ) yields Dyson’s conjecture in  $n - 1$  parameters immediately. Other summation identities concerning combinations of a few coefficients in the Dyson function  $\prod (1 - u_i/u_j)^{a_i}$  can now be obtained by looking at other integral values of  $z$ .

**2. The multiple integral.** Recall that Dyson’s conjecture asserts that  $(a_1 + \cdots + a_n)! / (a_1! a_2! \cdots a_n!)$  is the constant term in the expansion of

$$(2.1) \quad \prod_{1 \leq i \neq j \leq n} \left(1 - \frac{x_i}{x_j}\right)^{a_i}.$$

Applying the binomial theorem to each factor of (2.1) produces the expanded form

$$(2.2) \quad \sum^{\#} \prod_{1 \leq i \neq j \leq n} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} x_i^{m_{ij}} x_j^{-m_{ij}},$$

where  $\sum^{\#}$  denotes an  $(n^2 - n)$  fold sum over all integral values of indices  $m_{ij}$  with  $1 \leq i \neq j \leq n$ .

Hence the Dyson conjecture now reads:

$$(2.3) \quad \sum^* \prod_{1 \leq i \neq j \leq n} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} = \frac{(a_1 + a_2 + \dots + a_n)!}{a_1! a_2! \dots a_n!},$$

where  $\sum^*$  is again an  $(n^2 - n)$  fold sum over integral values of indices  $m_{ij}$  with  $1 \leq i \neq j \leq n$ , and subject to the  $n$  constraints:

$$(2.4) \quad m_{i1} + m_{i2} + \dots + m_{in} - m_{li} - m_{2i} - \dots - m_{ni} = 0.$$

Now note that (2.3) may be viewed as a polynomial identity in  $a_n$ . This is the case because the right-hand side is

$$\frac{(a_n + 1)(a_n + 2) \dots (a_n + a_1 + a_2 + \dots + a_{n-1})}{a_1! a_2! \dots a_{n-1!}},$$

while on the left-hand side the only nonzero terms involving  $\binom{a_n}{m_{nj}}$  must have  $m_{nj}$  subject to the inequalities

$$0 \leq m_{nj} \leq m_{j1} + m_{j2} + \dots + m_{jn} \leq (n - 1)a_j.$$

Thus (2.3) is a polynomial identity in  $a_n$  (where  $a_1, a_2, \dots, a_{n-1}$  are nonnegative integers) that is valid for all nonnegative integral  $a_n$ . Hence (2.3) is valid for arbitrary complex  $a_n$ . We therefore let  $a_n = z$ , and we assume for the moment that  $z$  is nonintegral. Let

$$(2.5) \quad M(i, n - 1) = m_{i1} + m_{i2} + \dots + m_{i,n-1} - m_{1i} - m_{2i} - \dots - m_{n-1,i}$$

and let  $\sum''$  denote summation over all integral values of the  $(n - 1)(n - 2)$  indices  $m_{ij}$  with  $1 \leq i \neq j \leq n - 1$ . Hence from (2.3),

$$\begin{aligned} & \frac{\Gamma(a_1 + \dots + a_{n-1} + z + 1)}{a_1! a_2! \dots a_{n-1}! \Gamma(z + 1)} \\ &= \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \sum_{m_{1n}, \dots, m_{n-1,n}} (-1)^{m_{1n} + \dots + m_{n-1,n}} \binom{a_1}{m_{1n}} \binom{a_2}{m_{2n}} \dots \binom{a_{n-1}}{m_{n-1,n}} \\ & \quad \cdot \binom{z}{M(1, n-1) + m_{1n}} \binom{z}{M(2, n-1) + m_{2n}} \dots \binom{z}{M(n-1, n-1) + m_{n-1,n}} \\ & \quad \cdot (-1)^{M(1, n-1) + M(2, n-1) + \dots + M(n-1, n-1) + m_{1n} + m_{2n} + \dots + m_{n-1,n}} \\ &= \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \binom{z + a_1}{M(1, n-1) + a_1} \binom{z + a_2}{M(2, n-1) + a_2} \\ & \quad \dots \binom{z + a_{n-1}}{M(n-1, n-1) + a_{n-1}} \end{aligned}$$

(by the Chu-Vandermonde summation [2, p. 3])

$$\begin{aligned}
 &= \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \prod_{s=1}^{n-1} \frac{\Gamma(z + a_s + 1)}{(M(s, n-1) + a_s)! \Gamma(z - M(s, n-1) + 1)} \\
 &= \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \prod_{s=1}^{n-1} \frac{\Gamma(z + a_s + 1) \Gamma(M(s, n-1) - z)}{(M(s, n-1) + a_s)!} \left( \frac{\sin \pi(-z)}{\pi} \right) \\
 &= \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \left\{ \prod_{s=1}^{n-1} B(z + a_s + 1, M(s, n-1) - z) \right\} (-1)^{n-1} \left( \frac{\sin \pi z}{\pi} \right)^{n-1} \\
 &\hspace{15em} \text{(where } B(p, q) \text{ is the } \beta\text{-function [10, p. 256])} \\
 &= \frac{(-1)^{n-1}}{(\Gamma(z)\Gamma(1-z))^{n-1}} \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} \right\} \\
 &\quad \cdot \prod_{s=1}^{n-1} \left\{ \int_{P_s}^{(1+,0+,1-,0-)} u_s^{M(s,n-1)-z-1} (1-u_s)^{z+a_s} du_s \times \frac{1}{4 \sin^2 \pi z} \right\} \\
 &= \frac{\Gamma(z)^{n-1} \Gamma(1-z)^{n-1}}{(2\pi i)^{2n-2}} \int_{P_1}^{(1+0+,1-,0-)} \cdots \int_{P_{n-1}}^{(1+,0+,1-,0-)} \\
 &\quad \cdot (u_1 \cdots u_{n-1})^{-z-1} (1-u_1)^{z+a_1} \cdots (1-u_{n-1})^{z+a_{n-1}} \\
 &\quad \cdot \sum'' \left\{ \prod_{1 \leq i \neq j \leq n-1} (-1)^{m_{ij}} \binom{a_i}{m_{ij}} u_i^{m_{ij}} u_j^{-m_{ij}} \right\} du_1 \cdots du_{n-1} \\
 &= \frac{\Gamma(z)^{n-1} \Gamma(1-z)^{n-1}}{(2\pi i)^{2n-2}} \int_{P_1}^{(1+,0+,1-,0-)} \cdots \int_{P_{n-1}}^{(1+,0+,1-,0-)} \\
 &\quad \cdot (u_1 \cdots u_{n-1})^{-z-1} (1-u_1)^{z+a_1} \cdots (1-u_{n-1})^{z+a_{n-1}} \\
 &\quad \cdot \prod_{1 \leq i \neq j \leq n-1} \left( 1 - \frac{u_i}{u_j} \right)^{a_i} du_1 \cdots du_{n-1}, \hspace{10em} \text{(by (2.2)).}
 \end{aligned}$$

Thus (1.6) is established.

**3. The recurrence for the constant term.** I. J. Good [5] proved the Dyson conjecture by proving that the entire expression given in (2.1) satisfies the same recurrence as the multinomial coefficient. This result together with some straightforward boundary conditions yields the full theorem. When one undertakes the same approach for (1.1), all sorts of complications arise. D. Zeilberger has carried out the modifications necessary to give a ‘‘Good’’ proof of the  $q$ -Dyson conjecture when  $n = 3$ ; however, his methods do not obviously extend to larger  $n$ .

Our object here is to modify Good’s approach in order to prove directly that the left side of (2.3) satisfies the defining recurrence and boundary conditions necessary to identify it with the right side.

Let  $L_n(a_1, a_2, \dots, a_n)$  denote the left side of (2.3). For each pair  $(h, k)$  where  $1 \leq h \neq k \leq n$ , we define

$$(3.1) \quad E_{hk} L_n(a_1, a_2, \dots, a_n) = \sum^* \prod_{1 \leq i \neq j \leq n} (-1)^{m_{ij} - \phi_{ij}} \binom{a_i}{m_{ij} - \phi_{ij}},$$

where  $\sum^*$  is subjected to the same summation constraints as before,  $\phi_{ij} = 1$  if  $i = h$  and  $j = k$ , and  $\phi_{ij} = 0$  otherwise. By the recurrence for binomial coefficients

$$(3.2) \quad (1 - E_{hk}) L_n(a_1, a_2, \dots, a_n) = \sum^* \prod_{1 \leq i \neq j \leq n} (-1)^{m_{ij}} \binom{a_i + \phi_{ij}}{m_{ij}}.$$

Consequently

$$(3.3) \quad \prod_{1 \leq h \neq k \leq n} (1 - E_{hk}) L_n(a_1 - 1, a_2 - 1, \dots, a_n - 1) = L_n(a_1, a_2, \dots, a_n).$$

Next we note that

$$(3.4) \quad E_{hk} E_{kh} L_n(a_1, \dots, a_n) = L_n(a_1, \dots, a_n),$$

because we may replace both  $m_{hk}$  and  $m_{kh}$  by  $m_{hk} + 1$  and  $m_{kh} + 1$  respectively without altering conditions of (2.4).

Furthermore, if  $h \neq g$ ,

$$(3.5) \quad E_{hg} E_{kg} L_n(a_1, \dots, a_n) = E_{hg} L_n(a_1, \dots, a_n),$$

because replacement of  $m_{hk}$  and  $m_{kg}$  by  $m_{hk} + 1$  and  $m_{kg} + 1$  respectively in the conditions of (2.4) yields new summation conditions identical to the conditions obtained when only  $m_{hg}$  is replaced by  $m_{hg} + 1$ .

The algebraic reduction laws given by (3.4) and (3.5) for the operators  $E_{ij}$  show that (when operating on  $L_n(a_1, \dots, a_n)$ ) the ring of polynomials in the  $E_{ij}$  with complex coefficients is isomorphic with the ring of polynomials of degree zero over the complex numbers generated by  $x_1/x_2, x_1/x_3, \dots, x_n/x_{n-2}, x_n/x_{n-1}$ , where the  $x_1, x_2, \dots, x_n$  are commutative variables. Indeed the isomorphism is merely

$$E_{ij} \rightarrow x_i x_j^{-1}.$$

Next we recall the fundamental identity from Good's proof

$$1 = \sum_{i=1}^n \prod_{\substack{j=1 \\ j \neq i}}^n \left(1 - \frac{x_i}{x_j}\right)^{-1},$$

which rationalizes to

$$(3.6) \quad \prod_{1 \leq i \neq j \leq n} \left(1 - \frac{x_i}{x_j}\right) = \sum_{h=1}^n \prod_{\substack{1 \leq i \neq j \leq n \\ i \neq h}} \left(1 - \frac{x_i}{x_j}\right).$$

Hence

$$\begin{aligned} L_n(a_1, \dots, a_n) &= \prod_{1 \leq i \neq j \leq n} (1 - E_{ij}) L_n(a_1 - 1, a_2 - 1, \dots, a_n - 1) && \text{(by(3.3))} \\ (3.7) \quad &= \sum_{h=1}^n \prod_{\substack{1 \leq i \neq j \leq n \\ i \neq h}} (1 - E_{ij}) L_n(a_1 - 1, a_2 - 1, \dots, a_n - 1) && \text{(by(3.6))} \\ &= L_n(a_1 - 1, a_2, \dots, a_n) + L_n(a_1, a_2 - 1, a_3, \dots, a_n) + \dots \\ &\quad + L_n(a_1, a_2, \dots, a_{n-1}, a_n - 1). \end{aligned}$$

Now it is obvious that

$$(3.8) \quad L_n(0, 0, \dots, 0) = 1.$$

Furthermore if a specific  $a_j = 0$ , then the only nonzero terms must have  $m_{j1} = m_{j2} = \dots = m_{jn} = 0$ . From (2.4) with  $i = j$  we see that the only nonnegative solutions possible also require  $m_{1j} = m_{2j} = \dots = m_{nj} = 0$ . Thus

$$(3.9) \quad L_n(a_1, \dots, a_{j-1}, 0, a_{j+1}, a_n) = L_{n-1}(a_1, \dots, a_{j-1}, a_{j+1}, \dots, a_n).$$

Since (3.7), (3.8), and (3.9) uniquely define the multinomial coefficient  $(a_1 + \cdots + a_n)! / (a_1! a_2! \cdots a_n!)$ , we must have

$$L(a_1, \dots, a_n) = \frac{(a_1 + \cdots + a_n)!}{(a_1! a_2! \cdots a_n!)},$$

which is the assertion of the Dyson conjecture.

**4. Conclusion.** The problem with the integral of § 2 is that for convergence we are forced to use the Pochhammer contour integrals instead of simple real integration on  $[0, 1]$ . This problem appears not to arise for Askey when he treats the  $q$ -analogue, since the  $q$ -analogue of the  $\beta$ -integral has a wider domain of validity.

The proof in § 3 can be immediately modified to provide an assault on the  $q$ -Dyson conjecture. A barrier to a proof now arises because no applicable analogue of (3.6) has been found. At least for  $n = 2, 3, 4$ , however, extensive reduction of the problem can be achieved using the algebra of the operators  $E_{hk}$  obtained from (3.4) and (3.5). Such reductions combined with various  $q$ -series transformations might be a useful approach to the  $q$ -Dyson conjecture.

#### REFERENCES

- [1] G. E. ANDREWS, *Problems and prospects for basic hypergeometric functions*, Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975.
- [2] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London, 1935 (Reprinted: Hafner, New York, 1964).
- [3] A. C. DIXON, *Summation of a certain series*, Proc. London Math. Soc., 35 (1903), pp. 285–289.
- [4] F. J. DYSON, *Statistical theory of energy levels of complex systems I*, J. Math. Phys., 3 (1962), pp. 140–156.
- [5] I. J. GOOD, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.
- [6] J. GUNSON, *Proof of a conjecture by Dyson in the statistical theory of energy levels*, J. Math. Phys., 3 (1962), pp. 752–753.
- [7] I. G. MACDONALD, *Affine root systems and Dedekind's  $\eta$ -function*, Inv. Math., 15 (1972), pp. 91–143.
- [8] M. L. MEHTA, *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.
- [9] A. SELBERG, *Bemerkninger om et multipelt integral*, Norsk Matematisk Tidsskrift, 26 (1944), pp. 71–78.
- [10] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, 4ed., Cambridge University Press, London, 1927.
- [11] K. WILSON, *Proof of a conjecture by Dyson*, J. Math. Phys. 3 (1962), pp. 1040–1043.

## ON A NONLINEAR HYPERBOLIC VOLTERRA EQUATION\*

OLOF J. STAFFANS†

**Abstract.** We study questions of existence, boundedness and asymptotic behavior of the solutions of the initial value problem

$$u_t(t, x) - \int_0^t a(t-s)\sigma(u_x(s, x))_x ds = f(t, x), \quad 0 < t < \infty, \quad x \in R,$$

(\*)

$$u(0, x) = u_0(x), \quad x \in R.$$

Here  $a: R^+ = [0, \infty) \rightarrow R$ ,  $\sigma: R \rightarrow R$ ,  $f: R^+ \times R \rightarrow R$ ,  $u_0: R \rightarrow R$  are given, sufficiently smooth functions, and the subscripts  $t$  and  $x$  denote partial derivatives. If  $a(t) \equiv 1$ , then (\*) reduces to a nonlinear wave equation, and it is well known that in this case classical solutions of (\*) do not in general exist for all time. However, we show that for a large class of kernels of physical importance equation (\*) has global classical solutions for small data. This class of kernels includes all those which are nonconstant, nonnegative, nonincreasing, convex and sufficiently smooth. We also analyze the asymptotic behavior of the solutions.

**1. Introduction.** We study questions of existence, boundedness and asymptotic behavior of the solutions of the initial value problem

$$u_t(t, x) - \int_0^t a(t-s)\sigma(u_x(s, x))_x = f(t, x), \quad 0 < t < \infty, \quad x \in R,$$

(E)

$$u(0, x) = u_0(x), \quad x \in R,$$

for small data  $f, u_0$ . Here  $a: R^+ = [0, \infty) \rightarrow R$ ,  $\sigma: R \rightarrow R$ ,  $f: R^+ \times R \rightarrow R$ ,  $u_0: R \rightarrow R$  are given, sufficiently smooth functions. The subscripts  $t$  and  $x$  denote partial derivatives.

The physical interpretation of (E) varies according to the properties of  $a$ . If  $a(\infty) = 0$ , then (E) represents a mathematical model for heat flow in an unbounded bar made out of a material with memory. If  $a(\infty) > 0$ , then (E) is the equation of motion of an unbounded viscoelastic bar. In particular, if  $a(t) \equiv 1$ , then (E) becomes the nonlinear wave equation

$$u_{tt}(t, x) - \sigma(u_x(t, x))_x = f_t(t, x), \quad 0 < t < \infty, \quad x \in R,$$

(1.1)

$$u(0, x) = u_0(x), \quad x \in R,$$

$$u_t(0, x) = u_1(x), \quad x \in R,$$

where  $u_1(x) = f(0, x)$ .

Although (1.1) is a special case of (E), it is too degenerate to be very representative. There is no damping mechanism included in (1.1), and (1.1) describes a physical system which does not lose any energy due to friction or viscosity. It is well known [2], [5] that the lack of damping in general prevents (1.1) from having classical solutions for all time independently of how small and how smooth the data is. If  $a$  satisfies some natural physical assumptions (which exclude the constant case), then (E) behaves more like the

\* Received by the editors July 12, 1979, and in revised form November 14, 1979.

† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

damped nonlinear wave equation

$$\begin{aligned}
 (1.2) \quad & u_{tt}(t, x) + \alpha u_t(t, x) - \sigma(u_x(t, x))_x = g(t, x), \\
 & u(0, x) = u_0(x), \quad x \in \mathbb{R}, \\
 & u_t(0, x) = u_1(x), \quad x \in \mathbb{R},
 \end{aligned}$$

where  $\alpha$  is some properly chosen positive constant (depending on  $a$ ), and  $g$  is some properly chosen function (here we have taken  $a(0) = 1$ ). Nishida [7] has shown that (1.2) (with  $g \equiv 0$ ) possesses classical global solutions for sufficiently small data. Nishida’s proof depends heavily on the concept of Riemann invariants, and it does not generalize to several space dimensions. Later Matsumura [6] applied an energy method to prove the existence of small global solutions of (1.2). The energy method has the advantage that it is not tied to one space dimension.

The fact that (E) behaves in the same way as (1.2) was discovered by MacCamy [3], [4]. He differentiates (E), transforms the resulting equation using the resolvent of the differentiated kernel, and ends up with a linear perturbation of (1.1), which we in the sequel call “the transformed equation” (see (3.4) below). Moreover, he develops estimates [3, §§ 3–4], [4, §§ 3–4] (in the following called “MacCamy’s damping estimates”) which show that there is a built-in damping mechanism in the transformed equation that makes it behave in the same way as (1.2). Especially in the viscoelastic case these estimates are quite complicated. MacCamy establishes the existence of a global classical solution of (E) for sufficiently small data, and he shows that this solution tends to zero as  $t$  tends to infinity. He omits the proof of existence of a local solution of the transformed equation, but such a proof was outlined by Nohel [8]. MacCamy follows Nishida and uses Riemann invariants, which means that the results are strictly “one-space-dimensional”. In a recent paper [1] Dafermos and Nohel combine MacCamy’s damping estimates with an energy method similar to the one of Matsumura to generalize and simplify MacCamy’s results (they also correct some errors in [3], [4]). In particular, Dafermos and Nohel are no longer tied to one space dimension. In this work we essentially use the same local existence result as in [1]. However, instead of using MacCamy’s damping estimates we develop new estimates, which are based directly on (E) rather than on the transformed equation. We use an energy method similar to the one in [1], but we replace differentiations with respect to  $t$  by differentiations with respect to  $x$ . As a result, we obtain a generalization of [1].

**2. Summary of results.** Our first goal is to develop a local existence theorem for (E). This theorem may be considered as a modification of the local existence theorem in [1]. The basic assumption is that the data are sufficiently smooth.

When the local existence of solutions has been established we develop global estimates, which enable us to continue the solution for all time, as well as give us information about the asymptotic behavior of solutions. These estimates require the data to be not only smooth but also small enough.

Before stating our local existence theorem we introduce some notations and assumptions.  $C^n$  stands for  $n$  times continuously differentiable functions, LAC for locally absolutely continuous functions, and  $L^1$ ,  $L^2$  and  $L^\infty$  are the usual Lebesgue spaces. In our notations for function spaces we throughout omit the domain of the independent variable (which is either  $\mathbb{R}$  or  $\mathbb{R}^+$ ) whenever no confusion is likely to arise. In particular, in  $(f_{loc})$  below  $L^2$  stands for  $L^2(\mathbb{R})$ , and  $L^1_{loc}(L^2)$  stands for the space of locally integrable functions on  $\mathbb{R}^+$  with values in  $L^2(\mathbb{R})$ . Whenever we write an



assumption containing conditions on derivatives we implicitly assume that the functions are smooth enough so that the needed derivatives can be computed (or alternatively, interpret the differentiations in the distribution sense). Besides the subscripts  $t$  and  $x$  we use a prime to denote the derivative of a function of one variable. In particular  $(a_{loc})$  below is equivalent to  $a, a' \in LAC, a(0) > 0$ .

In our local existence theorem we assume the following:

- $(a_{loc}) \quad a'' \in L^1_{loc}, \quad a(0) > 0,$
- $(\sigma_{loc}) \quad \sigma \in C^3, \quad \sigma(0) = 0, \quad 0 < p_0 \leq \sigma'(\xi) \leq p_1, \quad \xi \in R,$
- $(f_{loc}) \quad f(0, x) = u_1(x), \quad x \in R \text{ with } u_1, u_{1x}, u_{1xx} \in L^2,$   
 $f_b, f_{tx} \in L^1_{loc}(L^2), \text{ and } f \text{ can be written in the form}$   
 $f = g + h, \text{ where } g_{ttt}, h_{ttt} \in L^1_{loc}(L^2).$
- $(u_0) \quad u_{0x}, u_{0xx}, u_{0xxx} \in L^2.$

**THEOREM 1 (local existence).** *Let the assumptions  $(a_{loc}), (\sigma_{loc}), (f_{loc})$  and  $(u_0)$  hold. Then there exists a unique solution  $u$  of (E) defined on a maximal interval  $[0, T_0) \times R$ , where  $0 < T_0 \leq \infty$ , satisfying*

$$(2.1) \quad u_t, u_x, u_{tx}, u_{xx}, u_{ttx}, u_{xxx} \in L^\infty_{loc}([0, T_0); L^2).$$

Furthermore, if

$$(2.2) \quad u_t, u_x, u_{tx}, u_{xx}, u_{ttx}, u_{xxx} \in L^\infty([0, T_0); L^2),$$

then  $T_0 = \infty$ .

Compared to the local existence theorem in [1] our assumption  $(a_{loc})$  is substantially weaker than the corresponding assumption in [1] (which roughly requires  $a'', a''' \in C \cap L^1$  in addition to  $a(0) > 0$ ). Furthermore, we do not need any conditions on  $f_{tt}$  and  $f_{ttt}$  as Dafermos and Nohel do, and we can manage without conditions of  $f_{ttt}$  (i.e., take  $g = 0$  in  $(f_{loc})$ ). In (2.1) and (2.2) we have left out conditions on those derivatives of  $u$  which involve more than one differentiation with respect to  $t$ . There is no need to keep track of these derivatives in the proof of Theorem 1, because they can easily be estimated directly from (2.1), (2.2). As a matter of fact, (E),  $(a_{loc}), (\sigma_{loc}), (2.1)$  and (2.2) imply

$$u_{tt} - f_b, u_{ttt} - f_{tx}, u_{ttt} - f_{tt} \in L^\infty_{loc}([0, T_0); L^2)$$

(this is proved in the same manner as (2.5) below). Thus, if we add

$$f_b, f_{tx}, f_{tt} \in L^\infty_{loc}(R^+)$$

to the hypothesis of Theorem 1, then our local solution has the same regularity as the local solution in [1].

The proof of Theorem 1, which is given in § 3, follows very closely the proof of Theorem 3.1 in [1]. In particular, we work with the transformed equation (3.4) below. However, as soon as we have obtained local existence of solutions we discard the transformed equation and work directly with (E) in order to get global estimates. The

local conditions used in Theorem 1 are now replaced by global conditions:

- (a)  $a$  is strongly positive definite, and  $a', a'' \in L^1$ ;
- ( $\sigma$ )  $\sigma \in C^3$ ,  $\sigma(0) = 0$  and  $\sigma'(0) > 0$ ;
- (f)  $f(0, x) = u_1(x)$ ,  $x \in R$ , with  $u_1, u_{1x}, u_{1xx} \in L^2$ , and  $f = f_1 + f_2 + f_3$ ;
- (f<sub>1</sub>)  $f_1 \in L^\infty(L^2)$ ,  $f_{1x} \in L^1 \cap L^\infty(L^2)$ ,  $f_{1xx} \in L^2 \cap L^\infty(L^2)$ ,  $f_{1xxx} \in L^2(L^2)$ ;
- (f<sub>2</sub>)  $f_2, f_{2b}, f_{2x}, f_{2tx}, f_{2xx}, f_{2txx} \in L^2(L^2)$ ;
- (f<sub>3</sub>)  $f_3, f_{3x}, f_{3xx} \in L^\infty(L^2)$ ,  $f_{3b}, f_{3tx}, f_{3txx} \in L^1(L^2)$ .

THEOREM 2. Let (a), ( $\sigma$ ), (f<sub>10c</sub>), (f), (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) and (u<sub>0</sub>) hold. In addition, if  $a(\infty) = 0$ , then suppose that  $f_3 = 0$ . If the appropriate  $L^p$ -norms of the functions (and derivatives) listed in (f), (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) and (u<sub>0</sub>) are sufficiently small, then there exists a unique global solution  $u$  of (E), and it satisfies

- (2.3)  $u_b, u_x, u_{tx}, u_{xx}, u_{txx}, u_{xxx} \in L^\infty(L^2)$ ;
- (2.4)  $u_{tx}, u_{xx}, u_{txx}, u_{xxx} \in L^2(L^2)$ ;
- (2.5)  $u_{tt} - f_b, u_{txx} - f_{tx}, u_{ttt} - f_{tt} \in L^2 \cap L^\infty(L^2)$ ;
- (2.6)  $u_{xx}, u_{tt} - f_t \rightarrow 0$  in  $L^2$  as  $t \rightarrow \infty$ ;
- (2.7)  $u_x, u_{xx}, u_{tt} - f_t \rightarrow 0$  uniformly as  $t \rightarrow \infty$ .

Here the primary conclusions are (2.3), (2.4), and they imply global existence (by Theorem 1) as well as (2.5)–(2.7). There is some redundancy in the hypothesis of Theorem 2, as (f), (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) almost imply (f<sub>10c</sub>); i.e., the only part of (f<sub>10c</sub>) which is missing in (f), (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) is  $f_{1b}, f_{1tx}, f_{1txx} \in L^1_{loc}(L^2)$ .

A very small strengthening of the hypothesis of Theorem 2 yields two additional conclusions:

COROLLARY 3. In addition to the assumption of Theorem 2, suppose that  $f$  is uniformly continuous as a function of  $t$  with values in  $L^2$ . Then the solution  $u$  of (E) established in Theorem 2 satisfies

- (2.8)  $u_{tx} \rightarrow 0$  in  $L^2$  as  $t \rightarrow \infty$ ,
- (2.9)  $u_b, u_{tx} \rightarrow 0$  uniformly as  $t \rightarrow \infty$ .

Also, in the special case when  $a \in L^1$  one can draw additional conclusions:

COROLLARY 4. In addition to the assumption of Theorem 2, suppose that  $a \in L^1$ . Then the solution  $u$  of (E), established in Theorem 2, satisfies

- (2.10)  $u_t - f \in L^2(L^2)$ ,
- (2.11)  $u_t - f, u_{tx} - f_x \rightarrow 0$  in  $L^2$  and uniformly as  $t \rightarrow \infty$ .

The assumption of Theorem 2 is not the weakest possible, as we have tried to keep it reasonably simple. For example, (a) could be replaced by the hypothesis used on  $a$  in [14] combined with (a<sub>10c</sub>). Also (f), (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) do not cover the whole spectrum of conditions which could be used on  $f$ . One can, e.g., replace  $f_{1xxx} \in L^2(L^2)$  by  $f_{1xxx} \in L^1(L^2)$ , and one can use “mixed” versions of (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>). For example, in the case  $a(\infty) > 0$ , a function  $f$  satisfying

$$f, f_x \in L^\infty(L^2); \quad f_b, f_{tx} \in L^1(L^2); \quad f_{xx}, f_{txx} \in L^2(L^2)$$

can be allowed. This will be evident from the proof of Theorem 2, given in § 5. If one slightly strengthens the condition on  $a$ , then one can also replace some of the derivatives with respect to  $x$  in  $(f_1), (f_2), (f_3)$  by derivatives with respect to  $t$ . We discuss this further in § 6.

Theorem 2 generalizes Dafermos' and Nohel's global result for the heat flow equation [1, Thm. 4.1]. Dafermos and Nohel assume that  $a$  satisfies

$$(2.12) \quad a \in C^3; \quad a, a', a'', a''' \in L^\infty; \quad t^j a^{(m)}(t) \in L^1; \quad j = 0, 1, 2, 3; \quad m = 0, 1, 2, 3,$$

plus something which is equivalent to the strong positive definiteness of  $a$ . They also require  $f$  to satisfy

$$(2.13) \quad f, f_t, f_x, f_{tt}, f_{tx}, f_{xx}, f_{utt}, f_{utx}, f_{ttx} \in L^2(L^2),$$

which clearly implies  $(f_{loc}), (f)$  and  $(f_1), (f_2), (f_3)$  (e.g., take  $f_1 = f_3 = 0$ ). The conclusions of the two theorems are equivalent.

Theorem 2 can also be applied to the viscoelastic equation studied by Dafermos and Nohel, and it overlaps their global result for this equation [1, Thm. 5.1]. One gets the viscoelastic case by taking  $a(\infty) > 0$ . Dafermos and Nohel assume that  $a$  satisfies

$$(2.14) \quad \begin{aligned} &a \in C^3; \quad a, a', a'', a''' \in L^\infty; \quad a(\infty) > 0; \\ &(-1)^m a^{(m)} \geq 0, \quad m = 0, 1, 2; \quad a' \neq 0; \\ &t^j [a - a(\infty)]^{(m)} \in L^1, \quad j = 0, 1, 2, 3, \quad m = 0, 1, 2, 3. \end{aligned}$$

As is well known, this implies (a) (see [9, Cor. 2.2]). They replace our condition on  $f$  by

$$(2.15) \quad f_t, f_{tt} \in L^1(L^2), \quad f_{tx}, f_{utt}, f_{ttx} \in L^2(L^2).$$

Clearly, (2.15) does not directly fit into  $(f), (f_1), (f_2), (f_3)$  as it does not involve any second order derivatives with respect to  $x$ . However, when  $a(\infty) > 0$ , then one can also in our theorem replace  $(f), (f_1), (f_2), (f_3)$  by a weakened version of (2.15). We discuss this further in § 6.

The outline of the remainder of this paper is the following. In § 3 we prove Theorem 3, and state a lemma on how additional smoothness of the data is reflected in additional smoothness of the solution. This lemma is needed in the proof of Theorem 2. We have collected some inequalities for positive definite functions in § 4, and we prove Theorem 2 and Corollaries 3 and 4 in § 5. Finally, in § 6 we discuss possible modifications of Theorems 1 and 2.

**3. Proof of the local existence theorem.** To simplify the notations we normalize  $a$  so that  $a(0) = 1$ ; (i.e., divide  $a$  by  $a(0)$  and multiply  $\sigma$  by  $a(0)$ ).

In the proof of the local existence theorem we follow MacCamy and transform (E) into (1.1) with an additional linear perturbation. Let  $r$  be the resolvent of  $a'$ , i.e., the solution of the resolvent equation

$$(3.1) \quad r(t) + (a' * r)(t) = -a'(t), \quad t \in \mathbb{R}^+.$$

Here and below  $*$  stands for convolution with respect to the time variable, i.e.,

$$(a' * r)(t) = \int_0^t a'(t-s)r(s) ds.$$

By standard theory for Volterra equations,  $(a_{loc})$  implies

$$(r_{loc}) \quad r, r' \in L^1_{loc}(\mathbb{R}^+).$$

Differentiating (E) with respect to  $t$  one obtains

$$(3.2) \quad u_t = \sigma(u_x)_x + a' * \sigma(u_x)_x + f_t$$

(for simplicity we have omitted the arguments). Convolving (3.2) with  $r$ , adding the result to (3.2), and using (3.1) one gets

$$(3.3) \quad u_{tt} - \sigma(u_x)_x = f_t + r * f_t - r * u_{tt}$$

An integration by parts in the last two terms transforms this into

$$(3.4) \quad u_{tt} - \sigma(u_x)_x = k(u_t) + l,$$

where

$$(3.5) \quad k(u_t) = -r(0)u_t - r' * u_t,$$

$$(3.6) \quad l(t, x) = f_t(t, x) + r(0)f(t, x) + (r' * f)(t, x).$$

The crucial observation is that (3.4) differs from (1.1) only by a linear perturbation of lower order. As a consequence of this fact, a proof of local existence of solutions of (1.1) can be converted into a proof of local existence of solutions of (E). We do not give all the details of the proof as it is very similar to the proof of Theorem 3.1 in [1]. Instead we just point out the necessary modifications.

In a moment we shall fix  $T$  (sufficiently small) and look for a (local) solution of (3.4) on  $[0, T] \times R$ . Before doing so, let us introduce some more notations. We denote norms as follows:

- $\| \cdot \|$  is the norm of  $L^2$ ,
- $\| \cdot \|_p$  is the norm of  $L^p([0, T]; L^2)$ ,
- $\| \cdot \|_{\text{sup}}$  is the supremum norm over  $[0, T] \times R$ .

Here  $L^2$  throughout stands for  $L^2(R)$ , and  $p = 1, 2$  or  $\infty$ . By  $(f_{\text{loc}})$  and  $(u_0)$ , the constant  $U$  defined by

$$(3.7) \quad U^2 = \|u_1\|^2 + \|u_{1x}\|^2 + \|u_{1xx}\|^2 + p_1(\|u_{0x}\|^2 + \|u_{0xx}\|^2 + \|u_{0xxx}\|^2)$$

is finite (here  $p_1$  is the same constant as in  $(\sigma_{\text{loc}})$ ). The important properties of  $k, l$  defined in (3.5), (3.6) are the facts that  $k$  commutes with  $\partial/\partial x$ , and that

$$(3.8) \quad \|k(w)\|_{\infty} \leq K \|w\|_{\infty}, \quad w \in L^{\infty}([0, T]; L^2),$$

$$(3.9) \quad l = l_1 + l_2, \quad \text{and} \quad \|l\|_1^2 + \|l_x\|_1^2 + 2\|l_{1x}\|_{\infty}^2 + \|l_{1tx}\|_1^2 + \|l_{2xx}\|_1^2 \leq L^2$$

for some sufficiently large constants  $K$  and  $L$  (which depend on  $T$ ), and appropriately chosen functions  $l_1, l_2$ . That (3.8), (3.9) hold follows from  $(f_{\text{loc}})$ ,  $(r_{\text{loc}})$  and (3.5), (3.6). By  $(\sigma_{\text{loc}})$ , it is true that for each  $M > 0$ ,

$$(3.10) \quad |\sigma''(\xi)| + |\sigma'''(\xi)| \leq \bar{\sigma}, \quad |\xi| \leq M$$

for some sufficiently large constant  $\bar{\sigma}$  (which depends on  $M$ ).

For positive  $M$  and  $T$ , let  $X(M, T)$  denote the set of functions  $v \in C^1([0, T] \times R)$  with initial conditions  $v(0, x) = u_0(x)$ ,  $v_t(0, x) = u_1(x)$ , which satisfy

$$v, v_x, v_{tx}, v_{xx} \in \text{LAC}([0, T]; L^2);$$

$$v_{txx}, v_{xxx} \in L^{\infty}([0, T]; L^2),$$

and

$$(3.11) \quad \|\|v_t\|\|_\infty^2 + \|\|v_x\|\|_\infty^2 + \|\|v_{tx}\|\|_\infty^2 + \|\|v_{xx}\|\|_\infty^2 + \|\|v_{txx}\|\|_\infty^2 + \|\|v_{xxx}\|\|_\infty^2 \leq M^2.$$

Then  $X(M, T)$  is nonempty if  $M$  is sufficiently large. Moreover, (3.11) yields

$$(3.12) \quad \max\{\|\|v_t\|\|_{\text{sup}}, \|\|v_x\|\|_{\text{sup}}, \|\|v_{tx}\|\|_{\text{sup}}, \|\|v_{xx}\|\|_{\text{sup}}\} \leq M.$$

This follows from the fact that for any  $\psi \in L^2$  satisfying  $\psi_x \in L^2$ ,

$$(3.13) \quad \begin{aligned} |\psi(x)|^2 &\leq 2 \int_{-\infty}^x |\psi(x)| |\psi_x(x)| dx \leq 2\|\psi\| \|\psi_x\| \\ &\leq \|\psi\|^2 + \|\psi_x\|^2. \end{aligned}$$

Let  $S$  be the map which carries  $v \in X(T, M)$  into the solution  $u$  of the linear initial value problem

$$(3.14) \quad \begin{aligned} u_{tt} - \sigma'(v_x)u_{xx} &= k(v_t) + l, \\ u(0, x) &= u_0(x), \quad u_t(0, x) = u_1(x), \quad (x \in R). \end{aligned}$$

Our first goal is to show that for  $M$  large enough and  $T$  small enough  $S$  is a contraction mapping from  $X(T, M)$  into itself. It then has a unique fixed point, which will be a local solution of (3.4), hence of (E). We have defined  $S$  in a slightly different way than Dafermos and Nohel do in order to emphasize the fact that the damping factor  $\alpha$  in (1.2) plays no role in the local existence proof. However, the difference is so small that it does not affect the needed estimates (Dafermos and Nohel just replace  $r(0)v_t$  by  $r(0)u_t$ ; cf. (3.5)).

As in [1], let us assume temporarily that  $\sigma, r, u_0, f$  and  $v$  are  $C^\infty$  smooth on their domains of definition, and that  $u_{0x}, f$  and  $v$  are compactly supported on  $R$ . Then the solution  $u$  of (3.14) will be  $C^\infty$  smooth on  $R^+ \times R$ , and  $u_x$  will have compact support in  $R$  for  $t \in R^+$ .

Multiplying (3.14) by  $u_t$  and integrating over  $[0, s] \times R$  ( $0 \leq s \leq T$ ) we obtain (c.f. [1])

$$(3.15) \quad \begin{aligned} &\frac{1}{2} \int_{-\infty}^{\infty} u_t^2(s, x) dx + \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(v_x(s, x))u_x^2(s, x) dx \\ &= \frac{1}{2} \int_{-\infty}^{\infty} u_1^2 dx + \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_{0x})u_{0x}^2 dx \\ &\quad + \frac{1}{2} \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x)v_{tx}u_x^2 dx dt - \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x)v_{xx}u_t u_x dx dt \\ &\quad + \int_0^s \int_{-\infty}^{\infty} k(v_t)u_t dx dt + \int_0^s \int_{-\infty}^{\infty} lu_t dx dt. \end{aligned}$$

Thus by more or less obvious estimates which use the usual norm inequalities in the different  $L^p$ -spaces together with  $(\sigma_{\text{loc}})$  and (3.7)–(3.12),

$$(3.16) \quad \begin{aligned} \frac{1}{2}\|u_t\|^2(s) + \frac{P_0}{2}\|u_x\|^2(s) &\leq \frac{1}{2}U^2 + \frac{1}{2}\bar{\sigma}MT\|\|u_x\|\|_\infty^2 + \bar{\sigma}MT\|\|u_t\|\|_\infty\|\|u_x\|\|_\infty \\ &\quad + KMT\|\|u_t\|\|_\infty + L\|\|u_t\|\|_\infty. \end{aligned}$$

One can continue this estimate by using the inequality

$$(3.17) \quad |\alpha\beta| \leq \lambda\alpha^2 + \frac{1}{4\lambda}\beta^2, \quad \alpha, \beta \in \mathbf{R}; \quad \lambda > 0$$

with  $\alpha = \|u_t\|_\infty$ ,  $\lambda = \frac{1}{12}$  to get

$$\begin{aligned} \frac{1}{2}\|u_t\|_\infty^2 + \frac{p_0}{2}\|u_x\|_\infty^2 &\leq 2 \sup_{0 \leq s \leq T} \left( \frac{1}{2}\|u_t\|^2(s) + \frac{p_0}{2}\|u_x\|^2(s) \right) \\ &\leq \frac{1}{4}\|u_t\|_\infty^2 + (\bar{\sigma}MT + 12\bar{\sigma}^2M^2T^2)\|u_x\|_\infty^2 \\ &\quad + U^2 + 12K^2M^2T^2 + 12L^2, \end{aligned}$$

or equivalently,

$$(3.18) \quad \frac{1}{4}\|u_t\|_\infty^2 + \left( \frac{p_0}{2} - \bar{\sigma}MT - 12\bar{\sigma}^2M^2T^2 \right) \|u_x\|_\infty^2 \leq U^2 + 12K^2M^2T^2 + 12L^2.$$

Now fix  $T > 0$  arbitrarily, then choose  $M$  so large that

$$(3.19) \quad U^2 + 12L^2 \leq \frac{1}{24} \min \{1, p_0\} M^2,$$

and finally reduce the size of  $T$  (if necessary) so that

$$(3.20) \quad 12K^2M^2T^2 \leq \frac{1}{24} \min \{1, p_0\} M^2, \quad \bar{\sigma}MT + 12\bar{\sigma}^2M^2T^2 \leq \frac{p_0}{4}.$$

Then (3.18), (3.19), (3.20) yield

$$(3.21) \quad \|u_t\|_\infty^2 + \|u_x\|_\infty^2 \leq \frac{M^2}{3},$$

and we have completed the first step in the proof of the fact that for  $M$  sufficiently large and  $T$  sufficiently small,  $S$  maps  $X(M, T)$  into itself.

We still need estimates on the norms  $\|u_{tx}\|_\infty, \|u_{xx}\|_\infty, \|u_{txx}\|_\infty$  and  $\|u_{xxx}\|_\infty$ . These are obtained in the same manner as above. The only difference is that instead of multiplying (3.14) by  $u_t$  one multiplies by either  $u_{tx}(\partial/\partial x)$  or  $u_{txx}(\partial^2/\partial x^2)$ . One could also follow Dafermos and Nohel and instead multiply by  $u_{tt}(\partial/\partial t), u_{tx}(\partial/\partial x), u_{tt}(\partial^2/\partial t^2)$ , and  $u_{tx}(\partial^2/\partial t \partial x)$ , but that approach requires more smoothness of the data, and it leads to double the amount of work. We skip the proof of the fact that for the same values of  $M$  and  $T$  as above,

$$(3.22) \quad \|u_{tx}\|_\infty^2 + \|u_{xx}\|_\infty^2 \leq \frac{M^2}{3}$$

(cf. [1, line (3.8)]). Instead we multiply (3.14) by  $u_{txx}\partial^2/\partial x^2$ , and integrate over

$[0, s] \times \mathbf{R}$ ,  $0 \leq s \leq T$  to obtain (cf. [1])

$$\begin{aligned}
 & \frac{1}{2} \int_{-\infty}^{\infty} u_{txx}^2(s, x) \, dx + \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(v_x(s, x)) u_{xxx}^2(s, x) \, dx \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} u_{1xx}^2 \, dx + \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_{0x}) u_{0xxx}^2 \, dx \\
 (3.23) \quad & + \frac{1}{2} \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x) v_{tx} u_{xxx}^2 \, dx \, dt + \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x) v_{xx} u_{txx} u_{xxx} \, dx \, dt \\
 & + \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x) v_{xxx} u_{xx} u_{txx} \, dx \, dt + \int_0^s \int_{-\infty}^{\infty} \sigma'''(v_x) v_{xx}^2 u_{xx} u_{txx} \, dx \, dt \\
 & + \int_0^s \int_{-\infty}^{\infty} k(v_{txx}) u_{txx} \, dx \, dt + \int_0^s \int_{-\infty}^{\infty} l_{xx} u_{txx} \, dx \, dt.
 \end{aligned}$$

Most of the terms in (3.23) can be estimated exactly as in (3.15), but the fifth term on the right-hand side deserves special attention. Use (3.11), (3.12), (3.13) and (3.22) to get

$$\begin{aligned}
 (3.24) \quad \left| \int_0^s \int_{-\infty}^{\infty} \sigma''(v_x) v_{xxx} u_{xx} u_{txx} \, dx \, dt \right| & \leq \bar{\sigma} M T \| \| u_{xx} \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| \\
 & \leq 2^{1/2} \bar{\sigma} M T \| \| u_{xx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \\
 & \leq 2^{1/2} \bar{\sigma} M^{3/2} T \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \|^{1/2}.
 \end{aligned}$$

In the last term of (3.23), write  $l_{xx} = l_{1xx} + l_{2xx}$ , and integrate by parts, first with respect to  $x$  and then with respect to  $t$  to obtain

$$\begin{aligned}
 \int_0^s \int_{-\infty}^{\infty} l_{xx} u_{txx} \, dx \, dt &= \int_0^s \int_{-\infty}^{\infty} l_{1tx} u_{xxx} \, dx \, dt \\
 &+ \int_0^s \int_{-\infty}^{\infty} l_{2xx} u_{txx} \, dx \, dt + \int_{-\infty}^{\infty} l_{1x}(0, x) u_{0xxx}(x) \, dx \\
 &- \int_{-\infty}^{\infty} l_{1x}(s, x) u_{xxx}(s, x) \, dx.
 \end{aligned}$$

Hence by (3.9)

$$(3.25) \quad \left| \int_0^s \int_{-\infty}^{\infty} l_{xx} u_{txx} \, dx \, dt \right| \leq L (\| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| ).$$

By using  $(\sigma_{loc})$ , (3.7), (3.8), (3.10), (3.11), (3.12), (3.22), (3.24), (3.25) and standard  $L^p$ -inequalities in (3.23) one gets

$$\begin{aligned}
 \frac{1}{2} \| \| u_{txx} \| \|^2(s) + \frac{p_0}{2} \| \| u_{xxx} \| \|^2(s) & \leq \frac{1}{2} U^2 + \frac{1}{2} \bar{\sigma} M T \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \\
 & + 2^{1/2} \bar{\sigma} M^{3/2} T \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \|^{1/2} \\
 & + \bar{\sigma} M^3 T \| \| u_{txx} \| \| u_{xxx} \| \\
 & + K M T \| \| u_{txx} \| \| u_{xxx} \| + L (\| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| \| u_{txx} \| \| u_{xxx} \| ).
 \end{aligned}$$

This inequality is of the same type as (3.16), although it contains more terms. A similar computation to the one following (3.16) shows that by first increasing the size of  $M$  (which may necessitate a decrease of the size of  $T$  because of (3.20)), and then further decreasing the size of  $T$  one obtains

$$\| \|u_{txx}\| \|_{\infty}^2 + \| \|u_{xxx}\| \|_{\infty}^2 \leq \frac{M^2}{3}.$$

This together with (2.21), (2.22) shows that  $S$  maps a dense subset of  $X(M, T)$  (i.e., the subset of  $C^\infty$ -functions with compact support) into itself, provided the data is smooth and compactly supported on  $R$ .

The remainder of the proof of Theorem 1 is an almost exact copy of the proof of Theorem 3.1 in [1], and for this reason we do not give any further details. One equips  $X(M, T)$  with a complete metric, and shows that by further reducing the size of  $T$  (if necessary) one can make  $S$  a strict contraction. This together with the fact that  $S$  maps a dense subset of  $X(M, T)$  into  $X(M, T)$  implies that  $S$  maps all of  $X(M, T)$  into itself, provided the data is smooth and compactly supported. One gets rid of the assumption that the data is smooth and compactly supported on  $R$  by showing that  $S$  depends continuously on the data (this particular step is skipped in [1], but it involves the same type of estimates as the proof of the fact that  $S$  is a contraction). Thus,  $S$  has a unique fixed point in  $X(M, T)$ . This fixed point is a solution of (3.4) on  $[0, T] \times R$  with the right initial data, and so it is also a solution of (E) on the same strip. Define  $T_0 \leq \infty$  as the length of the maximal interval of existence of a solution of (E) which satisfies (2.1). Then  $u$  is the unique solution of (E) on  $[0, T_0)$ . If (2.2) holds and  $T_0 < \infty$ , then the solution can be extended beyond  $T_0$  (cf. [1]) and so (2.2) implies  $T_0 = \infty$ .

To get the global estimates needed in the proof of Theorem 2 we have to do the same thing as above, namely approximate the exact solution of (E) by a smoother solution, so that certain computations can be carried out. This is possible due to the following lemma.

LEMMA 3.1. *Let the assumption of Theorem 1 be satisfied. Under the additional hypothesis*

$$\begin{aligned} \sigma &\in C^4; \\ u_{0xxxx}, \quad u_{1xxx} &\in L^2; \\ g_{tuxx}, \quad h_{txxx} &\in L^1_{loc}(L^2) \end{aligned}$$

*the solution  $u$  of (E), established in Theorem 1, has the additional property*

$$u_{txxx}, u_{xxxx} \in L^\infty_{loc}([0, T_0); L^2).$$

We omit the proof of Lemma 3.1, which is very similar to the proof of the corresponding Theorem 3.2 in [1]. Basically, one assumes that the function  $v$  in (3.11), (3.14) satisfies

$$\| \|v_{txxx}\| \|_{\infty}^2 + \| \|v_{xxxx}\| \|_{\infty}^2 \leq N^2,$$

multiplies (3.14) by  $u_{txxx}(\partial^3/\partial x^3)$ , and makes a computation similar to those above to get

$$\| \|u_{txxx}\| \|_{\infty}^2 + \| \|u_{xxxx}\| \|_{\infty}^2 \leq N^2.$$

The crucial observation is that the constant  $T$  can be chosen independently of  $N$ . This implies that the length of the maximal interval of the smoother solution is controlled solely by the derivatives listed in (2.1), and so the solution stays smooth for as long as it exists.



**4. Inequalities for positive definite functions.** Our global estimates are based on a number of inequalities for positive definite functions. All of these have been used earlier elsewhere, but for the reader's convenience we list them below, and outline some of the proofs.

In this section, let  $H$  be a complex Hilbert space, and let  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$  denote the inner product and norm of  $H$ . Later on, when applying the inequalities listed here, we take  $H = L^2(\mathbb{R})$ . We throughout suppose that  $\varphi$  is a function satisfying  $\varphi \in L^1_{loc}(\mathbb{R}^+; H)$ , and define

$$(4.1) \quad Q(\varphi, T, a) = \int_0^T \left\langle \varphi(t), \int_0^t a(t-s)\varphi(s) ds \right\rangle dt$$

for  $T > 0$ .

LEMMA 4.1. *Let  $a$  be continuous and positive definite. Then, for each  $\varphi \in L^1_{loc}(\mathbb{R}^+; H)$  and  $T > 0$ ,*

$$\left\| \int_0^T a(T-s)\varphi(s) ds \right\|^2 \leq 2a(0)Q(\varphi, T, a).$$

This lemma is the same as [11, Lemma 6.1]. In [11] it is only formulated for the case when  $H = \mathbb{R}$ , but the same proof applies in the general case.

LEMMA 4.2. *Let  $k$  satisfy  $k, k' \in L^1(\mathbb{R}^+)$ . Then, for each  $\varphi \in L^1_{loc}(\mathbb{R}^+; H)$  and  $T > 0$ .*

$$\int_0^T \left\| \int_0^t k(t-s)\varphi(s) ds \right\|^2 dt \leq C_k Q(\varphi, T, e),$$

where  $C_k = [\int_0^\infty |k(t)| dt]^2 + 4[\int_0^\infty |k'(t)| dt]^2$ , and  $e(t) = e^{-t}$  ( $t \in \mathbb{R}^+$ ).

The scalar version of Lemma 4.2 has been used in several places; see e.g., [13, Lemma 2.2].

*Outline of proof.* Define

$$(4.2) \quad \varphi_T(t) = \begin{cases} \varphi(t), & 0 \leq t \leq T, \\ 0, & \text{otherwise.} \end{cases}$$

By the Plancherel identity and the fact that convolution is mapped into pointwise multiplication by the Fourier transform,

$$(4.3) \quad \begin{aligned} \int_0^T \left\| \int_0^t k(t-s)\varphi(s) ds \right\|^2 dt &\leq \int_0^\infty \left\| \int_0^t k(t-s)\varphi_T(s) ds \right\|^2 dt \\ &= \frac{1}{2\pi} \int_{-\infty}^\infty |\hat{k}(\omega)|^2 \|\hat{\varphi}_T(\omega)\|^2 d\omega. \end{aligned}$$

Now

$$\begin{aligned} |\hat{k}(\omega)| &= \left| \int_0^\infty e^{-i\omega t} k(t) dt \right| \leq \int_0^\infty |k(t)| dt, \\ |\omega \hat{k}(\omega)| &= \left| \int_0^\infty (e^{-i\omega t} - 1)k'(t) dt \right| \leq 2 \int_0^\infty |k'(r)| dt. \end{aligned}$$

Square these two inequalities, and add them to get

$$|\hat{k}(\omega)|^2 \leq \frac{C_k}{1 + \omega^2}.$$

This combined with (4.3) yields

$$\begin{aligned} \int_0^T \left\| \int_0^t k(t-s)\varphi(s) ds \right\|^2 dt &\leq \frac{C_k}{2\pi} \int_{-\infty}^{\infty} \frac{\|\hat{\varphi}_T(\omega)\|^2}{1+\omega^2} d\omega \\ &= \frac{1}{2} C_k \int_{-\infty}^{\infty} \left\langle \varphi_T(t), \int_{-\infty}^{\infty} e^{-|t-s|} \varphi_T(s) \right\rangle dt \\ &= C_k Q(\varphi, T, e). \end{aligned}$$

LEMMA 4.3. *Let  $f, f' \in L^2(\mathbb{R}^+; H)$ . Then, for each  $\varphi \in L^1_{\text{loc}}(\mathbb{R}^+; H)$  and  $T > 0$ ,*

$$\left| \int_0^T \langle \varphi(t), f(t) \rangle dt \right|^2 \leq C_f Q(\varphi, T, e),$$

where  $C_f = 2 \int_0^{\infty} (\|f(t)\|^2 + \|f'(t)\|^2) dt$ , and  $e(t) = e^{-t}$ ,  $t \in \mathbb{R}^+$ .

This is essentially the same lemma as [12, Lemma 4.1].

*Outline of proof.* Define  $\varphi_T$  as in (4.2), and let  $f_1$  be the even extension to  $\mathbb{R}$  of  $f$ . By Parseval’s identity and the Schwarz inequality,

$$\begin{aligned} \left| \int_0^T \langle \varphi(t), f(t) \rangle dt \right| &= \left| \int_{-\infty}^{\infty} \langle \varphi_T(t), f_1(t) \rangle dt \right| \\ (4.4) \quad &= \frac{1}{2\pi} \left| \int_{-\infty}^{\infty} \langle \hat{\varphi}_T(\omega), \hat{f}_1(\omega) \rangle d\omega \right| \\ &= \frac{1}{2\pi} \left| \int_{-\infty}^{\infty} \langle (1+\omega^2)^{-1/2} \hat{\varphi}_T(\omega), (1+\omega^2)^{1/2} \hat{f}_1(\omega) \rangle d\omega \right| \\ &\leq \frac{1}{2\pi} \left[ \int_{-\infty}^{\infty} \frac{\|\hat{\varphi}_T(\omega)\|^2}{1+\omega^2} d\omega \right]^{1/2} \left[ \int_{-\infty}^{\infty} (1+\omega^2) \|\hat{f}_1(\omega)\|^2 d\omega \right]^{1/2}. \end{aligned}$$

By the Plancherel identity,

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\hat{f}_1(\omega)\|^2 d\omega &= \int_{-\infty}^{\infty} \|f_1(t)\|^2 dt = 2 \int_0^{\infty} \|f(t)\|^2 dt, \\ \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\omega^2 \hat{f}_1(\omega)\|^2 d\omega &= \int_{-\infty}^{\infty} \|f'_1(t)\|^2 dt = 2 \int_0^{\infty} \|f'(t)\|^2 dt. \end{aligned}$$

This substituted into (4.4) yields

$$\left| \int_0^T \langle \varphi(t), f(t) \rangle dt \right|^2 \leq \frac{C_f}{2\pi} \int_{-\infty}^{\infty} \frac{\|\hat{\varphi}_T(\omega)\|^2}{1+\omega^2} d\omega = C_f Q(\varphi, T, e).$$

LEMMA 4.4. *Let  $a$  be positive definite with  $a(\infty) > 0$ , and let  $f' \in L^1(\mathbb{R}^+; H)$ . Then for each  $\varphi \in L^1_{\text{loc}}(\mathbb{R}^+; H)$  and  $T > 0$ ,*

$$\left| \int_0^T \langle \varphi(t), f(t) \rangle dt \right|^2 \leq \left[ \frac{C_f}{a(\infty)} \right] \sup_{0 \leq t \leq T} Q(\varphi, t, a),$$

where

$$C_f = 2 \left[ \sup_{0 \leq t < \infty} \|f(t)\| + \int_0^{\infty} \|f'(t)\| dt \right]^2.$$

*Outline of proof.* Define  $b = a - a(\infty)$ . Then  $b$  is positive definite, and

$$Q(\varphi, t, a) = Q(\varphi, t, b) + \frac{a(\infty)}{2} \left\| \int_0^t \varphi(s) ds \right\|^2;$$

hence,

$$(4.5) \quad \left\| \int_0^t \varphi(s) ds \right\|^2 \leq \frac{2}{a(\infty)} Q(\varphi, t, a), \quad (t > 0).$$

Integrate by parts to get

$$\begin{aligned} & \left| \int_0^T \langle \varphi(t), f(t) \rangle dt \right| \\ &= \left| \left\langle \int_0^T \varphi(t) dt, f(T) \right\rangle - \int_0^T \left\langle \int_0^t \varphi(s) ds, f'(t) \right\rangle dt \right| \\ &\leq \left( \|f(T)\| + \int_0^T \|f'(t)\| dt \right) \sup_{0 \leq t \leq T} \left\| \int_0^t \varphi(s) ds \right\| \\ &= \left( \frac{C_f}{2} \right)^{1/2} \sup_{0 \leq t \leq T} \left\| \int_0^t \varphi(s) ds \right\|. \end{aligned}$$

This combined with (4.5) yields the conclusion of Lemma 4.4.

**5. Proof of Theorem 2.** Our local arguments in § 3 throughout aimed at getting bounds on various  $L^\infty(L^2)$ -norms. In some of the perturbation terms it would have been more natural to use  $L^2(L^2)$ -norms instead, but these could be replaced by  $L^\infty(L^2)$ -norms due to the fact that we were working on a finite interval. Also, one could make various badly behaving terms small by choosing  $T$ , the length of the time interval, small. When one wants to get global bounds one has to use  $L^2(L^2)$ -norms, and in order to make badly behaving terms sufficiently small one has to assume that the data is small.

We want to be able to apply the local existence theorem, and therefore we restrict the range of  $u_x$  to a set where  $\sigma' > 0$ . Choose any constant  $c_0 > 0$  such that

$$0 < p_0 \leq \sigma'(\xi) \leq p_1 \quad (\xi \in [-c_0, c_0])$$

for some constants  $p_0$  and  $p_1$ . If necessary, redefine  $\sigma$  outside of the interval  $[-c_0, c_0]$  so that  $(\sigma_{loc})$  holds. Then Theorem 1 applies, but of course we have to show that for all time,  $|u_x| \leq c_0$ , so that the fact that we redefined  $\sigma$  outside of the interval  $[-c_0, c_0]$  does not affect (E).

Fix  $T > 0$ , and assume for the moment that the solution of (E) exists on  $[0, T]$ . By Theorem 1, this assumption is justified as soon as we can show that the  $L^\infty([0, T]; L^2)$ -norms of the derivatives listed in (2.3) are finite.

We shall use Dafermos' and Nohel's terminology and say that a quantity is controllably small if it can be made arbitrarily small, independently of  $T$ , by making the appropriate  $L^p$ -norms of the functions listed in (f),  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$  and  $(u_0)$  sufficiently small. We use the letters  $C$  and  $\gamma$  to denote constants.  $C$  stands for an a priori constant, independent of  $T$ , and  $\gamma$  stands for a constant which is controllably small.

Our basic strategy is the same as in [1]; namely, we show that there exists a (small) number  $\mu$ ,  $0 < \mu \leq c_0$ , such that if

$$(5.1) \quad |u_x(t, x)|, |u_{tx}(t, x)|, |u_{xx}(t, x)| \leq \mu,$$

then the appropriate  $L^p$ -norms of the derivatives listed in (2.3) and (2.4) are controllably small. However, if the norms in (2.3) are small, then (5.1) must hold, and so the loop closes.

By the strong positive definiteness of  $a$ ,  $a(t) - \epsilon e^{-t}$  is positive definite for some  $\epsilon > 0$  and  $e(t) = e^{-t}$ ,  $t \in R^+$ . This means that

$$(5.2) \quad Q(\varphi, T, e) \leq CQ(\varphi, T, a), \quad \varphi \in L^1_{loc}(L^2), \quad T > 0,$$

where  $C = \epsilon^{-1}$ .

To simplify the notations, define

$$(5.3) \quad \begin{aligned} \varphi &= -\sigma(u_x)_x, \\ W(\xi) &= \int_0^\xi \sigma(\eta) \, d\eta, \quad \xi \in R. \end{aligned}$$

Multiply (E) by  $\varphi(t, x)$ , and integrate over  $[0, s] \times R$ ,  $0 \leq s \leq T$ , to obtain

$$(5.4) \quad \int_{-\infty}^\infty W(u_x(s, x)) \, dx + Q(\varphi, s, a) = \int_{-\infty}^\infty W(u_{0x}) \, dx + \int_0^s \int_{-\infty}^\infty \varphi f \, dx \, dt.$$

Use  $(\sigma_{loc})$ ,  $(f)$ ,  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$ , (5.2), (5.3) and Lemmas 4.3 and 4.4 to estimate the last term in (5.4) as follows (recall that  $a(\infty) = 0$  implies  $f_3 = 0$ ):

$$\begin{aligned} \left| \int_0^s \int_{-\infty}^\infty \varphi f \, dx \, dt \right| &\leq \left| \int_0^s \int_{-\infty}^\infty \sigma(u_x) f_{1x} \, dx \, dt \right| + \left| \int_0^s \int_{-\infty}^\infty \varphi [f_2 + f_3] \, dx \, dt \right| \\ &\leq \gamma \left( \|u_x\|_\infty + \sup_{0 \leq s \leq T} [Q(\varphi, s, a)]^{1/2} \right). \end{aligned}$$

Substitute this into (5.4), take the supremum of the left-hand side over  $0 \leq s \leq T$ , and use  $(\sigma_{loc})$  and (5.3) to conclude that  $\|u_x\|_\infty$  and  $\sup_{0 \leq s \leq T} Q(\varphi, s, a)$  are controllably small, i.e.,

$$(5.5) \quad \|u_x\|_\infty^2 + \sup_{0 \leq s \leq T} Q(\varphi, s, a) \leq \gamma^2.$$

Combining this with (E), (a), (f),  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$ , (5.3) and Lemma 4.1 we further conclude that  $\|u_t\|_\infty$  is controllably small, i.e.,

$$(5.6) \quad \|u_t\|_\infty \leq \gamma.$$

At this point one could make the interesting observation that so far we have used the smallness assumption (5.1) only when replacing  $(\sigma)$  by  $(\sigma_{loc})$ , and so (5.5), (5.6) hold also for large data, provided  $(\sigma_{loc})$  is satisfied. In this case  $\gamma$  represents the same thing as  $C$ , i.e., a possibly large constant independent of  $T$ .

Next we turn to the estimates for the second order derivatives  $u_{tx}$  and  $u_{xx}$ . Multiply (E) by  $\varphi_x(\partial/\partial x)$ , and integrate over  $[0, s] \times R$ ,  $0 \leq s \leq T$ , to obtain

$$\begin{aligned}
 & \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_x(s, x)) u_{xx}^2(s, x) dx + Q(\varphi_s, s, a) \\
 (5.7) \quad & = \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_{0x}) u_{0xx}^2 dx + \frac{1}{2} \int_0^s \int_{-\infty}^{\infty} \sigma''(u_x) u_{tx} u_{xx}^2 dx dt \\
 & + \int_0^s \int_{-\infty}^{\infty} \varphi_x f_x dx dt.
 \end{aligned}$$

Compared to (5.4) we have one new bad term on the right-hand side. Thanks to  $(\sigma_{loc})$  and (5.1) we can estimate

$$\left| \int_0^s \int_{-\infty}^{\infty} \sigma''(u_x) u_{tx} u_{xx}^2 dx dt \right| \leq C\mu \| \|u_{xx}\| \|_2^2.$$

The remainder of (5.7) is treated analogously to (5.4), and one gets

$$\begin{aligned}
 & \| \|u_{xx}\| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi, s, a) \\
 & \leq \gamma \left( \gamma + \| \|u_{xx}\| \|_2 + \left[ \sup_{0 \leq s \leq T} Q(\varphi_s, s, a) \right]^{1/2} \right) + C\mu \| \|u_{xx}\| \|_2^2.
 \end{aligned}$$

Use (3.17) with  $\alpha = \gamma$ ,  $\lambda = 1$  on the right-hand side to simplify this into

$$(5.8) \quad \| \|u_{xx}\| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi_s, s, a) \leq \gamma^2 + \gamma \| \|u_{xx}\| \|_2 + C\mu \| \|u_{xx}\| \|_2^2.$$

This time (5.8) does not in itself imply boundedness of the left-hand side, because we have yet no estimate on  $\| \|u_{xx}\| \|_2$ . Define

$$\begin{aligned}
 (5.9) \quad & e(t) = e^{-t}, \quad t \in R^+, \\
 & w_1 = (a - (a(0) + 1)e) * \varphi, \\
 & w_2 = (a' + (a(0) + 1)e) * \varphi, \\
 & w_3 = (a(0) + 1)e * \varphi,
 \end{aligned}$$

and subtract  $w_3$  from both sides of (E) to get

$$(5.10) \quad u_t + w_1 = f - w_3.$$

Multiply (5.10) by  $u_{tx}(\partial/\partial x)$ , and integrate over  $[0, T] \times R$  to obtain

$$\int_0^T \int_{-\infty}^{\infty} u_{tx}^2 dx dt + \int_0^T \int_{-\infty}^{\infty} u_{tx} w_{1x} dx dt = \int_0^T \int_{-\infty}^{\infty} u_{tx} (f_x - w_{3x}) dx dt.$$

After a number of integrations by parts this becomes

$$\begin{aligned}
 & \int_0^T \int_{-\infty}^{\infty} u_{tx}^2 dx dt + \int_0^T \int_{-\infty}^{\infty} \sigma'(u_x) u_{xx}^2 dx dt \\
 &= \int_0^T \int_{-\infty}^{\infty} u_{tx} [f_{1x} - w_{3x}] dx dt \\
 (5.11) \quad &+ \int_0^T \int_{-\infty}^{\infty} u_{xx} [f_{2t} + f_{3t} - w_2] dx dt \\
 &- \int_{-\infty}^{\infty} u_{xx}(T, x) [f_2(T, x) + f_3(T, x) - w_1(T, x)] dx \\
 &+ \int_{-\infty}^{\infty} u_{0xx}(x) [f_2(0, x) + f_3(0, x)] dx.
 \end{aligned}$$

Here one can use (a),  $(\sigma_{10c})$ ,  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$ , (5.2), (5.5), (5.9) and Lemmas 4.1, 4.2 to obtain

$$\begin{aligned}
 & \| \| u_{tx} \| \|_2^2 + p_0 \| \| u_{xx} \| \|_2^2 \\
 & \leq \| \| u_{tx} \| \|_2 (\| \| f_{1x} \| \|_2 + \| \| w_{3x} \| \|_2) + \| \| u_{xx} \| \|_2 (\| \| f_{2t} \| \|_2 + \| \| w_2 \| \|_2) \\
 & \quad + \| \| u_{xx} \| \|_{\infty} (\| \| f_2 \| \|_{\infty} + \| \| f_3 \| \|_{\infty} + \| \| f_{3t} \| \|_1 + \| \| w_1 \| \|_{\infty}) \\
 & \quad + \| \| u_{0xx} \| \| (\| \| f_2 \| \|_{\infty} + \| \| f_3 \| \|_{\infty}) \\
 & \leq \gamma (\gamma + \| \| u_{tx} \| \|_2 + \| \| u_{xx} \| \|_2 + \| \| u_{xx} \| \|_{\infty}) + C \| \| u_{tx} \| \|_2 [Q(\varphi_x, T, a)]^{1/2}.
 \end{aligned}$$

Use (3.17) to rewrite this as follows:

$$(5.12) \quad \| \| u_{tx} \| \|_2^2 + \| \| u_{xx} \| \|_2^2 \leq \gamma^2 + \gamma \| \| u_{xx} \| \|_{\infty} + CQ(\varphi_x, T, a).$$

Divide (5.12) by a sufficiently large constant, and add the result to (5.8) to get

$$\| \| u_{tx} \| \|_2^2 + \| \| u_{xx} \| \|_2^2 + \| \| u_{xx} \| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi_x, s, a) \leq \gamma^2 + \gamma (\| \| u_{xx} \| \|_2 + \| \| u_{xx} \| \|_{\infty}) + C\mu \| \| u_{xx} \| \|_2^2.$$

Thus, if  $\mu$  is sufficiently small, then

$$(5.13) \quad \| \| u_{tx} \| \|_2^2 + \| \| u_{xx} \| \|_2^2 + \| \| u_{xx} \| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi_x, s, a) \leq \gamma^2.$$

Again, by combining this with (E), (a), (f),  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$ , (5.3) and Lemma 4.1 we further obtain

$$(5.14) \quad \| \| u_{tx} \| \|_{\infty} \leq \gamma.$$

The estimates for  $u_{txx}$  and  $u_{xxx}$  are obtained in exactly the same manner as the estimates for  $u_{tx}$  and  $u_{xx}$ . Suppose for the moment that the data is smooth enough so that Lemma 3.1 can be applied. Multiply (E) by  $\varphi_{xx}(\partial^2/\partial x^2)$ , and integrate over  $[0, x] \times R$ ,

$0 \leq s \leq T$ , to obtain

$$\begin{aligned}
 & \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_x(s, x)) u_{xxx}^2(s, x) dx + Q(\varphi_{xx}, s, a) \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} \sigma'(u_{0x}) u_{0xxx}^2 dx + \frac{1}{2} \int_0^s \int_{-\infty}^{\infty} \sigma''(u_x) u_{tx} u_{xxx}^2 dx dt \\
 (5.15) \quad &+ 2 \int_0^s \int_{-\infty}^{\infty} \sigma''(u_x) u_{xx} u_{txx} u_{xxx} dx dt + \int_0^s \int_{-\infty}^{\infty} \sigma'''(u_x) u_{xx}^3 u_{txx} dx dt \\
 &+ \int_0^s \int_{-\infty}^{\infty} \varphi_{xx} f_{xx} dx dt.
 \end{aligned}$$

We treat (5.15) in the same way as (5.4), (5.7) to get

$$\begin{aligned}
 & \| \| u_{xxx} \| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi_{xx}, s, a) \\
 & \leq \gamma(\gamma + \| \| u_{xxx} \| \|_2 + \sup_{0 \leq s \leq T} [Q(\varphi_{xx}, s, a)]^{1/2}) \\
 & + C\mu(\| \| u_{xxx} \| \|_2^2 + \| \| u_{txx} \| \|_2 \| \| u_{xxx} \| \|_2 + \mu \| \| u_{xx} \| \|_2 \| \| u_{txx} \| \|_2).
 \end{aligned}$$

This combined with (3.17), (5.13) yields

$$(5.16) \quad \| \| u_{xxx} \| \|_{\infty}^2 + \sup_{0 \leq s \leq T} Q(\varphi_{xx}, s, a) \leq \gamma^2 + \gamma \| \| u_{xxx} \| \|_2 + C\mu(\| \| u_{txx} \| \|_2^2 + \| \| u_{xxx} \| \|_2^2).$$

This time we need estimates on both  $\| \| u_{txx} \| \|_2$  and  $\| \| u_{xxx} \| \|_2$  to complement (5.16), but these are obtained in the same way as before. Multiply (5.10) by  $u_{txx}(\partial^2/\partial x^2)$ , integrate over  $[0, T] \times R$ , and do the same integrations by parts which were used to arrive at (5.11). This yields

$$\begin{aligned}
 & \int_0^T \int_{-\infty}^{\infty} u_{txx}^2 dx dt + \int_0^T \int_{-\infty}^{\infty} \sigma'(u_x) u_{xxx}^2 dx dt \\
 &= \int_0^T \int_{-\infty}^{\infty} u_{txx} [f_{1xx} - w_{3xx}] dx dt \\
 &+ \int_0^T \int_{-\infty}^{\infty} u_{xxx} [f_{2tx} + f_{3tx} - w_{2x} - \sigma''(u_x) u_{xx}^2] dx dt \\
 &- \int_{-\infty}^{\infty} u_{xxx}(T, x) [f_{2x}(T, x) + f_{3x}(T, x) - w_{1x}(T, x)] dx dt \\
 &+ \int_{-\infty}^{\infty} u_{0xxx}(x) [f_{2x}(0, x) + f_{3x}(0, x)] dx dt.
 \end{aligned}$$

Again, use this together with (a),  $(\sigma_{loc})$ ,  $(f_1)$ ,  $(f_2)$ ,  $(f_3)$ , (5.1), (5.2), (5.9), (5.13) and

Lemmas 4.1, 4.2 to obtain

$$\begin{aligned}
 & \|u_{txx}\|_2^2 + p_0 \|u_{xxx}\|_2^2 \\
 & \leq \|u_{txx}\|_2 (\|f_{1xx}\|_2 + \|w_{3xx}\|_2) \\
 & \quad + \|u_{xxx}\|_2 (\|f_{2tx}\|_2 + \|w_{2x}\|_2 + C\mu \|u_{xx}\|_2) \\
 & \quad + \|u_{xxx}\|_\infty (\|f_{2x}\|_\infty + \|f_{3x}\|_\infty + \|f_{3tx}\|_1 + \|w_{1x}\|_\infty) \\
 & \quad + \|u_{0xxx}\| (\|f_{2x}\|_\infty + \|f_{3x}\|_\infty) \\
 & \leq \gamma (\gamma + \|u_{txx}\|_2 + \|u_{xxx}\|_2 + \|u_{xxx}\|_\infty) \\
 & \quad + C \|u_{txx}\|_2 [Q(\varphi_{xx}, T, a)]^{1/2}.
 \end{aligned}$$

Thus, by further using (3.17) one gets

$$(5.17) \quad \|u_{txx}\|_2^2 + \|u_{xxx}\|_2^2 \leq \gamma^2 + \gamma \|u_{xxx}\|_\infty + CQ(\varphi_{xx}, T, a).$$

Divide (5.17) by a sufficiently large constant, add the result to (5.16), and choose  $\mu$  small enough to obtain

$$(5.18) \quad \|u_{txx}\|_2^2 + \|u_{xxx}\|_2^2 + \|u_{xxx}\|_\infty^2 + \sup_{0 \leq s \leq T} Q(\varphi_{xx}, s, a) \leq \gamma^2.$$

As before, this implies

$$(5.19) \quad \|u_{txx}\|_\infty \leq \gamma.$$

We have derived (5.18), (5.19) under the additional assumption that the data was smooth. As a matter of fact,  $Q(\varphi_{xx}, s, a)$  in (5.18) is not even well defined unless  $\varphi_{xx} \in L^1_{loc}(L^2)$ , i.e.,  $u_{xxx} \in L^1_{loc}(L^2)$ . Of course, (5.18) implies

$$(5.20) \quad \|u_{txx}\|_2^2 + \|u_{xxx}\|_2^2 + \|u_{xxx}\|_\infty^2 \leq \gamma^2.$$

Now (5.19), (5.20) make sense even for the less smooth solution which Theorem 1 establishes. As the solution depends continuously on the data and (5.19), (5.20) hold for smooth solutions, also the solution established by Theorem 1 must satisfy (5.19), (5.20).

By now the proof of Theorem 2 is almost complete. Combining (5.5), (5.6), (5.13), (5.14), (5.19) and (5.20) we observe that the  $L^p$ -norms of the derivatives listed in (2.3), (2.4) are controllably small. But this together with (3.13) implies that one can make (5.1) hold for arbitrarily small  $\mu$  by choosing the data small enough. Hence by Theorem 1, we get a global solution satisfying (2.3), (2.4) for small data. That (2.5) holds follows from (E) combined with (a), ( $\sigma$ ), (2.3), (2.4), and (5.1). By (2.4), (2.5) both  $u_{xx}$  and  $u_t - f_t$  belong to  $L^2(L^2)$  together with their derivatives with respect to  $t$ , and so (2.6) holds. The final claim (2.7) follows from (2.3), (2.5), (2.6) and (3.13).

*Proof of Corollary 3.* Let the assumption of Corollary 3 hold. Then so do (2.3)–(2.7). By (2.5) and the fact that  $f_x$  is uniformly continuous with values in  $L^2$ , so is  $u_{tx}$ . Combined with (2.4) this yields (2.8), which in turn combined with (2.3) and (3.13) yields (2.9):

$$\begin{aligned}
 |u_t(t, x)|^2 & \leq 2 \|u_t\|_\infty \|u_{tx}\|(t) \rightarrow 0, & t \rightarrow \infty, \\
 |u_{tx}(t, x)|^2 & \leq 2 \|u_{tx}\|(t) \|u_{txx}\|_\infty \rightarrow 0, & t \rightarrow \infty.
 \end{aligned}$$

*Proof of Corollary 4.* Use (E), ( $\sigma$ ), (2.4) and (5.1) together with the fact that convolution with an  $L^1$ -function maps  $L^2$  into itself to get (2.10). Combined with (2.5) this implies that  $u_t - f \rightarrow 0$  in  $L^2$  as  $t \rightarrow \infty$ . In this case clearly  $a(\infty) = 0$ ; hence,  $f_3 = 0$  and



so (f), (f<sub>1</sub>), (f<sub>2</sub>) imply  $f_x \in L^2(L^2)$ . Combined with (2.4) this yields  $u_{tx} - f_x \in L^2(L^2)$ , which in turn combined with (2.5) implies that  $u_{tx} - f_x \rightarrow 0$  in  $L^2$ . The remaining claims concerning the uniform convergence can again be deduced from the  $L^2$ -convergence by use of (3.13):

$$\begin{aligned} |u_t(t, x) - f(t, x)|^2 &\leq 2\|u_t - f\|_\infty \|u_{tx} - f_x\|(t) \rightarrow 0, \quad t \rightarrow \infty, \\ |u_{tx}(t, x) - f_x(t, x)|^2 &\leq 2\|u_{tx} - f_x\|(t)\|u_{txx} - f_{xx}\|_\infty \rightarrow 0, \quad t \rightarrow \infty. \end{aligned}$$

**6. Modifications of the main results.** When  $a$  satisfies (a) and in addition either  $a \in L^1(\mathbb{R}^+)$  or  $a(\infty) > 0$ , then one can replace some of the derivatives with respect to  $x$  in (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) by derivatives with respect to  $t$ . More specifically, one can show that (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) can be replaced by

$$\begin{aligned} (f_1)^\sim & f_1 \in L^\infty(L^2), \quad f_{1x} \in L^1 \cap L^\infty(L^2), \quad f_{1tx} \in L^2 \cap L^\infty(L^2), \quad f_{1txx} \in L^2(L^2); \\ (f_2)^\sim & f_2, f_{2t}, f_{2xt}, f_{2tt}, f_{2tx}, f_{2ttx} \in L^2(L^2); \\ (f_3)^\sim & f_3 \in L^\infty(L^2), \quad f_{3t} \in L^1 \cap L^\infty(L^2), \quad f_{3tx} \in L^2 \cap L^\infty(L^2), \quad f_{3tb} f_{3ttx} \in L^1(L^2). \end{aligned}$$

Also some mixed versions of (f<sub>1</sub>)<sup>~</sup>, (f<sub>2</sub>)<sup>~</sup>, (f<sub>3</sub>)<sup>~</sup> could be used, so that, e.g., (2.15) is a perfectly good condition when  $a(\infty) > 0$  (even without the requirement  $f_{tt} \in L^2(L^2)$ ). We shall only give a brief outline of the fact that (f<sub>1</sub>), (f<sub>2</sub>), (f<sub>3</sub>) can be replaced by (f<sub>1</sub>)<sup>~</sup>, (f<sub>2</sub>)<sup>~</sup>, (f<sub>3</sub>)<sup>~</sup> in Theorem 2 and Corollaries 3 and 4 when either  $a(\infty) > 0$  or  $a \in L^1(\mathbb{R}^+)$ . This proof is similar to the proof of Theorem 2 given in § 5, but it is more complicated, and it requires a substantial amount of additional work. The estimates on the first order derivatives are obtained in the same way as in § 5. To get the required bounds on the second order derivatives one first multiplies (E) by  $\varphi_t(\partial/\partial t)$  and integrates over  $[0, s] \times \mathbb{R}$ ,  $0 \leq s \leq T$ , to obtain an inequality of the type

$$(6.1) \quad \|u_{tx}\|_\infty^2 + \sup_{0 \leq s \leq T} Q(\varphi_b, s, a) \leq \gamma^2 + \gamma \|u_{tx}\|_2 + C\mu \|u_{tx}\|_2^2.$$

In (5.11) one integrates by parts to replace  $w_{3x}$  by  $(a(0) + 1)e * \varphi_b$ , and gets the inequality

$$(6.2) \quad \|u_{tx}\|_2^2 + \|u_{xx}\|_2^2 \leq \gamma^2 + \|u_{tx}\|_\infty^2 + \|u_{xx}\|_\infty^2 + CQ(\varphi_b, T, a).$$

Unfortunately, this time we have not yet obtained any bound on  $\|u_{xx}\|_\infty$ , and to get this bound we have to impose conditions on  $a$  which imply that the resolvent  $r$  of  $a'$  defined in (3.1) satisfies at least

$$(r') \quad r' \in L^1(\mathbb{R}^+).$$

If  $a(\infty) > 0$ , then by the standard Paley-Wiener theorem

$$(r) \quad r \in L^1(\mathbb{R}^+),$$

and this together with (a) and (3.1) yields (r'). If  $a \in L^1(\mathbb{R}^+)$ , then it is also possible to show that (r') holds (one argues in the spirit of [10]). By using (r') one can show that

$$(6.3) \quad \|u_{xx}\|_\infty^2 \leq \gamma^2 + C \sup_{0 \leq s \leq T} Q(\varphi_b, s, a).$$

Combining (6.1), (6.2), (6.3) one gets the desired bounds on the second order derivatives.

To get the bounds on the third order derivatives one first multiplies (E) by  $\varphi_{tx}(\partial^2/\partial t \partial x)$ , and integrates over  $[0, s] \times \mathbb{R}$ ,  $0 \leq s \leq T$  to obtain

$$(6.4) \quad \|u_{txx}\|_\infty^2 + \sup_{0 \leq s \leq T} Q(\varphi_{tx}, s, a) \leq \gamma^2 + \gamma \|u_{txx}\|_2 + C\mu (\|u_{txx}\|_2^2 + \|u_{txx}\|_2^2 + \|u_{xxx}\|_2^2).$$

Next one multiplies (E) by  $u_{txx} \partial^2/\partial t \partial x$ , and does the same type of calculation which gave rise to (5.11) to obtain

$$(6.5) \quad \left\| \|u_{txx}\|_2^2 + \|u_{txx}\|_2^2 \right\| \leq \gamma^2 + \gamma \| \|u_{txx}\|_\infty + CQ(\varphi_{tx}, T, a).$$

Finally, one uses (E), (r) and (r') to get

$$(6.6) \quad \| \|u_{xxx}\|_p \leq \gamma + C \| \|u_{txx}\|_p, \quad p = 2, \infty, \quad \| \|u_{txx}\|_\infty \leq \gamma + C \sup_{0 \leq s \leq T} [Q(\varphi_{tx}, s, a)]^{1/2}.$$

Combining (6.4), (6.5), (6.6) one gets the needed bounds on the third order derivatives. The proof is completed exactly as in § 5.

Dafermos and Nohel mention several possible extensions of their analogues of Theorems 1 and 2. Similar extensions are possible here. Instead of working on an unbounded space domain one can study an initial-boundary value problem with either Neumann or Dirichlet boundary conditions. This basically amounts to replacing the space  $L^2(R)$  in §§ 2–6 by an  $L^2$ -space over a finite interval. However, it should be pointed out that the fact that we perform one more differentiation and integration by parts with respect to  $x$  than in [1] causes some additional difficulties. In particular, in the Neumann problem, one has to strengthen the assumption on  $f$  to make some boundary terms vanish. Also, the fact that we worked in only one space dimension was not important. The same method works for equations with several space variables. For details, see [1]. Observe that by using our estimates in §§ 3 and 5 rather than those in [1] one does not need more than one derivative with respect to  $t$ , and so there is no need to strengthen the assumption on  $a$  when one works in higher dimensions.

#### REFERENCES

- [1] C. M. DAFERMOS AND J. A. NOHEL, *Energy methods for nonlinear hyperbolic Volterra integrodifferential equations*, Comm. Partial Differential Equations, 4 (1979), pp. 219–278.
- [2] P. D. LAX, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Mathematical Phys., 5 (1964), pp. 611–613.
- [3] R. C. MACCAMY, *An integro-differential equation with application in heat flow*, Quart. Appl. Math., 35 (1977), pp. 1–19.
- [4] R. C. MACCAMY, *A model for one-dimensional, nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21–33.
- [5] R. C. MACCAMY AND V. MIZEL, *Existence and nonexistence in the large of solutions of quasilinear wave equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 299–320.
- [6] A. MATSUMURA, *Global existence and asymptotics of the solutions of the second-order quasilinear hyperbolic equations with first-order dissipation*, to appear.
- [7] T. NISHIDA, *Global smooth solutions for the second-order quasilinear wave equation with the first-order dissipation*, unpublished.
- [8] J. A. NOHEL, *A forced quasilinear wave equation with dissipation*, Proceedings of EQUADIFF 4, Lecture notes in Mathematics 703, Springer, Berlin, 1979, pp. 318–327.
- [9] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Advances in Math., 22 (1976), pp. 278–304.
- [10] 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.
- [11] O. J. STAFFANS, *Positive definite measures with applications to a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 219–237.
- [12] O. J. STAFFANS, *Boundedness and asymptotic behavior of solutions of a nonlinear Volterra equation*, Michigan Math. J., 24 (1977), pp. 77–95.
- [13] O. J. STAFFANS, *A nonlinear Volterra equation with rapidly decaying solutions*, Trans. Amer. Math. Soc., to appear.
- [14] O. J. STAFFANS, *A Volterra equation with square integrable solutions*, Proc. Amer. Math. Soc., to appear.

## SPLITTING THE SPECTRUM OF A RIEMANNIAN MANIFOLD\*

DAVID D. BLEECKER† AND LESLIE C. WILSON†

**Abstract.** The eigenvalues of a Riemannian manifold have been calculated mostly only for spaces having a high degree of symmetry. In these cases the eigenvalues generally have large multiplicities. However, for most metrics the eigenvalues are all simple. We give a proof of this using perturbation theory. In particular, we characterize those functions  $f$  for which the perturbation  $\exp(\epsilon f)g$  will completely split a given eigenvalue of the Laplace operator of the metric  $g$ . We give examples in which such an  $f$  is explicitly calculated.

**1. Introduction.** In this paper we show that given a closed, connected  $C^\infty$  manifold  $M$  of dimension  $n \geq 2$  and an arbitrary Riemannian metric  $g$  on  $M$ , there is a dense (in fact residual) set of  $f \in C^\infty(M)$  for which the eigenvalues of the Laplacian of the conformally related metric  $\exp(f)g$  all have multiplicity one. The proof relies on the perturbation theory of linear operators pioneered by Franz Rellich and subsequently extended by many others [6], [7]. A key observation which we make is that it is "formally" possible to split an eigenvalue of  $\Delta$  by a perturbation of the metric of the form  $\exp(\epsilon f)g$ , where  $f$  is a suitable eigenfunction of  $\Delta$ . The perturbation theory makes the formal procedure valid and allows us to show that the subsets  $F_i$  (of functions  $f \in C^\infty(M)$  such that the first  $i$  eigenvalues of the Laplacian of  $\exp(f)g$  are of multiplicity 1) are open and dense. Then, as  $F_\infty = \bigcap_i F_i$ ,  $F_\infty$  is residual.

Since finishing our paper we have learned of the important work of Uhlenbeck [8]. Using infinite dimensional transversality theory, she shows that among certain classes of elliptic operators on a compact manifold, the following properties are generic: eigenvalues have multiplicity one, zero is not a critical value of the eigenfunction restricted to the interior of the domain of the operator, and the eigenfunctions are Morse functions on the interior of the domain of the operator. In her paper, generic means of second category, which is weaker than residual. Aside from this, her results include ours as a special case. We offer our paper anyway since our methods are quite different and are more constructive (see Remarks 5.3.2 and 5.3.3). The reader should also consult the article of Albert [1] for similar results about elliptic operators perturbed by adding a smooth function to the operator.

**2. Notation and statement of the main theorem.** Let  $M$  be a closed, connected  $C^\infty$   $n$ -dimensional manifold,  $n \geq 2$ , and let  $C^\infty(M)$  be the set of all real valued  $C^\infty$  functions on  $M$ . We give  $C^\infty(M)$  the topology of uniform convergence of derivatives of all orders. A subset of  $C^\infty(M)$  is called *residual* if it is the countable intersection of open and dense subsets. Since  $C^\infty(M)$  is a Frechet space (a vector space having a complete, translation-invariant metric; see [5]), residual sets are dense.

Let  $x_1, \dots, x_n$  be a coordinate system defined on some open set  $U \subset M$ . Let  $\partial_i$ , be the Laplace operator on  $C^\infty(M)$  relative to the metric  $\exp(f)g$ . In § 5 we will prove:

**THEOREM 2.1.** *The set of all  $f \in C^\infty(M)$  for which every eigenvalue of  $f\Delta$  has multiplicity one is residual.*

Let  $x_1, \dots, x_n$  be a coordinate system defined on some open set  $U \subset M$ . let  $\partial_i$ ,  $1 \leq i \leq n$ , denote the  $i$ th coordinate vector field on  $U$  or, by standard abuse, the partial derivative with respect to the coordinate  $x_i$ . For the fixed metric  $g$ , let  $\bar{\omega}$  be the volume element associated with  $g$ , and define  $g_{ij} : U \rightarrow \mathbb{R}$  by  $g_{ij} = g(\partial_i, \partial_j)$ . Let  $\omega : U \rightarrow \mathbb{R}$  be given

\* Received by the editors November 29, 1979.

† University of Hawaii at Manoa, Department of Mathematics, 2565 The Mall, Honolulu, Hawaii, 96822. This work was supported in part by the National Science Foundation under Grant MCS 76-08216.

by  $\omega = (\det g_{ij})^{1/2}$ . The volume element on  $U$  is then  $\bar{\omega} = \omega dx_1 \wedge \cdots \wedge dx_n$ . In terms of these local coordinates, the Laplacian of the metric  $g$ , denoted  $\Delta: C^\infty(M) \rightarrow C^\infty(M)$  is given by  $\Delta(h) = -(1/\omega) \partial_i(\omega g^{ij} \partial_j h)$ , where  $g^{ij} g_{jk} = \delta_j^i$  defines  $g^{ij}$ , and we have used the Einstein summation convention. For a given function  $f \in C^\infty(M)$  and real number  $\varepsilon$ , we use  $g_{ij}(\varepsilon)$ ,  $\omega(\varepsilon)$ ,  $\Delta(\varepsilon)$ , etc., to denote the corresponding objects derived from the metric  $\exp(\varepsilon f)g$ . Note that  $g_{ij}(\varepsilon) = \exp(\varepsilon f)g_{ij}$  and  $\omega(\varepsilon) = \exp(\frac{1}{2}n\varepsilon f)\omega$ . We define an operator  ${}_fA(\varepsilon): C^\infty(M) \rightarrow C^\infty(M)$  by  ${}_fA(\varepsilon)(h) = \exp(\frac{1}{4}n\varepsilon f) \Delta(\varepsilon)[\exp(-\frac{1}{4}n\varepsilon f)h]$ . The purpose of using  ${}_fA(\varepsilon)$ , instead of  $\Delta(\varepsilon)$  by itself, is that  $\Delta(\varepsilon)$  is not necessarily self-adjoint with respect to the  $L^2$  inner product  $\langle h, k \rangle = \int_M h k \bar{\omega}$  associated with the original metric.

**PROPOSITION 2.2.**  ${}_fA(\varepsilon): C^\infty(M) \rightarrow C^\infty(M)$  is self-adjoint relative to  $\langle \cdot, \cdot \rangle$ ; that is  $\langle {}_fA(\varepsilon)h, k \rangle = \langle h, {}_fA(\varepsilon)k \rangle$  for  $h, k \in C^\infty(M)$ .

*Proof.*  $\langle {}_fA(\varepsilon)h, k \rangle = \int_M {}_fA(\varepsilon)[h]k\bar{\omega} = \int_M \exp(\frac{1}{4}n\varepsilon f)\Delta(\varepsilon) \cdot [\exp(-\frac{1}{4}n\varepsilon f)h]k\bar{\omega} = \int_M \Delta(\varepsilon)[\exp(-\frac{1}{4}n\varepsilon f)h] \exp(-\frac{1}{4}n\varepsilon f)k\bar{\omega}(\varepsilon)$  which is symmetric in  $h$  and  $k$  because of the self-adjointness of  $\Delta(\varepsilon)$  relative to the  $L^2$  inner product arising from  $\bar{\omega}(\varepsilon)$ .  $\square$

**3. Some formulas.** While the formulas here are not becoming, they will be of importance in establishing certain estimates needed for the application of perturbation theory.

**PROPOSITION 3.1.** For  $h \in C^\infty(M)$  we have

$${}_fA(\varepsilon)(h) = \exp(-\varepsilon f)[\Delta h + \varepsilon(\nabla f \cdot \nabla h - \frac{1}{4}n(\Delta f)h) + \varepsilon^2(\frac{1}{4}n - 1)\frac{1}{4}n|\nabla f|^2 h].$$

*Proof.* The proof is a straight-forward, but most unpleasant, computation. We provide checkpoints for the reader who wishes to verify it. In local coordinates, we have  ${}_fA(\varepsilon)(h) = -\exp(\frac{1}{4}n\varepsilon f)\omega(\varepsilon)^{-1} \partial_i\{\omega(\varepsilon)g^{ij} \partial_j[\exp(-\frac{1}{4}n\varepsilon f)h]\}$ . Applying the product rule for each of the partial derivatives (performing the inner partial first) we obtain four terms:

$$\begin{aligned} &\exp(-\varepsilon f)(\frac{1}{4}n - 1)\frac{1}{4}n\varepsilon^2|\nabla f|^2 h, && \exp(-\varepsilon f)(\frac{1}{4}n\varepsilon)[\nabla f \cdot \nabla h - (\Delta f)h], \\ &-\exp(-\varepsilon f)(\frac{1}{4}n - 1)\varepsilon \nabla f \cdot \nabla h, && \text{and } \exp(-\varepsilon f) \Delta h. \end{aligned}$$

Adding these, we get the formula for  ${}_fA(\varepsilon)(h)$ .  $\square$

**COROLLARY 3.2.**  ${}_fA(\varepsilon)(h) = {}_fA^{(0)}(h) + {}_fA^{(1)}(h)\varepsilon + {}_fA^{(2)}(h)\varepsilon^2 + \cdots$  where  ${}_fA^{(0)}(h) = \Delta h$ ,  ${}_fA^{(1)}(h) = \nabla f \cdot f \Delta h - f \Delta h - \frac{1}{4}nh \Delta f$ , and  ${}_fA^{(m)}(h) = [(-1)^m/(m - 2)!]f^{m-2}[f^2 \Delta h/m(m - 1) + \frac{1}{4}n(\frac{1}{4}n - 1)|\nabla f|^2 h - f(\nabla f \cdot \nabla h - \frac{1}{4}nh \Delta f)/(m - 1)]$ .

*Proof.* Use the formula for  ${}_fA(\varepsilon)(h)$ , expand  $\exp(-\varepsilon f)$  in a power series in  $\varepsilon$ , and gather like powers.  $\square$

**4. Perturbation theory for Laplacians.** Most of the terminology and theory we use comes from Rellich [7].

A linear operator defined on a dense subspace  $U$  of a complex Hilbert space  $H$  is said to be *Hermitian* if  $\langle v, Au \rangle = \langle Av, u \rangle$  for all  $u, v \in U$ .  $A$  is called *essentially self-adjoint* if  $A$  is Hermitian and  $A - i$  and  $A + i$  map  $U$  onto dense subspaces of  $H$ .

We may consider  $\Delta$  and  ${}_fA(\varepsilon)$  as operators on complex-valued  $C^\infty$  functions on  $M$ ,  $C^\infty(M) \otimes C$ . For  $h, k \in C^\infty(M) \otimes C$ , we let  $h, k = \int_M h \bar{k} \bar{\omega}$ . Then Proposition 2.2 is still valid, as well as the formulas of § 3, when properly interpreted.

**PROPOSITION 4.1.**  $\Delta$  is essentially self-adjoint on  $L^2(M, \bar{\omega})$ , the completion of  $C^\infty(M) \otimes C$  relative to  $\langle \cdot, \cdot \rangle$ .

*Proof.* Since the eigenfunctions of  $\Delta$  span a dense subset of  $L^2(M, \bar{\omega})$ , we need only show that for any eigenfunction  $u$ , we can find  $\phi, \psi \in C^\infty(M)$  satisfying  $\Delta \phi + i\phi = u$  and  $\Delta \psi - i\psi = u$ . Simply let  $\phi = (\lambda + i)^{-1}u$  and  $\psi = (\lambda - i)^{-1}u$ , where  $\lambda$  is the eigenvalue of  $u$ .  $\square$

PROPOSITION 4.2. *The various  $fA^{(n)}$  of § 3 are Hermitian.*

*Proof.* Since  $fA(\varepsilon)$  is Hermitian for all  $\varepsilon$ , we have  $0 = (d^n/d\varepsilon^n)(\langle fA(\varepsilon)h, k \rangle - \langle h, fA(\varepsilon)k \rangle)|_{\varepsilon=0} = n!(\langle fA^{(n)}h, k \rangle - \langle h, fA^{(n)}k \rangle)$ .  $\square$

PROPOSITION 4.3. *There are constants  $p, a, b > 0$  such that for all  $h \in C^\infty(M)$ , we have  $\|fA^{(m)}h\| \leq p^{m-1}(a\|h\| + b\|fA^{(0)}h\|)$ ,  $m = 1, 2, 3, \dots$ .*

*Proof.* We use Corollary 3.2. Note that  $\int_M |\nabla h|^2 \bar{\omega} = \int_M h \Delta h \bar{\omega} \leq (\int_M h^2 \bar{\omega})^{1/2} (\int_M |\Delta h|^2 \bar{\omega})^{1/2}$ . In other words,  $\|\nabla h\|^2 \leq \|h\| \|\Delta h\| \leq \frac{1}{2}(\|h\| + \|\Delta h\|)^2$  so  $\|\nabla h\| \leq 2^{-1/2}(\|h\| + \|\Delta h\|)$ . Now  $|f|$ ,  $|\nabla f|$ , and  $|\Delta f|$  have finite maximums on  $M$ . This, together with our bound on  $\|\nabla h\|$ , yields  $\|fA^{(m)}h\| \leq p^{m-2}/(m-2)! \cdot (\bar{a}\|h\| + \bar{b}\|fA^{(0)}h\|)$ , where  $p = \max |f|$  and  $m \geq 2$ . (The case  $m = 1$  is handled separately.) Finally, set  $a = \bar{a}/p$  and  $b = \bar{b}/p$ .  $\square$

THEOREM 4.4. *Suppose  $\lambda$  is an eigenvalue of finite multiplicity  $k \geq 1$  of  $\Delta$  on  $C^\infty(M)$ . Let  $d_1, d_2 > 0$  be such that  $(\lambda - d_1, \lambda + d_2)$  contains no other eigenvalue of  $\Delta$ . Then, for any  $f \in C^\infty(M)$ , there exist power series  $\lambda_1(\varepsilon), \dots, \lambda_k(\varepsilon)$  with values in  $\mathbb{R}$ , and power series  $\phi_1(\varepsilon), \dots, \phi_k(\varepsilon)$  with values in  $C^\infty(M)$  such that the following hold.*

- (1) *The  $\lambda_i(\varepsilon)$  are convergent in a neighborhood of  $0 \in \mathbb{R}$ .*
- (2) *The  $\phi_i(\varepsilon)$  converge for  $\varepsilon$  in a neighborhood of  $0 \in \mathbb{R}$  in the sense that the partial sums converge in  $L^2(M, \bar{\omega})$  to a function in  $C^\infty(M)$ .*
- (3)  *$fA(\varepsilon)\phi_i(\varepsilon) = \lambda_i(\varepsilon)\phi_i(\varepsilon)$ ,  $i = 1, \dots, k$ .*
- (4)  *$\lambda_i(0) = \lambda$ ,  $i = 1, \dots, k$ .*
- (5)  *$\langle \phi_i(\varepsilon), \phi_j(\varepsilon) \rangle = \delta_{ij}$ ,  $i, j = 1, \dots, k$ .*
- (6) *For every  $d'_1$  and  $d'_2$  with  $0 < d'_1 < d_1$  and  $0 < d'_2 < d_2$ , there is an  $\alpha > 0$  such that  $|\varepsilon| < \alpha$  implies that the only eigenvalues of  $fA(\varepsilon)$  in  $(\lambda - d'_1, \lambda + d'_2)$  are  $\lambda_1(\varepsilon), \dots, \lambda_k(\varepsilon)$ .*

*Proof.* Propositions 4.1–4.3 imply that  $fA(\varepsilon)$  satisfies criterion 3 [7, p. 78]. Thus, by definition 2 [7, p. 78], the hypermaximal extension of  $fA(\varepsilon)$  is regular in the sense of definition 1 [7, p. 71]. Thus, Theorem 3 [7, p. 74] applies with  $fA(\varepsilon)$  replacing his  $A(\varepsilon)$ .  $\square$

The eigenfunctions  $\phi_i(\varepsilon)$  and eigenvalues  $\lambda_i(\varepsilon)$  of  $fA(\varepsilon)$  are easily related to those of  $\Delta(\varepsilon)$ .

PROPOSITION 4.5.  *$u(\varepsilon)$  is an eigenfunction of  $\Delta(\varepsilon)$  with eigenvalue  $\lambda(\varepsilon)$  iff  $w(\varepsilon) \equiv \exp(\frac{1}{2}\varepsilon f)u(\varepsilon)$  is an eigenfunction of  $fA(\varepsilon)$  with eigenvalue  $\lambda(\varepsilon)$ .*

*Proof.* Both  $\Delta(\varepsilon)$  and  $fA(\varepsilon)$  are elliptic, and so their eigenfunctions are  $C^\infty$ . If  $\Delta(\varepsilon)u(\varepsilon) = \lambda(\varepsilon)u(\varepsilon)$ , then  $fA(\varepsilon)(w(\varepsilon)) = \exp(\frac{1}{2}\varepsilon f)\Delta(\varepsilon)(\exp(-\frac{1}{2}\varepsilon f)w(\varepsilon)) = \lambda(\varepsilon)w(\varepsilon)$ , and the converse is just as easy.  $\square$

*Remark.* The functions  $\lambda_i(\varepsilon)$  can be defined for all real  $\varepsilon$  by considering  $\exp(\varepsilon_0 f)g$  as the initial metric and taking the variation to be  $\varepsilon \rightarrow \exp((\varepsilon + \varepsilon_0)f)g$ . Piecing together the functions  $\lambda_i$  for various  $\varepsilon_0$  presents no problem.

PROPOSITION 4.5.  *$u(\varepsilon)$  is an eigenfunction of  $\Delta(\varepsilon)$  with eigenvalue  $\lambda(\varepsilon)$  iff  $fA^{(1)}$  be as in Corollary 3.2. Then  $\langle fA^{(1)}h, k \rangle = \int_M [(\frac{1}{2} - \frac{1}{4}\varepsilon) \Delta f - \lambda f] h k \bar{\omega}$ .*

*Proof.*  $\langle fA^{(1)}h, k \rangle = \langle \nabla f \cdot \nabla h - f \Delta h - \frac{1}{4}\varepsilon \Delta f, k \rangle = \langle \frac{1}{2}(h \Delta f + f \Delta h - \Delta(fh)) - f \lambda h - \frac{1}{4}\varepsilon \Delta f, k \rangle = \langle \frac{1}{2} h \Delta f - \frac{1}{2} f \lambda h - \frac{1}{4}\varepsilon \Delta f, k \rangle - \frac{1}{2} \langle \Delta(fh), k \rangle = \langle [(\frac{1}{2} - \frac{1}{4}\varepsilon) \Delta f - \lambda f] h, k \rangle$ .  $\square$

PROPOSITION 4.7. *Using the notation of Theorem 4.4, we have  $\lambda'_i(0) = \langle \phi_i(0), fA^{(1)}\phi_i(0) \rangle$  and  $0 = \langle \phi_i(0), fA^{(1)}\phi_j(0) \rangle$  if  $i \neq j$ .*

*Proof.*  $fA(\varepsilon)\phi_i(\varepsilon) = \lambda_i(\varepsilon)\phi_i(\varepsilon)$  implies  $\langle fA(\varepsilon)\phi_i(\varepsilon), \phi_j(\varepsilon) \rangle = \langle \lambda_i(\varepsilon)\phi_i(\varepsilon), \phi_j(\varepsilon) \rangle$ . Differentiating with respect to  $\varepsilon$  and then setting  $\varepsilon = 0$ , we have  $\langle fA^{(1)}\phi_i(0), \phi_j(0) \rangle + \langle fA(0)\phi'_i(0), \phi_j(0) \rangle + \langle fA(0)\phi_i(0), \phi'_j(0) \rangle = \langle \lambda'_i(0)\phi_i(0), \phi_j(0) \rangle + \langle \lambda\phi'_i(0), \phi_j(0) \rangle + \langle \lambda\phi_i(0), \phi'_j(0) \rangle$  or  $\langle fA^{(1)}\phi_i(0), \phi_j(0) \rangle = \lambda'_i(0)\delta_{ij}$ . One point to observe is that  $\phi'_i(0)$  is  $C^\infty$  since it satisfies the elliptic equation

$$fA^{(0)}\phi'_i(0) - \lambda_i(0)\phi'_i(0) = \lambda'_i(0)\phi_i(0) - fA^{(1)}\phi_i(0).$$

Thus,  $fA^{(0)}\phi'_i(0)$  makes sense in the above calculation.  $\square$

**COROLLARY 4.8.** *Let  $P$  be the orthogonal projection of  $L^2(M, \bar{\omega})$  onto the eigenspace  $V_\lambda$  of the eigenvalue  $\lambda$  of  $\Delta$ . Then, in the notation of Theorem 4.4,  $\{\lambda'_1(0), \dots, \lambda'_k(0)\}$  is the set of eigenvalues of  $P \circ fA^{(1)}|_{V_\lambda}$ .*

**COROLLARY 4.9.** *If  $P \circ fA^{(1)}|_{V_\lambda}$  is not just a scalar multiple of the identity, then there is at least one pair  $\lambda_i(\epsilon), \lambda_j(\epsilon)$  such that  $\lambda_i(\epsilon) \neq \lambda_j(\epsilon)$  for all sufficiently small  $\epsilon \neq 0$ .*

**5. Proof of the main theorem.**

**LEMMA 5.1.** *Let  $\lambda$  be an eigenvalue of multiplicity  $k \geq 2$  of  $\Delta$  and let  $\lambda_1(\epsilon), \dots, \lambda_k(\epsilon)$  be as in Theorem 4.4. Suppose  $n$ , the dimension of  $M$ , is at least two. Then there is a  $C^\infty$  function  $f$  such that  $\lambda_i(\epsilon) \neq \lambda_j(\epsilon)$  for all sufficiently small  $\epsilon > 0$ , for some pair  $i, j, 1 \leq i, j \leq k$ .*

*Proof.* Let  $u_1, \dots, u_k$  be an orthonormal basis of the eigenspace  $V_\lambda$ . By Corollary 4.9, we must find an  $f \in C^\infty(M)$  such that  $fA^{(1)}$  is not a scalar multiple of the identity. It is sufficient to find an  $f$  such that  $\langle fA^{(1)}u_1, u_1 \rangle \neq \langle fA^{(1)}u_2, u_2 \rangle$ . If  $u_1^2 = u_2^2$ , then there is an open set on which  $u_1 = u_2$  or  $u_1 = -u_2$ . However, the theory of second order elliptic equations [3] then implies that  $u_1 = u_2$  or  $u_1 = -u_2$  on all of  $M$ , contradicting the independence of  $u_1$  and  $u_2$ . As  $u_1^2 \neq u_2^2$ , there is a nonconstant eigenfunction  $v$  with eigenvalue  $\mu \neq 0$  such that  $\int_M v u_1^2 \bar{\omega} \neq \int_M v u_2^2 \bar{\omega}$ . From Proposition 4.6 we have

$$(5.2) \quad \langle vA^{(1)}u_i, u_i \rangle = \int_M [(\frac{1}{2} - \frac{1}{4}n)\mu - \lambda] v u_i^2 \bar{\omega}, \quad i = 1, 2.$$

Since  $n \geq 2, \mu \geq 0$ , and  $\lambda > 0$  ( $\lambda = 0$  has multiplicity one, so is excluded), we have  $(\frac{1}{2} - \frac{1}{4}n)\mu - \lambda < 0$ ; hence  $\langle vA^{(1)}u_1, u_1 \rangle \neq \langle vA^{(1)}u_2, u_2 \rangle$ .  $\square$

*Remarks 5.3* (about the above proof).

- (1) Suppose  $M$  is one-dimensional; we may as well assume  $M = T^1 = \mathbb{R}(\text{mod } 2\pi)$ . If  $\lambda = k^2$  in the above proof, then  $u_1 = \cos kt, u_2 = \sin kt, u_1^2 = \frac{1}{2}(1 + \cos 2kt), u_2^2 = \frac{1}{2}(1 - \cos 2kt), v = \cos 2kt, \mu = 4k^2$ , and so  $(\frac{1}{2} - \frac{1}{4}n)\mu - \lambda = 0$ . Thus the proof does fail, as it should, if  $\dim M = 1$ .
- (2) When we know the form of the eigenfunctions, we may be able to discover an explicit  $f$  such that the eigenvalue splits completely under the variation  $\exp(\epsilon f)g$ , for small  $\epsilon$ . For instance, let  $M$  be the flat 3-torus,  $T^1 \times T^1 \times T^1$ . The eigenvalue 3 has multiplicity 8; an orthogonal basis is  $u_1 = \sin x \sin y \sin z, u_2 = \sin x \sin y \cos z$ , etc. The  $u_i^2$  are sums and products of 1,  $\cos 2x, \cos 2y$  and  $\cos 2z$ ; the  $u_i u_j, i \neq j$ , all have as a factor  $\sin 2x, \sin 2y$  or  $\sin 2z$ . Let  $f = \cos 2x + 2\cos 2y + 4\cos 2z$ . The matrix  $(\langle fA^{(1)}u_i, u_j \rangle)$  is diagonal with distinct diagonal entries. Thus  $\exp(\epsilon f)g$  splits the eigenvalue 3.
- (3) The calculations are simpler when  $\dim M = 2$ , for  $\langle fA^{(1)}u_i, u_j \rangle = \int_M f u_i u_j \bar{\omega}$ , without assuming  $f$  to be an eigenfunction. This helps, for instance, in the case  $M = T^1 \times T^1$  and  $\lambda = 5$  (which has multiplicity 8). An orthogonal basis for the eigenspace is  $u_1 = \sin 2x \sin y, u_2 = \sin x \sin 2y, u_3 = \sin 2x \cos y$ , etc. Let  $f = \cos 2x + \cos 2y + 2\cos 4x + 3\cos 4y$ . Then the matrix  $(\int_M f u_i u_j \bar{\omega})$  is diagonal with distinct diagonal entries. Thus  $\exp(\epsilon f)g$  splits the eigenvalue 5.

Let  $F_i$  be the set of all  $f \in C^\infty(M)$  whose first  $i$  eigenvalues, arranged in increasing order, have multiplicity one. We will now prove that each  $F_i$  is open and dense; this will establish Theorem 1.1.

**LEMMA 5.4.**  *$F_i$  is dense.*

*Proof.* Suppose  $f \notin F_i$ . Let  $U$  be a neighborhood of  $f$ . Without loss of generality we may assume  $f \equiv 0$ , for otherwise we may replace  $g$  by  $(\exp f)g$  in what follows. By Lemma 5.1, there is an  $f$  such that, among the first  $i$  eigenvalues  $\lambda_1(\epsilon), \dots, \lambda_i(\epsilon)$  of  $f\Delta(\epsilon)$ , at least two are unequal for all sufficiently small  $\epsilon > 0$ , yet equal for  $\epsilon = 0$ .

Moreover, by Theorem 4.4.6, those of the above eigenvalues having multiplicity one will remain of multiplicity one for all sufficiently small  $\varepsilon > 0$ . Let  $\varepsilon > 0$  be sufficiently small in both these respects and also be such that  $\varepsilon f \in U$ . We may repeat the above process with the new metric  $\exp(\varepsilon f)g$  and after a finite number of steps the sum of the functions  $\varepsilon f$  produced will lie in  $F_i$ .  $\square$

In order to prove that  $F_i$  is open in  $C^\infty(M)$ , we need to establish a bound on how fast a given eigenvalue of  $\Delta$  can grow under the variation of metric  $\exp(\varepsilon f)g$ . Although the next Lemma follows from the general perturbation theory (see [6, p. 391]) and the results of the previous section, an ad hoc proof in our special setting seems appropriate here because it is important to our proof that  $F_i$  is open, and it is not particularly intuitively evident.

**LEMMA 5.5.** *Let  $\lambda$  be an eigenvalue of  $\Delta$  of multiplicity  $k$ . There is a neighborhood  $U$  of  $0 \in C^\infty(M)$ , an interval  $[-\alpha, \alpha]$ ,  $\alpha > 0$ , and positive constants  $a$  and  $b$  ( $U, \alpha, a$ , and  $b$  independent of  $\lambda$ ) such that, for all  $f \in U$  and  $\varepsilon$  in  $[-\alpha, \alpha]$ , we have that  $|\lambda_i(\varepsilon) - \lambda| \leq (a + b\lambda)(e^{b|\varepsilon|} - 1)/b$ , where  $\lambda_i(\varepsilon)$  ( $1 \leq i \leq k$ ) are as in Theorem 4.4.*

*Proof.* Let  $\delta$  be a real number,  $f \in C^\infty(M)$ , and  $\langle \cdot, \cdot \rangle_{\delta f}$  the  $L^2$  inner product on  $C^\infty(M)$  relative to the metric  $\exp(\delta f)g$ . Let  $\lambda(\varepsilon)$  be an analytic continuation of  $\lambda$  with respect to the variation  $\exp(\varepsilon f)g$  of the original metric  $g$ . Then  $\mu(\varepsilon) = \lambda(\varepsilon + \delta)$  is an analytic continuation of  $\lambda(\delta)$  with respect to the variation  $\exp((\varepsilon + \delta)f)g$  of the metric  $\exp(\delta f)g$ . Let  ${}_rA^\delta(\varepsilon) = \exp(\frac{1}{4}\varepsilon f)\Delta(\varepsilon + \delta)[\exp(-\frac{1}{4}\varepsilon f)h]$ ; note this differs from  ${}_rA(\varepsilon + \delta)$ , but has the same eigenvalues (the latter is not used because we want to get the Laplace operator when  $\varepsilon = 0$ ). Let  $\psi(\varepsilon)$  be an eigenfunction of  ${}_rA^\delta(\varepsilon)$  for the eigenvalue  $\mu(\varepsilon)$ , as given in Theorem 4.4. Then, according to Proposition 4.3, we have (where  $h = \psi(0)$ )

$$(5.6) \quad \| [{}_rA^\delta]^{(1)} h \|_{\delta f} \leq a(\delta, f) \| h \|_{\delta f} + b(\delta, f) \| \Delta(\delta) h \|_{\delta f}$$

where  $a(\delta, f)$  and  $b(\delta, f)$  are constants which, by inspection of the proof of Proposition 4.3, can be made jointly continuous in  $\delta$  and  $f$ . Hence, this inequality remains valid if we replace  $a(\delta, f)$  and  $b(\delta, f)$  by  $a \equiv a(0, 0) + 1$  and  $b \equiv b(0, 0) + 1$ , respectively, provided that  $(\delta, f) \in I \times U \subset \mathbb{R} \times C^\infty(M)$ , where  $I \times U$  is a suitably small neighborhood of  $(0, 0)$ . Using Proposition 4.7 the Cauchy-Schwarz inequality with respect to  $\langle \cdot, \cdot \rangle_{\delta f}$ , and (5.6), we have  $|\lambda'(\delta)| = |\mu'(0)| \leq a + b\lambda(\delta)$ , which implies the desired inequality.  $\square$

**LEMMA 5.7.**  $F_i$  is open.

*Proof.* Without loss of generality, we may assume  $f \neq 0$ , for otherwise we may replace  $g$  by  $\exp(f)g$  in the following argument. Thus, we are assuming that the first  $i$  eigenvalues of  $\Delta$ , labeled  $\lambda_1, \dots, \lambda_i$ , are of multiplicity one. Let  $I_1, \dots, I_i$  be disjoint closed intervals centered on  $\lambda_1, \dots, \lambda_i$  of lengths  $2d_1, \dots, 2d_i$ , respectively. Moreover, we may assume  $\lambda_{i+1} \notin I_i$ . Let  $a$  and  $b$  be the constants in the previous Lemma. Then select  $\beta > 0$  such that  $|\varepsilon| \leq \beta$  implies  $(a + b\lambda_j)(e^{b|\varepsilon|} - 1)/b \leq d_j$  for all  $j$ ,  $1 \leq j \leq i$ . Choose  $\gamma > 0$  such that  $|\varepsilon| \leq \gamma$  implies  $0 < a(e^{b|\varepsilon|} - 1)/b + \lambda_i + d_i < (2 - e^{b|\varepsilon|})\lambda_{i+1}$ . Note  $2 - e^{b|\varepsilon|}$  is necessarily positive, so the inequality is preserved if we replace  $\lambda_{i+1}$  by  $\lambda_p$  for any  $p > i$ . Thus for all  $p > i$  and  $|\varepsilon| \leq \gamma$ , we have  $(a + \lambda_p b)(e^{b|\varepsilon|} - 1)/b < \lambda_p - \lambda_i - d_i$ . Let  $\alpha$  and  $U$  be as in the previous Lemma and set  $\delta = \min\{\alpha, \beta, \gamma\}$ . Then  $\varepsilon f \in F_i$  for all  $f \in U$  and  $|\varepsilon| \leq \delta$ . Thus,  $\delta U$  is an open neighborhood of 0 contained in  $F_i$ .  $\square$

By the remark before Lemma 5.4, the proof of Theorem 2.1 is now complete.

**COROLLARY 5.8.** *For  $f$  in the residual set  $F_\infty$ , we have that the eigenvalues of  $\Delta(\varepsilon)$  are all of multiplicity one for all real  $\varepsilon$  outside a countable subset of  $\mathbb{R}$ .*

*Proof.* Let  $\lambda_i(\varepsilon)$  be the analytic continuations of the eigenvalues of  $\Delta$  under the variation  $\exp(\varepsilon f)g$  (see Remark after Proposition 4.5). Since  $f$  is in  $F_\infty$ , we know that no

pair of the  $\lambda_i(\varepsilon)$  agree at  $\varepsilon = 1$ , and hence the  $\lambda_i(\varepsilon)$  are distinct analytic functions. Thus any two of them agree on at most a countable set. Taking the countable union of these countable sets over all pairs of  $\{\lambda_i(\varepsilon)\}$ , we still have just a countable set.  $\square$

## REFERENCES

- [1] J. H. ALBERT, *Genericity of simple eigenvalues for elliptic PDE's*, Proc. Amer. Math. Soc., 48 (1975), pp. 413–418.
- [2] M. BERGER, *Le Spectre d'une Variété Riemannienne*, Lecture Notes in Mathematics, no. 194, Springer-Verlag, Berlin, 1971.
- [3] A. P. CALDERÓN, *Uniqueness in the Cauchy problem for partial differential equations*, Amer. J. Math., 80 (1958), pp. 16–36.
- [4] B. H. DRISCOLL, *The Multiplicity of the Eigenvalues of a Symmetric Drum*, Dissertation, Northwestern University, 1978.
- [5] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Springer-Verlag, New York Inc., 1973.
- [6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag New York Inc., 1966.
- [7] F. RELICH, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
- [8] K. UHLENBECK, *Generic properties of eigenfunctions*, Amer. J. Math., 98, no. 4 (1976), pp. 1059–1078.



## THE EIGENVALUES OF AN EQUILATERAL TRIANGLE\*

MARK A. PINSKY†

**Abstract.** Let  $D$  be an equilateral triangle of side 1. We consider solutions of  $\Delta u + \lambda u = 0$  in  $D$  with either the boundary condition  $u = 0$  or  $\partial u / \partial n = 0$ . Let  $n(\lambda)$  be the number of distinct eigenvalues  $\leq \lambda$ ,  $N(\lambda)$  be the total number of eigenvalues  $\leq \lambda$ , including multiplicities. Theorem 1 states that for either boundary condition,  $\lambda_{mn} = (16\pi^2/27)(m^2 + n^2 - mn)$ , where  $m + n \equiv 0 \pmod{3}$ . In the first case it is further required that  $m \neq 2n$ ,  $n \neq 2m$ . Theorem 2 states that  $\lim_{\lambda \rightarrow \infty} (N(\lambda)/n(\lambda)) = \infty$ . The proof uses the representation of  $\lambda_{mn}$  as the norm of an integer in the quadratic number field  $k(\omega)$ , where  $\omega$  is a primitive cube root of unity. These results contrast with the generic results for domains with  $Z_3$  symmetry obtained by V. Arnold (Functional Anal. Appl., 1972).

**1. Introduction.** Let  $D$  be an equilateral triangle of unit side. We are concerned with the eigenvalue problem

$$\begin{aligned} \Delta f + \lambda f &= 0 && \text{in } D, \\ f &= 0 && \text{on } D. \end{aligned}$$

This problem was studied by Lamé [7, p. 131–136], who obtained an integral quadratic form for the eigenvalues. In trying to understand Lamé’s work, we found the following shortcomings: (i) the complete list of eigenvalues is not clearly specified, either by the allowable entries of the quadratic form or by any intrinsic characterization, and (ii) he gives no methods for computing the multiplicities of the eigenvalues. Thus it is of interest to give an up-to-date treatment covering these two points.

In §§ 2 and 3 we give a self-contained elementary derivation of Lamé’s result, both for Dirichlet and Neumann boundary conditions. In § 4 we make the previously unpublished observation that any eigenvalue can be written as the norm of an integer in the quadratic number field  $Q(\sqrt{-3})$ , which allows us to find a formula for the multiplicity of any eigenvalue. Finally, this is used in § 5 to prove that the “average multiplicity” when suitably defined becomes infinite when  $\lambda \rightarrow \infty$ .

These questions have also been of interest recently in connection with the non-generic behavior in domains with  $120^\circ$  rotational symmetry [3], [4]. For any such domain the eigenfunctions can be chosen to be symmetric ( $Rf = f$ ) or complex ( $Rf = e^{\pm 2\pi i/3} f$ ), where  $R$  denotes a rotation of  $120^\circ$ . It is shown below that for a given eigenvalue of the equilateral triangle, the eigenfunctions are either all symmetric or all complex (no hybrid eigenvalues are possible). The dimensions of the eigenspaces can be arbitrarily large, for both symmetric and complex eigenvalues. This is in marked contrast to the unit disk, where the eigenvalues are all of multiplicity one or two.

**2. Eigenvalues of the Dirichlet problem.** We classify the eigenvalues of the Laplacian on the triangular domain

$$(2.1) \quad D = \{(x, y) : 0 < y < x\sqrt{3}, y < \sqrt{3}(1-x)\},$$

with zero boundary conditions.

**THEOREM 1.** *The eigenvalues of  $\Delta$  on  $D$  are the numbers*

$$(2.2) \quad \lambda_{mn} = \left(\frac{16\pi^2}{27}\right)(m^2 + n^2 - mn), \quad m, n = 0, \pm 1, \dots,$$

\* Received by the editors August 16, 1979, and in revised form November 29, 1979.

† Department of Mathematics, Northwestern University, Evanston, Illinois, 60201.

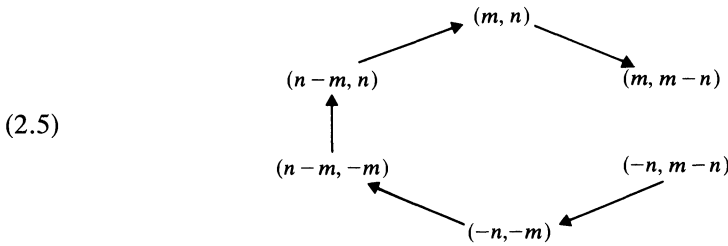
with the following conditions:

- (2.3) (A)  $m + n$  is a multiple of 3,  
 (B)<sup>1</sup>  $m \neq 2n$ ,  
 (C)<sup>1</sup>  $n \neq 2m$ .

The multiplicity of  $\lambda_{mn}$  is  $(1/6) \times$  the number of times which it appears in the lattice (2.2). The eigenfunctions are of the form

$$(2.4) \quad f(x, y) = \sum_{(m,n)} \pm \exp\left(\frac{2\pi i}{3}\right) \left( nx + \frac{(2n-m)y}{\sqrt{3}} \right).$$

In this sum  $(m, n)$  range over  $\mathcal{S} \subseteq \mathbb{Z}^2$ , where  $|\mathcal{S}| = 6$  and  $\pm$  is determined by the following rules:



Each transition induces a change of sign in the  $(m, n)$  entry of (2.4).

Note. Lamé’s formula is  $\lambda_{\mu\nu} = (16\pi^2/9)(\mu^2 + \nu^2 + \mu\nu)$ . This can be obtained from our formula by means of the substitution  $m = 2\mu + \nu, n = \mu - \nu$ . In these variables, (B) and (C) state that  $\mu \neq 0, \mu + \nu \neq 0$ .

We introduce the rotation operator by

$$R: (x, y) \rightarrow \left( 1 - \frac{x}{2} - \frac{y\sqrt{3}}{2}, \frac{x\sqrt{3}}{2} - \frac{y}{2} \right).$$

An eigenfunction is said to be symmetric if  $R \circ f = f, \vec{x} \in D$ . An eigenfunction is said to be complex if  $R \circ f = \exp(\pm 2\pi i/3)f, \vec{x} \in D$ .

COROLLARY 1. The eigenvalue  $\lambda_{mn}$  belongs to a symmetric eigenfunction iff the following additional condition is met:

- (D)  $m$  is a multiple of 3,

(equivalently, the associated eigenfunction is periodic in  $x$  with period 1). The eigenvalue  $\lambda_{mn}$  belongs to a complex eigenfunction iff  $m$  is congruent to  $+1$  or  $-1$ , modulo 3. In particular, an eigenvalue cannot belong to both a complex eigenfunction and a symmetric eigenfunction.

COROLLARY 2. The following are symmetric eigenfunctions and give a complete list of the simple eigenvalues.

$$(2.6) \quad f_p(x, y) = \sin(2\pi p\bar{d}_1) + \sin(2\pi p\bar{d}_2) + \sin(2\pi p\bar{d}_3), \quad p = 1, 2, \dots,$$

<sup>1</sup> It is required that (B) and (C) be satisfied for all pairs  $(m, n)$  which appear in (2.5).

where

$$\bar{d}_1 = \frac{2y}{\sqrt{3}}$$

$$\bar{d}_2 = x - \frac{y}{\sqrt{3}}$$

$$\bar{d}_3 = 1 - x - \frac{y}{\sqrt{3}}$$

$\bar{d}_1, \bar{d}_2, \bar{d}_3$  are the normalized altitudes of the point  $(x, y)$  in the triangle  $D$ . They satisfy the normalization  $\bar{d}_1 + \bar{d}_2 + \bar{d}_3 = 1$  and the reflection laws  $R_i \circ d_i = -d_i, i = 1, 2, 3$ . The eigenvalues are obtained by the formula

$$\lambda_{3p,0} = \left(\frac{16\pi^2}{27}\right)(9p^2).$$

To prove these results, we introduce the parallelogram

$$\tilde{D} = \left\{ (x, y) : 0 < y < \frac{3\sqrt{3}}{2}, \frac{y}{\sqrt{3}} < x < 3 + \frac{y}{\sqrt{3}} \right\},$$

and the reflection operators

$$R_1: (x, y) \rightarrow (x, -y),$$

$$R_2: (x, y) \rightarrow \left(-\frac{x}{2} + \frac{y\sqrt{3}}{2}, \frac{x\sqrt{3}}{2} + \frac{y}{2}\right),$$

$$R_3: (x, y) \rightarrow \left(\frac{3}{2} - \frac{x}{2} - \frac{y\sqrt{3}}{2}, \frac{y}{2} + \frac{\sqrt{3}}{2} - \frac{x\sqrt{3}}{2}\right).$$

There is a canonical isomorphism between  $L^2(D)$  and the following subspace of  $L^2(\tilde{D})$ ,

$$(2.7) \quad H = \{f \in L^2(\tilde{D}) : R_i f = -f, i = 1, 2, 3\},$$

obtained by  $f \rightarrow f|_D$  for  $f \in H$ . Thus, any eigenfunction of  $\Delta$  on  $D$  can be obtained by solving the equation on  $H$ . The restriction to  $D$  will automatically satisfy the Dirichlet boundary conditions. By a standard result in [1, p. 148], we obtain a complete list of the eigenfunctions of  $\Delta$  on  $D$  given by linear combinations of

$$(2.8) \quad \tilde{f}(x, y) = \exp(i(\alpha x + \beta y)),$$

where  $(\alpha, \beta)$  are in the dual lattice. This requires that  $3\alpha = 2n\pi, 3\alpha/2 + 3\sqrt{3}\beta/2 = 2m\pi$  for integers  $(m, n)$ . Thus  $\alpha = 2n\pi/3, \beta = 2\pi(2m - n)/(3\sqrt{3})$ . It is readily verified that, corresponding to this,

$$(2.9) \quad \begin{aligned} \lambda_{mn} &= \alpha^2 + \beta^2, \\ &= \left(\frac{2n\pi}{3}\right)^2 + (2\pi)^2 \left(\frac{(2m - n)^2}{27}\right), \\ &= \left(\frac{16\pi^2}{27}\right)(m^2 + n^2 - mn). \end{aligned}$$

The eigenfunction is therefore of the form

$$(2.10) \quad \tilde{f} = \sum_{(m,n)} A_{mn} \exp\left(\frac{2\pi i}{3}\right) \left[ nx + \frac{(2m-n)y}{\sqrt{3}} \right],$$

where the sum is over  $(m, n)$  with  $\lambda_{mn} = \lambda$ . To satisfy the reflection conditions, we write

$$\begin{aligned} R_1 \circ \tilde{f} &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right) \left( n'x - \frac{(2m'-n')y}{\sqrt{3}} \right), \\ &= \sum_{mn} A_{m-n,n} \exp\left(\frac{2\pi i}{3}\right) \left( nx + \frac{(2m-n)y}{\sqrt{3}} \right), \end{aligned}$$

where we have made the substitution  $m' = n - m, n' = n$ . Hence  $R_1 \circ \tilde{f} = -\tilde{f}$  requires that

$$(2.11) \quad A_{mn} = -A_{n-m,n}.$$

The second reflection operator is

$$\begin{aligned} R_2 \circ \tilde{f} &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right) \left[ n' \left( \frac{-x}{2} + \frac{y\sqrt{3}}{2} \right) + (2m' - n') \left( \frac{x}{2} + \frac{y}{2\sqrt{3}} \right) \right], \\ &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right) \left[ (m - n')x + \frac{(m' + n')y}{\sqrt{3}} \right], \\ &= \sum_{mn} A_{m,m-n} \exp\left(\frac{2\pi i}{3}\right) \left( nx + \frac{(2m-n)y}{\sqrt{3}} \right). \end{aligned}$$

Therefore, we must have

$$(2.12) \quad A_{mn} = -A_{m,m-n}.$$

The third reflection operator is

$$\begin{aligned} R_3 \circ \tilde{f} &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right) \left[ n' \left( \frac{3-x}{2} - \frac{y\sqrt{3}}{2} \right) + (2m' - n') \left( \frac{1}{2} - \frac{x}{2} + \frac{y}{2\sqrt{3}} \right) \right], \\ &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right) \left[ (n' + m') \exp\left(\frac{2\pi i}{3}\right) \left( -m'x + \frac{(m' - 2n')y}{\sqrt{3}} \right) \right], \\ &= \sum_{mn} A_{-n,-m} \exp\left(\frac{2\pi i}{3}\right) (-m - n) \exp\left(\frac{2\pi i}{3}\right) \left( mx + \frac{(2m-n)y}{\sqrt{3}} \right). \end{aligned}$$

Thus we must have

$$(2.13) \quad A_{m,n} = -A_{-n,-m} \exp\left(\frac{2\pi i}{3}\right) (-m - n).$$

Now if for a fixed  $(m_0, n_0)$  we have  $A_{m_0,n_0} = A$ , then by iterating (2.11) and (2.12) and referring to the graph (2.5), we see that  $\exp(2\pi i/3)(-m-n) = 1$ , i.e.,  $m+n$  is a multiple of 3, which proves condition (A). To prove condition (B) assume to the contrary that  $m_0 = 2n_0$ . This is the same as  $(m_0, n_0) = (m_0, m_0 - n_0)$ . Property (2.12) therefore requires that  $A = -A$ , i.e.,  $A = 0$ . Similarly, to prove condition (C), we note that  $n_0 = 2m_0$  is the same as  $(m_0, n_0) = (n_0 - m_0, n_0)$ , which by property (2.13) requires

that  $A = 0$ . Conversely, we verify directly that any sum of the form (2.4) satisfies the reflection conditions (2.7) and is therefore an eigenfunction. Thus we have proved the theorem.

To prove Corollary 1 we study the effect of the rotation operator on (2.10):

$$\begin{aligned} R \circ \tilde{f} &= \sum_{m'n'} A_{m'n'} \exp\left(\frac{2\pi i}{3}\right)(n') \exp\left(\frac{2\pi i}{3}\right)\left[(m' - n')x - \frac{(m' + n')y}{\sqrt{3}}\right], \\ &= \sum_{mn} A_{n-m,-m} \exp\left(\frac{2\pi i}{3}\right)(-m) \exp\left(\frac{2\pi i}{3}\right)\left[mx + \frac{(2m - n)y}{\sqrt{3}}\right]. \end{aligned}$$

But (2.11) and (2.12) require that  $A_{n-m,-m} = A_{m,n}$ . Therefore, we must have  $\exp(2\pi i/3)(-m) = 1$ , i.e.,  $m$  is a multiple of 3. Similarly, if  $m$  is congruent to  $\pm 1$  modulo 3, it is clear that  $R \circ \tilde{f} = e^{\pm 2\pi i/3} \tilde{f}$ , i.e.,  $f$  is a complex eigenfunction.

To prove Corollary 2 we first note that for the choice  $m = 3p, n = 0$ , the formula (2.4) yields the symmetric eigenfunction (2.6). We will now show that these correspond to the only simple eigenvalues of  $\Delta$  on  $D$ . Indeed, suppose that  $\lambda_{mn}$  is a simple eigenvalue corresponding to the set  $\mathcal{S}$  described in (2.5). If  $(m, 0) \in \mathcal{S}$  for some  $m$ , then by Theorem 1,  $m = 3p$  and the result is proved. Therefore, we may suppose that  $\mathcal{S}$  contains no pair of the form  $(m, 0)$ . Thus we may write

$$\mathcal{S} = \{(m_0, n_0), (m_0, m_0 - n_0), (-n_0, m_0 - n_0), (-n_0, -m_0), (n_0 - m_0, -m_0), (n_0 - m_0, n_0)\}.$$

By hypothesis we must have  $n_0 \neq 0, m_0 \neq 0, n_0 \neq m_0$ . Define a new set

$$\tilde{\mathcal{S}} \equiv \{(n_0, m_0), (n_0, n_0 - m_0), (-m_0, n_0 - m_0), (-m_0, n_0 - m_0), (m_0 - n_0, n_0), (m_0 - n_0, m_0)\}.$$

Clearly  $\mathcal{S} \cap \tilde{\mathcal{S}} = \emptyset$ , and therefore  $\tilde{\mathcal{S}}$  can be used to manufacture a new eigenfunction according to formula (2.4) with the same eigenvalue—a contradiction. Therefore  $(m, 0) \in \mathcal{S}$  for some  $m$  and necessarily  $m = 3p$  by Theorem 1. This again leads to formula (2.6), which was to be proved.

**3. Eigenvalues of the Neumann problem.** The methods of §§ 1 and 2 can also be applied to enumerate the eigenvalues of the problem

$$\Delta f + \lambda f = 0, \quad x \in D, \quad \frac{\partial f}{\partial n} \Big|_{\partial D} = 0.$$

Indeed, given  $f$  on  $D$ , we lift  $f$  to a function  $\tilde{f}$  on  $\tilde{D}$  satisfying  $R_i \circ \tilde{f} = +\tilde{f}, x \in D, i = 1, 2, 3; \tilde{f}|_D = f$ .  $\tilde{f}$  will still be an eigenfunction on  $D$  and hence a linear combination of (2.8) with the same values of  $(\alpha, \beta)$ . Applying the reflection operation, we have the following result.

**PROPOSITION 3.** *The eigenvalues of  $\Delta$  on  $D$  with Neumann boundary conditions are given by the numbers*

$$\lambda_{mn} = \left(\frac{16\pi^2}{27}\right)(m^2 + n^2 - mn),$$

where  $m + n$  is a multiple of 3. The eigenfunctions are of the form

$$f(x, y) = \sum_{(m,n)} \exp\left(\frac{2\pi i}{3}\right)\left(nx + \frac{(2m - n)y}{\sqrt{3}}\right),$$

where  $(m, n)$  range over  $\mathcal{S} \subset Z^2$  where  $|\mathcal{S}| = 6$  determined by the transformation (2.5).

**4. Number-theoretic discussion of the eigenvalues.** The eigenvalues can be classified using the factorization theory in the ring  $Z(\omega)$ , where  $\omega = (-1 + \sqrt{-3})/2$  is a primitive cube root of unity [5, Chapters 14–15]. Indeed, we can write

$$(4.1) \quad \begin{aligned} \tilde{\lambda} &= m^2 + n^2 - mn, \\ &= |m + n\omega|^2. \end{aligned}$$

We refer to  $\tilde{\lambda}$  as a *normalized eigenvalue*. It is shown that  $Z(\omega)$  has the unique factorization property:  $a = m + n\omega = p_1^{\alpha_1} \cdots p_r^{\alpha_r} u$  where  $u$  is one of the six possible units  $(\pm\omega^k)$ ,  $k = 0, 1, 2$ , and the primes  $(p_i)$  are one of the following types:

- (1)  $p = \pi_3 \equiv (1 - \omega)$ ;
- (2)  $p$  is a rational prime of the form  $3s + 2$ ;
- (3)  $p$  is a factor  $m + n\omega$  of a rational prime of the form  $3s + 1$ ;

these primes are denoted  $\pi_7, \bar{\pi}_7, \pi_{13}, \bar{\pi}_{13}, \dots$ .

**PROPOSITION 4.** *Let  $\tilde{\lambda} = m^2 + n^2 - mn$  be a normalized eigenvalue of the equilateral triangle with Dirichlet boundary conditions. Then the prime factorization of  $\tilde{\lambda}$  has the form*

$$(4.2) \quad \tilde{\lambda} = 3^{\alpha_0} 2^{2\alpha_1} 5^{2\alpha_2} \dots 7^{\gamma_1} 13^{\gamma_2} \dots,$$

where

- (1)  $\alpha_0 \geq 1$ ,
- (2) if  $\alpha_0 = 1$  then  $\gamma_i > 0$  for some  $i$ ,
- (3) multiplicity  $(\tilde{\lambda}) = \begin{cases} 1 & \gamma_i \equiv 0, \\ \prod_i (1 + \gamma_i) & \alpha_0 = 1 \text{ or } \gamma_i \text{ odd for some } i, \\ \prod_i (1 + \gamma_i) - 1 & \alpha_0 = 1 \text{ and } \gamma_i \text{ even for all } i, \end{cases}$
- (4)  $\tilde{\lambda}$  is symmetric iff  $\alpha_0 \geq 2$ .

Conversely, any integer of the form (4.2) satisfying (1) and (2) is a normalized eigenvalue. For Neumann boundary conditions we need only require  $\alpha_0 \geq 1$ . In this case, the multiplicity  $(\tilde{\lambda}) = \prod_i (1 + \gamma_i)$  in all cases.

*Proof.* We first translate the known conditions on  $(m, n)$ .

*Fact 1.*  $m + n \equiv 0 \pmod{3}$  iff  $\alpha_0 \geq 1$ .

Indeed, if  $\alpha_0 \geq 1$ , then  $a = (1 - \omega)(\tilde{m} + \tilde{n}\omega) = (\tilde{m} + \tilde{n}) + (2\tilde{n} - \tilde{m})\omega = (m + n\omega)$  and clearly  $m + n = 3\tilde{n} \equiv 0 \pmod{3}$ . Conversely, if  $m + n\omega \in Z(\omega)$ ,  $(m + n\omega)/(1 - \omega) = [\frac{1}{3}(m + n) - n] + \omega[(m + n)/3]$ . Hence, if  $m + n \equiv 0 \pmod{3}$ , then the quotient is again  $\in Z(\omega)$  and thus  $\alpha_0 \geq 1$ .

*Fact 2.*  $m \equiv 0 \pmod{3}$ ,  $n \equiv 0 \pmod{3}$  iff  $\alpha_0 \geq 2$ .

Indeed, if  $\alpha_0 \geq 2$ , then  $a = (1 - \omega)^2(\tilde{m} + \tilde{n}\omega) = 3\tilde{n} + \bar{\omega}(3\tilde{n} - 3\tilde{m})$  and clearly  $m \equiv 0 \pmod{3}$ ,  $n \equiv 0 \pmod{3}$ . Conversely, if  $m + n\omega \in Z(\omega)$ ,  $(m + n\omega)/(1 - \omega)^2 = -\frac{1}{3}[n + m\bar{\omega}]$ . If  $m \equiv 0 \pmod{3}$ ,  $n \equiv 0 \pmod{3}$ , this quotient is again  $\in Z(\omega)$  and thus  $\alpha_0 \geq 2$ .

*Fact 3.*  $m = 2n$  iff  $a$  is an integral multiple of  $\pi_3\bar{\omega}$ , and  $n = 2m$  iff  $a$  is an integral multiple of  $\pi_2\omega$ .

Indeed, if  $m = 2n$ , then  $a = m + n\omega = n(2 + \omega) = n\pi_3\bar{\omega}$ . Conversely, if  $a = n\pi_3\bar{\omega}$ , then  $a = 2n + n\omega$  and thus  $m = 2n$ . In case  $n = 2m$ , we write  $a = m + n\omega = m(1 + 2\omega) = m\pi_3\omega$ , to obtain the same conclusion.

Putting these facts together proves parts (1), (2), (4) of the proposition. Indeed,  $m \neq 2n, n \neq 2m$  requires that  $a/\pi_3$  not be an integer. This requires that the product  $\pi_7^{\beta_1} \bar{\pi}_7^{\beta_1} \cdots$  not be an integer, which is equivalent to  $\beta_i \neq \tilde{\beta}_i$  for some  $i$ . Finally,  $m + n \equiv 0 \pmod{3}$  requires that  $\alpha_0 \equiv 1$ , by Fact 1.

To determine the multiplicities, we must first determine the nonuniqueness of (4.2). For this purpose, let  $a, a'$  be two possible representations:  $\tilde{\lambda} = |a|^2 = |a'|^2$  where

$$a = \pi_3^{\alpha_0} 2^{\alpha_1} 5^{\alpha_2} 11^{\alpha_3} \cdots \pi_7^{\beta_1} \bar{\pi}_7^{\tilde{\beta}_1} \cdots u,$$

$$a' = \pi_3^{\alpha'_0} 2^{\alpha'_1} 5^{\alpha'_2} 11^{\alpha'_3} \cdots \pi_7^{\beta'_1} \bar{\pi}_7^{\tilde{\beta}'_1} \cdots u'.$$

The equation  $|a|^2 = |a'|^2$  requires that

$$3^{\alpha_0} 2^{2\alpha_1} 5^{2\alpha_2} \cdots 7^{\beta_1 + \tilde{\beta}_1} = 3^{\alpha'_0} 2^{2\alpha'_1} 5^{2\alpha'_2} \cdots 7^{\beta'_1 + \tilde{\beta}'_1}.$$

By the unique factorization for rational integers, we must have  $\alpha_0 = \alpha'_0, \alpha_1 = \alpha'_1, \cdots, \beta_1 + \tilde{\beta}_1 = \beta'_1 + \tilde{\beta}'_1, \cdots$ . Therefore, the multiplicity of a given eigenvalue is equal to the number of ways of redistributing the exponents  $\beta_i, \tilde{\beta}_i$  subject to those conditions. If  $\alpha_0 \equiv 2$ , clearly this is equal to the product  $\prod_i (1 + \gamma_i)$ . If  $\alpha_0 = 1$ , we must exclude the case  $\beta_i = \tilde{\beta}_i$  for all  $i$ . This can only happen if  $\gamma_i$  is even for all  $i$ .

**COROLLARY.** *For an integer  $k$ , there exists an eigenvalue  $\lambda$  of multiplicity  $k$ ;  $\tilde{\lambda}$  may be chosen to be symmetric.*

*Proof.* Let  $a = \pi_3^2 \pi_7^{k-1}$  be a possible representation for  $\tilde{\lambda}$ . The remaining possible representatives are  $\pi_3^2 \pi_7^{k-2} \bar{\pi}_7, \pi_3^2 \pi_7^{k-2} \bar{\pi}_7^2, \cdots, \pi_3^2 \bar{\pi}_7^{k-1}$ . Each of these is admissible and corresponds to an eigenfunction. For this choice,  $\tilde{\lambda} = 9 \cdot 7^{k-1}$ .

*Example.* For  $k = 3$ , we have the values  $a = 15 - 9\omega, a = -21\omega, a = -15 - 24\omega$ , with  $\tilde{\lambda} = 441$ . This gives the simplest example of a degenerate eigenvalue of odd multiplicity.

*Note.* Formula (3) of Proposition 4 also follows from [8, Satz 204, p. 144].

**5. Multiplicity of the eigenvalues.** In this section we obtain some qualitative estimates on the number of multiple eigenvalues. The main result is that the “average multiplicity” becomes infinite in the limit  $\lambda \rightarrow \infty$ .

To prove this, we introduce the functions

$$(5.1) \quad n_k(\lambda) = \#\{j: \lambda_j \leq \lambda, \lambda_j \text{ is a } k\text{-fold eigenvalue}\},$$

$$(5.2) \quad n(\lambda) = \sum_1^\infty n_k(\lambda),$$

$$(5.3) \quad N(\lambda) = \sum_1^\infty k n_k(\lambda).$$

Since there are only a finite number of eigenvalues in each finite interval, both sums (5.2) and (5.3) contain only a finite number of nonzero terms;  $n(\lambda)$  counts the number of (unrepeated) eigenvalues  $\leq \lambda$ ,  $N(\lambda)$  counts the total number of eigenvalues, including multiplicities. The theorem of Hermann Weyl states that

$$(5.4) \quad N(\lambda) \sim \frac{A}{4\pi} \lambda, \quad \lambda \rightarrow \infty,$$

where  $A = \sqrt{3}/4$ . To study the average multiplicity, we define

$$m = \lim_{\lambda \rightarrow \infty} \frac{N(\lambda)}{n(\lambda)}.$$

**THEOREM 2.**  $m = \infty$ .

To prepare the proof, we study certain sets of integers. Let  $P = (p_j)$  be a set of prime numbers such that

$$\sum_j \frac{1}{p_j} = \infty.$$

Let  $\mathcal{S} = \{n: P_j \nmid n \text{ for all } j\}$ ,  $f(x) = \sum_{n \in \mathcal{S}, n \leq x} 1$ .

LEMMA.  $f(x) = o(x)$ ,  $x \rightarrow \infty$ .

Proof. Set  $\mathcal{S}_N = \{n: p_j \nmid n \text{ for } 1 \leq j \leq N\}$ ,

$$f_N(x) = \sum_{n \in \mathcal{S}_N, n \leq x} 1,$$

Clearly,  $\mathcal{S} \subseteq \mathcal{S}_N$  for each  $N$  and thus  $f(x) \leq f_N(x)$ . But it is well known that

$$\frac{1}{x} f_N(x) \rightarrow \prod_{j=1}^N \left(1 - \frac{1}{p_j}\right), \quad x \rightarrow \infty, \quad N = 1, 2, \dots$$

Thus

$$\limsup_{x \rightarrow \infty} \frac{f(x)}{x} \leq \prod_{j=1}^N \left(1 - \frac{1}{p_j}\right), \quad N = 1, 2, \dots$$

But  $\sum 1/p_j = \infty$  implies that the product  $\prod_{j=1}^N (1 - (1/p_j))$  tends to zero when  $N \rightarrow \infty$ . Therefore  $\limsup_{x \rightarrow \infty} f(x)/x = 0$ , which was to be proved.

Proof of the theorem. Let  $P_1 = \{2, 5, 11, 17, \dots\}$ ,  $P_2 = \{3, 7, 13, 19, \dots\}$ .  $P_1$  contains all primes of the form  $3n + 2$ , whereas  $P_2$  contains all primes of the form  $3n + 1$  together with the prime 3. It is known [6, p. 52] that  $\sum 1/p_j = \infty$  for both  $P_1, P_2$ . Let  $\mathcal{S}_i (i = 1, 2)$  be the sets formed from  $P_i, i = 1, 2$ , in the above manner. Now recall that if  $\lambda$  is an eigenvalue,  $\lambda = (16\pi^2/27)\tilde{\lambda}$ , where

$$\begin{aligned} \tilde{\lambda} &= 3^{\alpha_0} 2^{2\alpha_1} 5^{2\alpha_2} \dots 7^{\gamma_1} 13^{\gamma_2} \dots, \\ &= u^2 v, \end{aligned}$$

where  $u \in \mathcal{S}_2, v \in \mathcal{S}_1$ . Thus,

$$\begin{aligned} \#(\lambda \leq x) &\leq \sum_{\{u \in \mathcal{S}_2, v \in \mathcal{S}_1, u^2 v \leq x\}} 1, \\ &= \sum_{v \in \mathcal{S}_1} \sum_{u \leq \sqrt{x}/v, u \in \mathcal{S}_2} 1, \\ &= \int_1^x f_2\left(\frac{\sqrt{x}}{v}\right) df_1(v). \end{aligned}$$

Now let  $\epsilon > 0$ . Thus  $f_2(u) < \epsilon u$  for  $u \geq M$ . Let  $K = \max_{1 \leq x \leq M} f(x)$ . Thus

$$\#(\tilde{\lambda} \leq x) = I + II,$$

where

$$\begin{aligned} I &= \int_1^{x/M^2} f_2\left(\frac{\sqrt{x}}{v}\right) df_1(v), \\ II &= \int_{x/M^2}^x f_2\left(\frac{\sqrt{x}}{v}\right) df_1(v). \end{aligned}$$



Now

$$\begin{aligned} I &\leq \varepsilon \int_1^{x/M^2} \sqrt{\frac{x}{v}} df_1(v), \\ &= \varepsilon \sqrt{x} \left\{ \frac{M}{\sqrt{x}} f_1\left(\frac{x}{M^2}\right) + \frac{1}{2} \int_1^{x/M^2} \frac{f_1(v)}{v^{3/2}} dv \right\}. \end{aligned}$$

But  $f_1(x) = o(x)$ ,  $x \rightarrow \infty$  implies that both terms are  $o(\sqrt{x})$ ,  $x \rightarrow \infty$ . Hence,  $I = o(x)$ ,  $x \rightarrow \infty$ . For the other term, we have

$$\begin{aligned} II &\leq K \left\{ f_1(x) - f_1\left(\frac{x}{M^2}\right) \right\} \\ &= o(x), \quad (x \rightarrow \infty). \end{aligned}$$

This completes the proof that  $\#(\tilde{\lambda} \leq x) = o(x)$ ,  $x \rightarrow \infty$ . Hence  $n(\lambda) = o(\lambda)$ ,  $\lambda \rightarrow \infty$ . But  $N(\lambda) \sim A\lambda/4\pi$ . Hence,

$$\frac{N(\lambda)}{n(\lambda)} \sim \text{const.} \frac{\lambda}{o(\lambda)}, \quad \lambda \rightarrow \infty.$$

The proof is complete.

**Acknowledgment.** We are indebted to the referees for several helpful suggestions.

*Note added in proof.* Using the methods of this paper, P. Bérard has obtained explicit formulas for the eigenvalues of the Laplacian of certain Euclidean domains associated with crystallographic groups. (P. Bérard, *Spectres et groupes cristallographiques*, C.R. Acad. Sci. Paris Sér. A, 288 (25 juin, 1979), pp. 1059–1060).

#### REFERENCES

- [1] V. ARNOLD, *Modes and Quasimodes*, Functional Anal. Appl., 6 (1972), pp. 12–20.
- [2] M. BERGER, P. GAUDUCHON AND E. MAZET, *Le Spectre d'une Variété Riemannienne*, Springer Verlag Lecture Notes in Mathematics, vol. 194, 1971.
- [3] B. DRISCOLL, *Eigenvalues of a symmetric drumhead*, Ph.D. Thesis, Northwestern University, 1978.
- [4] P. GARABEDIAN, *Partial Differential Equations*, McGraw Hill, 1964.
- [5] GELFAND AND LINNIK, *Elementary Methods in the Theory of Numbers*, MIT Press, 1966.
- [6] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford, 1960.
- [7] M. G. LAMÉ, *Leçons sur le Théorie Mathématique de l'Elasticité des Corps Solides*, Paris, Bachelier, 1852.
- [8] E. LANDAU, *Vorlesungen über Zahlentheorie*, vol. I, Leipzig, Herzel, 1927.

## EXACT ERROR TERMS IN THE ASYMPTOTIC EXPANSION OF A CLASS OF INTEGRAL TRANSFORMS I (OSCILLATORY KERNELS)\*

KUSUM SONI†

**Abstract.** In this paper we show how the Parseval relation for the Mellin transform can be used to obtain an explicit expression for the remainder in the asymptotic expansion of a class of integral transforms. The technique, with some modification, can be used to derive similar results for many other integral transforms which are not discussed here.

### 1. Introduction. Let

$$(1.1) \quad F(x) = \int_0^\infty K(xt)f(t) dt$$

be the  $K$ -transform of  $f$ . In this paper, under certain conditions on  $K$  and  $f$ , we give the exact error terms in the asymptotic expansion of  $F(x)$  when  $x \rightarrow \infty$ . In particular, our results are applicable to the Fourier transform. As far as we know, the earliest results of this type are due to Olver [12]. He obtained explicit expressions for the remainder in the asymptotic expansion of the Fourier integrals when the expansion is terminated after a finite number of terms. Later, Wong [19], and McClure and Wong [11], gave similar results for the Hankel and the Stieltjes transforms respectively. The techniques used by Olver and Wong are different from the one used here. We use the Parseval relation for the Mellin transform. This method was used by Handelsman and Lew [5], [6], to obtain the asymptotic expansions for a very broad class of functions. More recently, it has been used by Bleistein [1] to extend the work of Handelsman and Lew. The basic idea is as follows: Suppose that  $M[K, s]$ ,  $s = \sigma + i\tau$ , is the Mellin transform of  $K(t)$  evaluated at  $s$  and  $M[f, 1-s]$  is the Mellin transform of  $f(t)$  evaluated at  $1-s$ . Assume further that  $M[K, s]$  and  $M[f, 1-s]$  are analytic in a strip containing the line  $\text{Re } s = c$ , and the Parseval relation

$$(1.2) \quad \int_0^\infty K(xt)f(t) dt = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s} M[K, s] M[f, 1-s] ds$$

holds. If  $M[K, s]$  and  $M[f, 1-s]$  can be analytically continued to a meromorphic function in the right half plane and the line of integration in (1.2) can be shifted from  $\text{Re } s = c$  to  $\text{Re } s = d > c$ , then by the residue theorem,

$$(1.3) \quad F(x) = - \sum_{c < \text{Re } s < d} \text{Res}\{x^{-s} M[K, s] M[f, 1-s]\} + E,$$

where

$$(1.4) \quad E = \frac{1}{2\pi i} \int_{d-i\infty}^{d+i\infty} x^{-s} M[K, s] M[f, 1-s] ds.$$

If the integral (1.4) converges absolutely, then (1.3) provides a finite asymptotic expansion of  $F(x)$  as  $x \rightarrow \infty$  with exact remainder  $E$ . In some cases, it may be possible to obtain bounds for the error term directly from (1.4) when the Mellin transforms of  $f(t)$  and  $K(t)$  are known; however, in general it will be necessary to represent  $E$  in terms of

\* Received by the editors May 13, 1976, and in final revised form December 27, 1979.

† Mathematics Department, University of Tennessee, Knoxville, Tennessee 37916.

the functions  $f(t)$  and  $K(t)$  as in [11], [12], [19]. Our object is to develop a technique which will provide such a representation for a special class of kernels. The advantage of the Mellin transform is that, in these cases, we can anticipate the form of the remainder relevant to the functions involved. We would like to add that even though the kernels considered in this paper are of a very special type, the technique, with some modification, can be used successfully in various other cases. One such case is discussed in [15].

We prove two theorems. In the first theorem, we give the remainder explicitly in terms of the functions which are directly related to  $f(t)$  and  $K(t)$ . It is assumed that  $f(t)$  and  $t^p f^{(p)}(t)$  are absolutely integrable. These conditions are imposed so that the Parseval relation (1.2) can be justified in a comparatively simple manner. In the second theorem, we use an integral analogue of Abel's theorem [8, p. 151], [12, Lemma 2] to show that the conclusion of the first theorem remains valid under weaker assumptions on  $f(t)$ . Although our results can be applied directly to functions  $f(t)$  when

$$f(t) \sim \sum_{k,l=0}^{\infty} a_{kl} t^{\lambda_k} (\log t)^l, \quad t \rightarrow 0+,$$

where  $\operatorname{Re} \lambda_k \uparrow \infty$  as  $k \rightarrow \infty$  and  $\{l: a_{kl} \neq 0\}$  is finite for each  $k$ , we assume that for each  $k$ ,  $a_{kl} = 0$  for  $l > 1$ .

**2. Notation and basic assumptions.** The functions  $f(t)$  and  $K(t)$  are complex valued and locally integrable in  $[0, \infty)$ .  $F(x)$  is the  $K$ -transform of  $f(t)$  and is defined by (1.1). The variable  $x$  is real.

The variable  $s$  is complex; the real and the imaginary parts of  $s$  are denoted by  $\sigma$  and  $\tau$  respectively. The Mellin transform of a function  $\phi(t)$  evaluated at  $s$  is

$$(2.1) \quad M[\phi, s] = \int_0^{\infty} \phi(t) t^{s-1} dt, \quad s = \sigma + i\tau,$$

whenever the integral converges. As is usual,  $M[\phi, s]$  also denotes the function which is an analytic continuation of the function element defined by (2.1) in the complex  $s$ -plane. An integral which converges but not absolutely, is assumed to converge in the Cauchy sense [18, p. 9].

*Conditions on  $f(t)$ .*

$$(2.2) \quad f(t) = f_{n-1}(t) + R_n(t),$$

where

$$(2.3) \quad f_{n-1}(t) = \sum_{k=0}^{n-1} (a_k + b_k \log t) t^{\lambda_k},$$

$$-1 < \operatorname{Re} \lambda_0 \leq \operatorname{Re} \lambda_1 \leq \dots \leq \operatorname{Re} \lambda_{n-1}, \quad |a_0| + |b_0| \neq 0,$$

and

$$(2.4) \quad R_n(t) = O(t^{\lambda_n}), \quad t \rightarrow 0+, \quad \operatorname{Re} \lambda_{n-1} < \operatorname{Re} \lambda_n.$$

*Conditions on  $K(t)$ .*

(i) There exists a positive number  $\sigma_0$  such that

$$(2.5) \quad M[K, s] = \int_0^{\infty} K(t) t^{s-1} dt$$

converges for  $0 < \sigma < \sigma_0$ .  $M[K, s]$  is, therefore, analytic in this strip.

(ii)  $M[K, s]$  can be analytically continued to a meromorphic function in the right half plane  $\sigma > 0$ . Furthermore, for some  $\delta > 0$ ,

$$(2.6) \quad M[K, \sigma + i\tau] = O(|\tau|^{\sigma-\delta}), \quad |\tau| \rightarrow \infty, \quad 0 < \sigma < \operatorname{Re} \lambda_n + 1.$$

Without loss of generality we can assume  $\delta \leq \sigma_0$ .

(iii) There exist numbers  $C_1, C_2$ , and  $\alpha, 0 \leq \alpha < \operatorname{Re} \lambda_0 + 1$  such that for all  $T \geq 0$  and  $0 < C < \sigma_0$ ,

$$(2.7) \quad \left| \int_{C-iT}^{C+iT} t^{-s} M[K, s] ds \right| \leq C_1 t^{-\alpha} + C_2,$$

and

$$(2.8) \quad \lim_{T \rightarrow \infty} \int_{C-iT}^{C+iT} t^{-s} M[K, s] ds = K(t), \quad 0 < t < \infty.$$

The conditions on  $K(t)$  are patterned after some of the well-known kernels used in applications, in particular the Fourier kernel  $e^{it}$ . The condition (2.6) is satisfied by a certain class of oscillatory kernels considered by Handelsman and Lew [5], [6]. As mentioned earlier,  $x$  is real in (1.1). We can allow complex values of  $x$  only if we strengthen the condition (2.6) considerably. Finally, in order to express the remainder explicitly in terms of  $f(t)$ , we assume that for a nonnegative integer  $p$  satisfying (3.1) there exists a function  $K(t, p)$  which is the inverse Mellin transform of  $\Gamma(s)M[K, s + p]/\Gamma(s + p)$ . In general, such a function may only exist as a generalized function (see [20, p. 108]). However, by the condition (iv) below,  $K(t, p)$  satisfies conditions similar to those imposed on  $K(t)$ .

(iv) For some numbers  $C_3, C_4$  and  $\alpha', 0 \leq \alpha' < \operatorname{Re} \lambda_n + 1 - p$ , where  $p$  satisfies (3.1),

$$(2.9) \quad \left| \int_{C-iT}^{C+iT} t^{-s} M[K, s + p] \Gamma(s) (\Gamma(s + p))^{-1} ds \right| \leq C_3 t^{-\alpha'} + C_4$$

for all  $T \geq 0, 0 < C < \delta$ ;

$$(2.10) \quad \lim_{T \rightarrow \infty} \int_{C-iT}^{C+iT} t^{-s} M[K, s + p] \Gamma(s) (\Gamma(s + p))^{-1} ds = K(t, p);$$

and, the Mellin transform of  $K(t, p)$  converges in  $0 < \sigma < \delta$ .

**3. Statement of results.**

**THEOREM 1.** *If  $f(t)$  satisfies (2.2)–(2.4) and*

(i)  $f^{(p)}(t)$  *is continuous in  $(0, \infty)$ , where*

$$(3.1) \quad \max(\operatorname{Re} \lambda_{n-1}, \operatorname{Re} \lambda_{n-1} + 1 - \delta) < p < \operatorname{Re} \lambda_n + 1,$$

(ii)  $f^{(p)}(t) = f_{n-1}^{(p)}(t) + O(t^{\lambda_n - p}), \quad t \rightarrow 0+,$

(iii)  $f(t)$  and  $t^p f^{(p)} \in L(0, \infty)$ ;

then

$$(3.2) \quad \int_0^\infty K(xt) f(t) dt = - \sum_{0 < \sigma < \sigma(n,p)} \operatorname{Res}\{x^{-s} M[K, s] M[f, 1 - s]\} + E(x),$$

where

$$(3.3) \quad \sigma(n, p) = \min(p + \delta, \operatorname{Re} \lambda_n + 1)$$

and

$$(3.4) \quad E(x) = x^{-p} \int_0^{\infty} K(xt, p) R_n^{(p)}(t) dt.$$

Note that  $\delta$  and  $K(xt, p)$  are defined in (2.6) and (2.10) respectively.

**THEOREM 2.** *If  $f(t)$  satisfies (2.2)–(2.4) and*

(i)' *Conditions (i) and (ii) of Theorem 1 are satisfied,*

(ii)'  *$t^k f^{(k)}(t) = o(t^{p-1})$ ,  $t \rightarrow \infty$ ,  $k = 0, 1, \dots, p-1$ , and for every  $a > 0$ ,*

$$f^{(p)}(t) = o(e^{at}), \quad t \rightarrow \infty,$$

(iii)' *the integrals*

$$\int_1^{\infty} K(xt) f(t) dt \quad \text{and} \quad \int_1^{\infty} K(xt, p) f^{(p)}(t) dt$$

converge,

(iv)'  *$M(K, s)$  has no singularity in  $0 < \sigma < \text{Re } \lambda_n + 1$ ,*

then (3.2) holds provided that the finite sum on the right is replaced by

$$-\sum_{k=0}^{n-1} \text{Res} \left\{ x^{-s} M[K, s] \left( \frac{a_k}{\lambda_k + 1 - s} - \frac{b_k}{(\lambda_k + 1 - s)^2} \right) \right\}.$$

Condition (i) of Theorem 1 can be weakened slightly. It is not necessary that  $f^{(p)}(t)$  be continuous in  $(0, \infty)$ ; it is enough to assume that  $f^{(p)}(t)$  exists for all  $t$  in  $(0, \infty)$  and is bounded [17, (11.81)]. Again, it is quite possible that in certain cases Condition (3.1),  $\max(\text{Re } \lambda_{n-1}, \text{Re } \lambda_{n-1} + 1 - \delta) < p$ , can be replaced by some weaker condition, (see [14]). We need the condition as stated in the proof of Theorem 1, when we shift the line of integration to the right in the complex  $s$ -plane.

Admittedly, for a given  $n$  there may not be any nonnegative integer  $p$  which satisfies (3.1). In general, this does not present any problem because if

$$f(t) \sim \sum_{n=0}^{\infty} (a_n + b_n \log t) t^{\lambda_n}, \quad t \rightarrow 0+, \quad \text{Re } \lambda_n \uparrow \infty,$$

then for each  $p$ , we can find  $n$  and a positive number  $\varepsilon$  such that

$$\begin{aligned} \text{Re } \lambda_{n-1} &< \text{Re } \lambda'_n, & (\lambda'_n &= \lambda_n - \varepsilon), \\ R_n(t) &= O(t^{\lambda'_n}), & t &\rightarrow 0+, \end{aligned}$$

and (3.1) is satisfied when  $\lambda_n$  is replaced by  $\lambda'_n$ .

Finally, in Theorem 2 we assume that  $M[K, s]$  has no singularities in  $0 < \text{Re } s < \text{Re } \lambda_n + 1$ . This restriction can easily be removed and we can allow  $M[K, s]$  to have poles provided that we make an additional assumption on  $f$ , namely that

$$\lim_{\nu \rightarrow 0} \int_1^{\infty} e^{-\nu t} f(t) t^{-s} dt$$

exists and is meromorphic in  $0 < \sigma < \text{Re } \lambda_n + 1$ . Under the conditions of Theorem 2,  $M[f, 1-s]$  need not exist. However, by the above condition, the contribution from the poles of  $M[K, s]$  to the finite sum in (3.2) would be well defined.

**4. Some preliminary results.** In this section, we investigate the behavior of  $M[f, 1 - s]$ . If  $f(t) \in L(0, \infty)$  and satisfies (2.2)–(2.4), then

$$(4.1) \quad M[f, 1 - s] = \int_0^\infty f(t)t^{-s} dt,$$

converges absolutely in  $0 < \sigma < \text{Re } \lambda_0 + 1$  and represents an analytic function in this region. Furthermore, it is known [5], [6] that  $M[f, 1 - s]$  can be analytically continued into  $0 < \sigma < \text{Re } \lambda_n + 1$ , and its only singularities in this strip are poles, of order two at most, at  $s = \lambda_k + 1$ ,  $k = 0, 1, \dots, (n - 1)$ . For our purpose, this information is not enough. In the first place, we want to know the function represented by  $M[f, 1 - s]$  outside the strip of convergence of the integral (4.1). (For  $b_k = 0$ ,  $k = 0, 1, \dots, (n - 1)$ , the result is given in [16] and the extension to the present case is straightforward). We state this result in Lemma 1 and briefly indicate the proof. Secondly, we want to know the behavior of  $M[f, 1 - \sigma - i\tau]$  as  $|\tau| \rightarrow \infty$  in  $0 < \sigma < \text{Re } \lambda_n + 1$ , so that we may also be able to shift the line of integration in (1.2). This behavior is given in Lemma 2. (Similar results, given by Bleistein and Handelsman [2, pp. 226–229] are not applicable to the functions under consideration).

**LEMMA 1.** *If  $f(t) \in L(0, \infty)$  and satisfies (2.2)–(2.4), then  $M[f, 1 - s]$  is analytic in  $0 < \sigma < \text{Re } \lambda_n + 1$ ; its only singularities are poles, of order two at most, at  $s = \lambda_k + 1$ ,  $k = 0, 1, \dots, (n - 1)$ ; and in the strip  $\text{Re } \lambda_{n-1} + 1 < \sigma < \text{Re } \lambda_n + 1$ ,*

$$(4.2) \quad M[f, 1 - s] = \int_0^\infty R_n(t)t^{-s} dt.$$

*Proof.* Although we need to prove only (4.2), the other properties also follow in the process. The integral (4.1) converges absolutely in  $0 < \sigma < \text{Re } \lambda_0 + 1$ . By using (2.3),

$$(4.3) \quad \begin{aligned} \int_0^\infty t^{-s}f(t) dt &= \int_0^1 t^{-s}(f(t) - f_{n-1}(t)) dt \\ &+ \int_1^\infty t^{-s}f(t) dt \\ &+ \sum_{k=0}^{n-1} \left( \frac{a_k}{\lambda_k + 1 - s} - \frac{b_k}{(\lambda_k + 1 - s)^2} \right). \end{aligned}$$

The finite sum in (4.3) is obtained by integrating  $t^{-s}f_{n-1}(t)$  term by term in  $(0, 1)$ . The right side of (4.3) is meromorphic in  $0 < \sigma < \text{Re } \lambda_n + 1$  because, by (2.4), the first integral converges absolutely in  $\sigma < \text{Re } \lambda_n + 1$ . Thus (4.3) provides the analytic continuation of  $M[f, 1 - s]$  into the region  $0 < \sigma < \text{Re } \lambda_n + 1$ . The conclusion (4.2) follows from the fact that in the strip  $\text{Re } \lambda_{n-1} + 1 < \sigma < \text{Re } \lambda_n + 1$ ,

$$\sum_{k=0}^{n-1} \left( \frac{a_k}{\lambda_k + 1 - s} - \frac{b_k}{(\lambda_k + 1 - s)^2} \right) = - \int_1^\infty t^{-s}f_{n-1}(t) dt.$$

**LEMMA 2.** *If  $f(t)$  satisfies the conditions of Theorem 1, then*

$$(4.4) \quad M[f, 1 - s] = O(|\tau|^{-p}), \quad |\tau| \rightarrow \infty, \quad 0 \leq \sigma < \text{Re } \lambda_n + 1.$$

*Proof.* Since  $f(t)$  and  $t^p f^{(p)}(t) \in L(0, \infty)$ , (see Condition (iii)), it follows that

$$f(t) = \frac{(-1)^p}{\Gamma(p)} \int_t^\infty (u - t)^{p-1} f^{(p)}(u) du.$$

Therefore, for  $0 \leq \sigma < \min(\operatorname{Re} \lambda_0 + 1, 1)$ ,

$$\begin{aligned}
 M[f, 1-s] &= \frac{(-1)^p}{\Gamma(p)} \int_0^\infty f^{(p)}(u) \left( \int_0^u (u-t)^{p-1} t^{-s} dt \right) du \\
 (4.5) \qquad &= \frac{(-1)^p \Gamma(1-s)}{\Gamma(p+1-s)} \int_0^\infty f^{(p)}(u) u^{p-s} du.
 \end{aligned}$$

The interchange of the order of integration in (4.5) is justified by the absolute convergence of the double integral. Since the last integral converges absolutely, (4.4) holds in  $0 \leq \sigma < \operatorname{Re} \lambda_0 + 1$ . We have to show that this estimate holds in the larger strip  $0 \leq \sigma < \operatorname{Re} \lambda_n + 1$ .

$$\begin{aligned}
 \int_0^\infty f^{(p)}(u) u^{p-s} du &= \int_0^1 (f(u) - f_{n-1}(u))^{(p)} u^{p-s} du \\
 (4.6) \qquad &+ \int_1^\infty f^{(p)}(u) u^{p-s} du \\
 &+ \int_0^1 f_{n-1}^{(p)}(u) u^{p-s} du.
 \end{aligned}$$

The first integral on the right converges absolutely in  $\sigma < \operatorname{Re} \lambda_n + 1$  and the second integral converges absolutely in  $\sigma \geq 0$ . The third integral converges absolutely only in  $\sigma < \operatorname{Re} \lambda_0 + 1$ . But by using (2.3) we see that

$$(4.7) \qquad f_{n-1}^{(p)}(u) = \sum_{k=0}^{n-1} (\alpha_k + \beta_k \log u) u^{\lambda_k - p},$$

for some constants  $\alpha_k, \beta_k$  so that

$$(4.8) \qquad \int_0^1 u^p f_{n-1}^{(p)}(u) u^{-s} du = \sum_{k=0}^{n-1} \left( \frac{\alpha_k}{\lambda_k + 1 - s} - \frac{\beta_k}{(\lambda_k + 1 - s)^2} \right).$$

Therefore, by (4.6),  $M[u^p f_{n-1}^{(p)}(u), 1-s]$  is bounded as  $|\tau| \rightarrow \infty$  in  $0 \leq \sigma < \operatorname{Re} \lambda_n + 1$ , (not necessarily uniformly as  $\sigma \rightarrow \operatorname{Re} \lambda_n + 1$ ). This proves Lemma 2.

**5. Proof of Theorem 1.** We note that the Parseval relation (1.2) holds. To show this, let  $0 < c < \min(\sigma_0, \operatorname{Re} \lambda_0 + 1)$ . Then

$$(5.1) \qquad \frac{1}{2\pi i} \int_{c-iT}^{c+iT} x^{-s} M[K, s] M[f, 1-s] ds = \int_0^\infty f(t) \left( \frac{1}{2\pi i} \int_{c-iT}^{c+iT} (tx)^{-s} M[K, s] ds \right) dt.$$

The interchange of the order of integration is justified by the absolute convergence of the double integral. Now let  $T \rightarrow \infty$ . By (2.8) and the Lebesgue dominated convergence theorem, we obtain (1.2):

$$(5.2) \qquad \int_0^\infty K(xt) f(t) dt = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s} M[K, s] M[f, 1-s] ds.$$

Now shift the line of integration in the above integral from  $\sigma = c$  to  $\sigma = c'$ , where

$$(5.3) \qquad \max(p, \operatorname{Re} \lambda_{n-1} + 1) < c' < \min(p + \delta, \operatorname{Re} \lambda_n + 1).$$

Such a choice of  $c'$  is possible by (3.1). Note that  $M[K, s]M[f, 1-s]$  has no singularities in the strip  $\max(p, \operatorname{Re} \lambda_{n-1} + 1) < \sigma < \min(p + \delta, \operatorname{Re} \lambda_n + 1)$ . This follows from the observation that  $M[f, 1-s]$  has no singularities in the strip  $\operatorname{Re} \lambda_{n-1} + 1 < \sigma <$

Re  $\lambda_n + 1$ . Furthermore, the Mellin transform of  $K(t, p)$  converges in  $(0, \delta)$  so that by (2.10),  $M[K, s]$  has no singularities in the strip  $p < \sigma < p + \delta$ . By (2.6) and Lemma 2,

$$(5.4) \quad |x^{-s}M[K, s]M[f, 1-s]| = O(|\tau|^{\sigma-p-\delta}), \quad |\tau| \rightarrow \infty,$$

in  $c \leq \sigma \leq c'$ . Therefore, the integrand on the right side in (5.2) tends to zero as  $|\tau| \rightarrow \infty$  in this strip. By the residue theorem,

$$(5.5) \quad \int_0^\infty K(xt)f(t) dt = - \sum_{0 < \sigma < c'} \text{Res}\{s^{-s}M[K, s]M[f, 1-s]\} + E(x),$$

where

$$(5.6) \quad E(x) = \frac{1}{2\pi i} \int_{c'-i\infty}^{c'+i\infty} x^{-s}M[K, s]M[f, 1-s] ds.$$

We show that  $E(x)$  satisfies (3.4). Note that by Lemma 1,

$$(5.7) \quad E(x) = \frac{1}{2\pi i} \int_{c'-i\infty}^{c'+i\infty} x^{-s}M[K, s] \left( \int_0^\infty t^{-s}R_n(t) dt \right) ds.$$

Consider the inner integral in (5.7). If we apply integration by parts  $p$  times, use (5.3) together with (2.4) as  $t \rightarrow 0+$ , and use  $R_n(t) = O(|f(t)| + |t^{\lambda_{n-1}} \log t|)$  as  $t \rightarrow \infty$ , we obtain

$$(5.8) \quad \int_0^\infty t^{-s}R_n(t) dt = \frac{\Gamma(s-p)}{\Gamma(s)} \int_0^\infty t^{-s+p}R_n^{(p)}(t) dt.$$

Therefore,

$$(5.9) \quad E(x) = \frac{x^{-p}}{2\pi i} \int_{c'-p-i\infty}^{c'-p+i\infty} x^{-w}M[K, w+p] \frac{\Gamma(w)}{\Gamma(w+p)} M[R_n^{(p)}, 1-w] dw,$$

where  $w = s - p$ . Now let

$$\begin{aligned} R_n^{(p)}(t) &= g_1(t) + g_2(t) - g_3(t), \\ g_1(t) &= \begin{cases} R_n^{(p)}(t), & 0 < t < 1, \\ 0, & t \geq 1, \end{cases} \\ g_2(t) &= \begin{cases} 0, & 0 < t < 1, \\ f^{(p)}(t), & t \geq 1, \end{cases} \\ g_3(t) &= \begin{cases} 0, & 0 < t < 1, \\ f_{n-1}^{(p)}(t), & t \geq 1. \end{cases} \end{aligned}$$

By (5.3),  $0 < c' - p < \text{Min}(\delta, \text{Re } \lambda_n + 1 - p)$ . Since  $g_1(t)$  and  $g_1(t)t^{-\alpha'}$  (see Condition (iv) on  $K(t)$  in § 2) as well as  $g_2(t)$  and  $g_2(t)t^{-\alpha'}$  are absolutely integrable in  $(0, \infty)$ , we may use the technique used in the proof of (5.2) to show that, for  $j = 1, 2$ ,

$$(5.10) \quad \int_0^\infty K(xt, p)g_j(t) dt = \frac{1}{2\pi i} \int_{c'-p-i\infty}^{c'-p+i\infty} x^{-w}M[K, w+p] \frac{\Gamma(w)}{\Gamma(p+w)} M[g_j, 1-w] dw.$$

The Parseval relation (5.10) holds for  $g_3(t)$  also. To show this we use a result of Titchmarsh [18, Thm. 43]. The Mellin transform of  $K(xt, p)$  converges uniformly in every closed strip in the interior of  $0 < \text{Re } w < \delta$ , and the Mellin transform  $M[g_3, 1-w]$  converges uniformly in every closed half plane included in  $\text{Re } w > \text{Re } \lambda_{n-1} + 1 - p$ . Since  $\text{Re } \lambda_{n-1} + 1 - p < \delta$ , we can choose  $d$  such that  $\text{Re } w = d$  is in the common strip of uniform convergence. By (4.7),  $M[g_3, 1-w] = O(|\tau|^{-1})$  as  $|\tau| \rightarrow \infty$ . Therefore, the right



side of the integral (5.10) converges absolutely for  $g_j(t) = g_3(t)$ . If necessary, we can move the line of integration in (5.10) to  $\text{Re } w = d$ . The integral along this line converges. Since the Mellin transform of  $K(t, p)$  converges in  $0 < \sigma < \delta$ , we see that the integral

$$(5.11) \quad I_1(x) = \int_0^\infty K(xt, p)g_3(\xi t) dt$$

is uniformly bounded in the neighborhood of  $\xi = 1$  for each fixed  $x > 0$ . Therefore, by [18, Thm. 43], (5.10) holds for  $j = 3$ . By (5.9) and (5.10),  $j = 1, 2, 3$ ,

$$(5.12) \quad E(x) = x^{-p} \int_0^\infty K(xt, p)R_n^{(p)}(t) dt.$$

This completes the proof.

## 6. Proof of Theorem 2.

Let

$$(6.1) \quad h(t) = e^{-\nu t}f(t), \quad \nu > 0.$$

By using the power series expansion of  $e^{-\nu t}$ , we may write

$$(6.2) \quad h(t) = h_{m-1}(t) + H_m(t), \quad t \rightarrow 0+,$$

where

$$(6.3) \quad h_{m-1}(t) = \sum_{k=0}^{m-1} (c_k + d_k \log t)t^{\mu_k}$$

and

$$(6.4) \quad H_m(t) = O(t^{\mu_m}), \quad -1 < \text{Re } \mu_0 \leq \dots \leq \text{Re } \mu_{m-1} < \text{Re } \mu_m;$$

where the positive integer  $m$  is chosen so that

$$(6.5) \quad \max(\text{Re } \mu_{m-1}, \text{Re } \mu_{m-1} + 1 - \delta) < p < \text{Re } \mu_m + 1.$$

Clearly,  $\text{Re } \lambda_{n-1} \leq \text{Re } \mu_{m-1}$  and  $\text{Re } \mu_m \leq \text{Re } \lambda_n$ . By Condition (ii)',  $h(t)$  and  $t^p h^{(p)}(t) \in L(0, \infty)$ . If  $\alpha' < \text{Re } \mu_m + 1 - p$ , Condition (iv) on  $K(t)$  is satisfied, and by Theorem 1 we obtain

$$(6.6) \quad \int_0^\infty K(xt)h(t) dt = - \sum_{k=0}^{m-1} \text{Res} \left\{ x^{-s} M[K, s] \left( \frac{c_k}{\mu_k + 1 - s} - \frac{d_k}{(\mu_k + 1 - s)^2} \right) \right\} \\ + x^{-p} \int_0^\infty K(xt, p)H_m^{(p)}(t) dt.$$

Note that  $c_k$  and  $d_k$  tend to a limit as  $\nu \rightarrow 0+$ . If  $\mu_k \neq \lambda_j$  for any  $j, j = 0, 1, \dots, n-1$ , then  $c_k$  and  $d_k$  tend to zero, and if  $\mu_k = \lambda_j$  for some  $j$ , then  $c_k$  and  $d_k$  tend to  $a_j$  and  $b_j$  respectively. Now we use the integral analogue of Abel's limit theorem [8, Thm. 87]. As  $\nu \rightarrow 0+$ , by Condition (iii)',

$$(6.7) \quad \int_0^\infty K(xt)f(t) dt = - \sum_{k=0}^{n-1} \text{Res} \left\{ x^{-s} M[K, s] \left( \frac{a_k}{\lambda_k + 1 - s} - \frac{b_k}{(\lambda_k + 1 - s)^2} \right) \right\} \\ + x^{-p} \lim_{\nu \rightarrow 0+} \int_0^\infty K(xt, p)H_m^{(p)}(t) dt.$$

To complete the proof, we must show that the last integral approaches  $E(x)$  as  $\nu \rightarrow 0+$ .

By (2.9),  $K(t, p) = O(t^{-\alpha'})$ ,  $t \rightarrow 0+$ . Therefore, by the Lebesgue dominated convergence theorem,

$$(6.8) \quad \lim_{\nu \rightarrow 0+} \int_0^1 K(xt, p) H_m^{(p)}(t) dt = \int_0^1 K(xt, p) R_n^{(p)}(t) dt.$$

Also,

$$(6.9) \quad \lim_{\nu \rightarrow 0+} \int_1^\infty K(xt, p) h_{m-1}^{(p)}(t) dt = \int_1^\infty K(xt, p) f_{n-1}^{(p)}(t) dt.$$

Here we have used the fact that the Mellin transform of  $K(xt, p)$  converges in  $0 < \text{Re } s < \delta$  so that we can integrate each term on the left and then take the limit as  $\nu \rightarrow 0+$ . Therefore, we only need to prove that

$$(6.10) \quad \lim_{\nu \rightarrow 0+} \int_1^\infty K(xt, p) h^{(p)}(t) dt = \int_1^\infty K(xt, p) f^{(p)}(t) dt,$$

where

$$(6.11) \quad h^{(p)}(t) = \sum_{k=0}^p \binom{p}{k} (-\nu)^{p-k} e^{-\nu t} f^{(k)}(t).$$

By Condition (iii)',

$$(6.12) \quad \lim_{\nu \rightarrow 0+} \int_1^\infty e^{-\nu t} K(xt, p) f^{(p)}(t) dt = \int_1^\infty K(xt, p) f^{(p)}(t) dt.$$

Let  $\varepsilon > 0$ . By Condition (ii)', choose  $N$  such that for  $k = 0, 1, \dots, p-1$ ,

$$(6.13) \quad |f^{(k)}(t)| < \varepsilon t^{p-1-k}, \quad t \geq N.$$

By (2.9),  $K(xt, p)$  is bounded in  $[1, \infty)$ . Therefore, for  $k = 0, 1, \dots, p-1$ ,

$$\lim_{\nu \rightarrow 0+} (\nu)^{p-k} \int_1^N K(xt, p) e^{-\nu t} f^{(k)}(t) dt = 0,$$

and by (6.13),

$$\nu^{p-k} \int_N^\infty e^{-\nu t} |f^{(k)}(t)| dt < \varepsilon \nu^{p-k} \int_N^\infty e^{-\nu t} t^{p-1-k} dt < \varepsilon \Gamma(p-k).$$

Therefore,

$$\lim_{\nu \rightarrow 0+} \sum_{k=0}^{p-1} \binom{p}{k} (-\nu)^{p-k} \int_1^\infty e^{-\nu t} K(xt, p) f^{(k)}(t) dt = 0.$$

This proves (6.10) when  $\alpha' < \text{Re } \mu_m + 1 - p$ . If  $\text{Re } \mu_m + 1 - p \leq \alpha'$ , the proof of (5.10) for the function  $g_1(t)$  where

$$g_1(t) = \begin{cases} H_m^{(p)}(t), & 0 < t < 1, \\ 0, & 1 < t < \infty, \end{cases}$$

requires some modification. Let  $m'$  denote the positive integer greater than  $m$ , such that  $\text{Re } \mu_{m'-1} < \text{Re } \mu_{m'} = \text{Re } \lambda_n$ , and define

$$\xi(t) = \begin{cases} h_{m'-1}^{(p)}(t) - h_{m-1}^{(p)}(t), & 0 < t < 1, \\ 0, & t > 1. \end{cases}$$

Since  $0 < c' - p < \min(\delta, \operatorname{Re} \mu_m + 1 - p)$ , the function  $\xi(t)$  satisfies (5.10). Furthermore,  $g_1(t) - \xi(t)$  satisfies (5.10) because  $g_1(t) - \xi(t) = O(t^{\lambda_n} \log t)$ ,  $t \rightarrow 0+$ . Therefore  $g_1(t)$  also satisfies (5.10). This justifies (6.6). Since,

$$\lim_{\nu \rightarrow 0} \int_0^1 K(xt, p) \xi(t) dt = 0,$$

$H_m^{(p)}(t)$  in (6.8) can be replaced by  $H_m^{(p)'}(t)$  and the conclusion follows as above.

**7. Asymptotic nature of the expansion.** We prove the following:

If in Theorem 2, the integral

$$(7.1) \quad \mathcal{J}_1(x) = \int_1^\infty K(xt, p) f^{(p)}(t) dt$$

converges uniformly for  $x \geq X$  for some  $X$ , then the remainder  $E(x) = o(x^{-p})$ ,  $x \rightarrow \infty$ .

*Proof.* We may assume  $X > 1$ . Since the Mellin transform of  $K(xt, p)$  converges in  $0 < \operatorname{Re} s < \delta$ , by using (4.7) it follows that

$$(7.2) \quad \mathcal{J}_2(x) = \int_1^\infty K(xt, p) f_{n-1}^{(p)}(t) dt$$

converges uniformly in  $x \geq X$ . Let  $\varepsilon > 0$ . By the uniform convergence of (7.1) and (7.2), we can find  $N > 1$  such that

$$(7.3) \quad \left| \int_N^\infty K(xt, p) R_n^{(p)}(t) dt \right| < \varepsilon \quad \text{for } x \geq X.$$

Also, by (2.9),

$$(7.4) \quad \int_0^{1/x} K(xt, p) R_n^{(p)}(t) dt = O\left( \left| x^{-\alpha'} \int_0^{1/x} t^{-\alpha'} t^{\lambda_n - p} dt \right| \right) \\ = O(|x^{-(\lambda_n + 1 - p)}|), \quad x \rightarrow \infty.$$

Now define

$$(7.5) \quad K^*(t) = \begin{cases} K(t, p), & t \geq 1, \\ 0, & 0 < t < 1, \end{cases}$$

and

$$(7.6) \quad R(t) = \begin{cases} 0, & t \geq N, \\ R_n^{(p)}(t), & 0 < t \leq N. \end{cases}$$

$R(t) \in L(0, \infty)$ .  $K^*(t)$  is bounded, and since its Mellin transform converges for  $\operatorname{Re} s < \delta$ ,

$$(7.7) \quad \int_0^x K^*(t) dt = o(x), \quad x \rightarrow \infty.$$

Therefore, by a known result [9, Thm. 2.1.2],

$$(7.8) \quad \int_{1/x}^N K(xt, p) R_n^{(p)}(t) dt = \int_0^\infty K^*(xt) R(t) dt \\ = o(1), \quad x \rightarrow \infty.$$

By combining (7.3), (7.4), and (7.8),  $E(x) = o(x^{-p})$ ,  $x \rightarrow \infty$ .

**8. Applications.**

*Example 1.* Let

$$K(t) = e^{it}, \quad 0 < t < \infty.$$

The Mellin transform of  $e^{it}$  converges in  $0 < \sigma < 1$  and

$$M[K, s] = e^{is\pi/2}\Gamma(s).$$

By [17, p. 151],

$$M[K, s] = O(|\tau|^{\sigma-1/2}), \quad |\tau| \rightarrow \infty.$$

It is known that the integral (2.7) is uniformly bounded. (For  $t \geq 1$ , see [7, Lemma 5]; for  $0 < t < 1$ , shift the line of integration to  $\sigma = -\frac{3}{4}$ ). Also (2.8) is satisfied:

$$M[K(t, p), s] = e^{ip\pi/2} e^{is\pi/2}\Gamma(s),$$

and

$$K(t, p) = e^{ip\pi/2} e^{it}.$$

Clearly,  $K(t, p)$  satisfies (2.9) with  $\alpha' = 0$ , and (2.10). Therefore, if  $f(t)$  satisfies the conditions of Theorem 2, we obtain

$$(8.1) \quad \int_0^\infty e^{ixt}f(t) dt = \sum_{k=0}^{n-1} x^{-(\lambda_k+1)} e^{i(\lambda_k+1)\pi/2}\Gamma(\lambda_k+1) \cdot [a_k - b_k \log x + b_k\psi(\lambda_k+1) + ib_k\pi/2] + E(x),$$

where  $\psi$  is the logarithmic derivative of the gamma function and

$$(8.2) \quad E(x) = x^{-p} e^{ip\pi/2} \int_0^\infty e^{ixt}R_n^{(p)}(t) dt.$$

Conditions (ii)' of Theorem 2 are weaker than those given in [12, p. 20, (iv)]. However, if we assume that the integrals in Condition (iii)' converge uniformly for large  $x$ , so that the expansion in (8.1) is asymptotic as  $x \rightarrow \infty$ , our conditions are equivalent to those given in [12]. From (8.2),

$$|E(x)| \leq x^{-p} \int_0^\infty |R_n^{(p)}(t)| dt.$$

This provides an error bound when  $R_n^{(p)}(t)$  is absolutely integrable in  $(0, \infty)$ . If  $R_n^{(p)}(t)$  is of bounded variation in  $[0, \infty)$ , the convergence of the integral in (8.2) implies that  $R_n^{(p)}(t) \rightarrow 0$  as  $t \rightarrow \infty$ . We can write  $R_n^{(p)}(t)$  in (8.2) as the difference of two monotone functions decreasing to zero, and apply the second mean value theorem to obtain

$$|E(x)| \leq 2\mathcal{V}x^{-p-1}$$

where  $\mathcal{V}$  is the total variation of  $R_n^{(p)}(t)$  in  $[0, \infty)$ . For this estimate we do not require that  $R_n^{(p)}(t)$  be absolutely integrable in  $(0, \infty)$ . We must note, however, that in general  $R_n^{(p)}(t)$  may be neither absolutely integrable nor of bounded variation in  $[0, \infty)$ . In such a case, we can obtain an error bound by estimating the integral in (8.2) in the intervals  $(0, c)$  and  $(c, \infty)$  separately, for some appropriate  $c > 0$ . To illustrate this, we consider a function which satisfies the conditions of Theorem 2 but not of Theorem 1.

Let  $f(t) = Y_0(t)$  where  $Y_0$  is the Bessel function of the second kind and of order zero. The Fourier transform of  $Y_0(t)$  is known. By [3, pp. 47, 103],

$$\int_0^\infty e^{ixt} Y_0(t) dt = \frac{1}{\sqrt{x^2-1}} \{-1 + i2\pi^{-1} \log [x - \sqrt{x^2-1}]\}, \quad x > 1.$$

We use (8.1) to obtain the asymptotic expansion of the above transform with explicit remainder term when  $n = p = 4$ . By [4, p. 8, (33)],

$$(8.3) \quad \begin{aligned} Y_0(t) &= (2/\pi)(\gamma - \log 2 + \log t) \\ &\quad - (2\pi)^{-1}(\gamma - 1 - \log 2 + \log t)t^2 \\ &\quad + O(t^4 \log t), \quad t \rightarrow 0+. \end{aligned}$$

Let

$$\lambda_k = \begin{cases} k, & 0 \leq k \leq 3, \\ 4 - \varepsilon_1, & k = 4. \end{cases}$$

$\varepsilon_1$  is positive but can be chosen arbitrarily small.

$$(8.4) \quad R_4^{(iv)}(t) = Y_0^{(iv)}(t) - (\pi t^2)^{-1} + 12(\pi t^4)^{-1}.$$

To compute the successive derivatives of  $Y_0$  we use some well known relations for the Bessel functions of the second kind [4, pp. 11-12, (54)-(57)] and obtain

$$(8.5) \quad R_4^{(iv)}(t) = (3t^{-2} - 1)Y_2(t) - (\pi t^2)^{-1} + 12(\pi t^4)^{-1}.$$

Therefore, we have the following:

$$(8.6) \quad \begin{aligned} \int_0^\infty e^{ixt} Y_0(t) dt &= -x^{-1}(1 + 2i\pi^{-1} \log 2x) \\ &\quad - x^{-3}(\frac{1}{2} - \frac{1}{2}i\pi^{-1} + i\pi^{-1} \log 2x) + E(x), \end{aligned}$$

where

$$(8.7) \quad E(x) = x^{-4} \int_0^\infty e^{ixt} \{(3t^{-2} - 1)Y_2(t) - \pi^{-1}(t^{-2} - 12t^{-4})\} dt.$$

Note that  $R_4^{(iv)}(t)$  is neither absolutely integrable nor of bounded variation in  $[0, \infty)$ . To obtain an error bound, we write (8.7) as follows:

$$\begin{aligned} x^4 E(x) &= \int_0^2 e^{ixt} R_4^{(iv)}(t) dt \\ &\quad + \int_2^\infty e^{ixt} \{3t^{-2} Y_2(t) - \pi^{-1}(t^{-2} - 12t^{-4})\} dt \\ &\quad - \int_2^\infty e^{ixt} Y_2(t) dt \\ &= J_1 + J_2 + J_3. \end{aligned}$$

The integrals  $J_1$  and  $J_2$  converge absolutely.  $J_3$  does not converge absolutely, but it converges uniformly in  $x \geq c > 1$ , and an upper bound for it can be obtained by using the asymptotic expansion of  $Y_2(t)$ , or by integrating by parts.

In the following example we give only  $E(x)$ . The expansion (3.2) can be obtained by using (6.7).

Example 2. Let

$$(8.8) \quad K(t) = e^{-it^2/4} D_{-\nu}(e^{i\pi/4} t), \quad 0 < t < \infty,$$

where  $D_{-\nu}$  is the parabolic cylinder function and  $\text{Re } \nu \geq 0$ . By [4, p. 122],

$$(8.9) \quad K(t) \sim e^{-i\nu\pi/4} t^{-\nu} e^{-it^2/2}, \quad t \rightarrow \infty.$$

The Mellin transform of  $K(t)$  converges in  $0 < \sigma < \text{Re } \nu + 2$ , and by [3, p. 336],

$$(8.10) \quad M[K, s] = \pi^{1/2} 2^{-(s+\nu)/2} e^{-is\pi/4} \Gamma(s) [\Gamma((s + \nu + 1)/2)]^{-1}.$$

By [17, p. 151],

$$(8.11) \quad \begin{aligned} M[K, s] &= O(|\tau|^{(\sigma-1-\text{Re } \nu)/2}), \quad |\tau| \rightarrow \infty. \\ M[K(t, p), s] &= e^{-ip\pi/4} \pi^{1/2} 2^{-(s+\nu+p)/2} e^{-is\pi/4} \Gamma(s) [\Gamma((s + \nu + p + 1)/2)]^{-1}. \end{aligned}$$

By the uniqueness of the inverse Mellin transform and (8.10),

$$K(t, p) = e^{-ip\pi/4} e^{-it^2/4} D_{-\nu-p}(e^{i\pi/4} t).$$

$D_{-\nu}(z)$  is an entire function of  $z$ . We can use (8.9)–(8.11) to verify that  $K(t)$  and  $K(t, p)$  satisfy all the conditions. Therefore, Theorem 2 holds, and the remainder in (3.2) is

$$(8.12) \quad E(x) = (x e^{i\pi/4})^{-p} \int_0^\infty e^{-ix^2 t^2/4} D_{-\nu-p}(e^{i\pi/4} xt) R_n^{(p)}(t) dt.$$

We mention two particular cases. If  $\nu = 0$ ,  $K(t) = \exp(-it^2/2)$ , (see [10, p. 326]). In this case, (8.12) provides an alternative form of the remainder for the integrals of the type considered in Example 1. If  $\nu = \frac{1}{2}$ , by using a known result [10, p. 326],

$$K(t) = (2\pi)^{-1/2} e^{i\pi/8} t^{1/2} e^{-it^2/4} K_{1/4}(it^2/4).$$

In this case, the remainder  $E(x)$  can again be expressed in terms of the modified Bessel functions, but the form (8.12) in terms of the parabolic cylinder functions is more convenient for the computation of the error bound. To estimate the error when  $\nu$  is real, we use the inequality

$$|D_{-\nu}(e^{i\pi/4} t)| \leq \pi^{1/2} 2^{-\nu/2} (\Gamma[(1 + \nu)/2])^{-1}, \quad 0 < t < \infty, \quad \nu > 0,$$

which follows from the integral representation given in [4, p. 119, (1)]. If  $R_n^{(p)}(t)$  is absolutely integrable in  $(0, \infty)$ ,  $\nu \geq 0$ ,  $p > 0$ , we obtain

$$|E(x)| \leq x^{-p} \pi^{1/2} 2^{-(\nu+p)/2} (\Gamma[(1 + \nu + p)/2])^{-1} \int_0^\infty |R_n^{(p)}(t)| dt.$$

In particular, this estimate is valid when  $f(t)$  satisfies the conditions of Theorem 1 and  $\text{Re } \lambda_{n-1} - p + 1 < 0$ .

Finally, we give an example to indicate why in certain cases some modification in the form of the remainder is desirable and suggest what can be done to achieve this.

Example 3. Let

$$(8.13) \quad K(t) = J_\nu(t), \quad 0 < t < \infty,$$

where  $J_\nu$  is the Bessel function of the first kind. If  $\text{Re } \nu \geq 0$ , the Mellin transform of  $K(t)$  converges in  $0 < \sigma < \frac{3}{2}$ . (The condition on  $\nu$  can be relaxed). By [3, p. 326, (1)],

$$(8.14) \quad M[K, s] = 2^{s-1} \Gamma((s + \nu)/2) [\Gamma((\nu - s + 2)/2)]^{-1},$$

and by [17, p. 151],

$$(8.15) \quad M[K, s] = O(|\tau|^{\sigma-1}), \quad |\tau| \rightarrow \infty,$$

$$M[K(t, p), s] = \frac{2^{s+p-1} \Gamma(s) \Gamma((s+p+\nu)/2)}{\Gamma(s+p) \Gamma((\nu+s-p+2)/2)}.$$

The condition (2.9) is satisfied with  $\alpha' = 0$  and the kernel  $K(t, p)$  is well defined. However, this kernel is not as convenient for finding the error estimate as the Bessel function kernel  $J_{\nu+p}$  which appears in the remainder given by Wong [19], because the properties of  $K(t, p)$  have to be studied before we can use it effectively whereas the properties of the Bessel functions are well known. In [15], the differential operator  $t^{-1} d/dt$  is used to modify (5.8) so that the remainder can be expressed in the desired form. This approach can be used in other cases, particularly when the transform kernel is a well known special function and its Mellin transform involves gamma functions. In general, a judicious choice of the differential operator would result in a more useful representation for the remainder.

#### REFERENCES

- [1] N. BLEISTEIN, *Asymptotic expansions of integral transforms of functions with logarithmic singularities*, SIAM J. Math. Anal., 8 (1977), pp. 655–672.
- [2] N. BLEISTEIN AND R. A. HANDELSMAN, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, New York, 1975.
- [3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Tables of Integral Transforms*, vol. 1, McGraw-Hill, New York, 1954.
- [4] ———, *Higher Transcendental Functions*, vol. 2, McGraw-Hill, New York, 1953.
- [5] R. A. HANDELSMAN AND J. S. LEW, *Asymptotic expansion of a class of integral transforms via Mellin transforms*, Arch. Rational Mech. Anal., 35 (1969), pp. 382–396.
- [6] ———, *Asymptotic expansion of a class of integral transforms with algebraically dominated kernels*, J. Math. Anal. Appl., 35 (1971), pp. 405–433.
- [7] G. H. HARDY AND E. C. TITCHMARCH, *A class of Fourier kernels*, Proc. London Math. Soc., 35 (1933), pp. 116–155.
- [8] G. H. HARDY, *Divergent Series*, Oxford Univ. Press, Oxford, 1956.
- [9] T. KAWATA, *Fourier Analysis in Probability Theory*, Academic Press, New York, 1972.
- [10] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.
- [11] J. P. MCCLURE AND R. WONG, *Explicit error terms for asymptotic expansions of Stieltjes transforms*, J. Inst. Math. Appl., 22 (1978), pp. 129–145.
- [12] F. W. J. OLVER, *Error bounds for stationary phase approximation*, SIAM J. Math. Anal., 5 (1974), pp. 19–29.
- [13] ———, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
- [14] K. SONI, *On uniform asymptotic expansion of finite Laplace and Fourier integrals*, to appear.
- [15] ———, *Asymptotic expansion of the Hankel transform with explicit remainder terms* (under revision).
- [16] K. SONI AND R. P. SONI, *Slowly varying functions and asymptotic behavior of a class of integral transforms I*, J. Math. Anal. Appl., 49 (1975), pp. 166–179.
- [17] E. C. TITCHMARCH, *The Theory of Functions*, 2nd ed., Oxford Univ. Press, London, 1958.
- [18] ———, *Introduction to the Theory of Fourier Integrals*, 2nd ed., Oxford Univ. Press, London, 1948.
- [19] R. WONG, *Error bounds for asymptotic expansions of Hankel transforms*, SIAM J. Math. Anal., 7 (1976), pp. 799–808.
- [20] A. H. ZEMANIAN, *Generalized Integral Transformations*, John Wiley, New York, 1968.

## THE QUENCHING OF SOLUTIONS OF SOME NONLINEAR PARABOLIC EQUATIONS\*

HOWARD A. LEVINE† AND JOHN T. MONTGOMERY‡

**Abstract.** We consider the first initial-boundary value problem for  $u_t = u_{xx} + \phi(u)$ ,  $0 \leq x \leq l$  with  $\phi > 0$  on  $[0, a)$ ,  $\phi$  convex, monotone increasing and  $\lim_{u \rightarrow a} \phi(u) = \infty$ ,  $a < \infty$ , and with  $u(x, 0) \equiv 0$ . If  $\Phi(c) = \int_0^c \phi(\eta) d\eta$ ,  $\psi(c) = 2\sqrt{2} \int_0^{c^{1/2}} dy/\phi(\Phi^{-1}(c - y^2))$  and  $l_0 = \sup \{\Psi(c) | c \in (\text{Range } \Phi) \cap [0, \infty)\}$ , we prove the following: (a) if  $l < l_0$ ,  $u$  exists for all  $t > 0$  and approaches  $(t \rightarrow \infty)$ , the smallest stationary solution of the differential equation; (b) if  $l = l_0$  and  $l_0$  is taken by  $\Psi$ , then (a) holds; (c) if  $l_0$  is not taken and  $\text{Range } \Phi$  is bounded, then  $u$  approaches from below the smallest weak stationary solution of the differential equation and this weak solution is not a strong stationary solution,  $u_{xx}(l/2, t) \rightarrow -\infty$ , and  $u_t(l/2, t) \rightarrow 0$  as  $t \rightarrow \infty$ ; (d) if  $l = l_0$  and  $\text{Range } \Phi = [0, \infty)$  or (e)  $l > l_0$ , then the existence interval is finite and  $u(l/2, t) \rightarrow a$  as  $t \rightarrow T^-$  for some  $T < \infty$ .

1. In [3], Kawarada established the following interesting results for the initial-boundary value problem:

$$(1.1) \quad u_t = u_{xx} + 1/(1 - u) \quad \text{for } 0 \leq x \leq l, 0 \leq t < T,$$

$$(1.2) \quad u(0, t) = u(l, t) = 0 \quad \text{for } 0 \leq t < T, \text{ and}$$

$$(1.3) \quad u(x, 0) = 0 \quad \text{for } 0 \leq x \leq l.$$

**THEOREM 1.** *If  $l > 2\sqrt{2}$ , then  $u$  reaches 1 in a finite time along the line  $x = l/2$ .*

Along with this result, Kawarada was interested in *quenching*, and proved the following more difficult theorem:

**THEOREM 2.** *If the solution of (1.1)–(1.3) reaches one in finite time, then  $u$  is quenched in that time; that is,*

$$\limsup_{t \rightarrow T^-} \sup_x |u_t(x, t)| = \infty.$$

Equation (1.1) arises in the study of electric current transients in polarized ionic conductors.

Acker and Walter [1], [2], [5] have considerably sharpened and extended Theorem 1. Among other things, they have shown that for the more general equation (2.1) in the next section, there is a number  $l_0 < \infty$  such that (a) if  $l < l_0$ , the solution exists for all  $t \geq 0$ ; and (b) if  $l > l_0$ , the solution is defined only on a finite interval  $[0, T)$ , and  $u(l/2, t) \rightarrow 1^-$  as  $t \rightarrow T$  from below.

The behavior at  $l = l_0$  was not determined, however, and it is the purpose of this paper to do so. The result appears in Theorem 3.

This research also duplicates some of the results of Acker and Walter mentioned above. It was done independently before the authors learned of [1], [2], [5]. Additionally, our methods are somewhat different from those of [1], [2], [5].

The techniques we employ allow us to examine, for example, the equation  $u_t = u_{xx} + (1 - u)^{-1/2}$  at  $l_0 = 4\sqrt{2}/3$ . This shows that Kawarada's Theorem 2 is not true for the more general case, since although solutions of this equation exist for all time, we have  $\sup_x |u(x, t)| \rightarrow 1$  as  $t \rightarrow \infty$ ; furthermore, it is  $u_{xx}(l/2, t)$  which blows up, and not  $u_t(l/2, t)$ , as  $t \rightarrow \infty$ .

\* Received by the editors November 8, 1978 and in revised form November 21, 1979.

† 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.

‡ Department of Mathematics, University of Rhode Island, Kingston, Rhode Island 02881.



2. The equations are:

$$(2.1) \quad u_t = u_{xx} + \phi(u),$$

$$(2.2) \quad u(0, t) = 0, \quad u(l, t) = 0 \quad \text{for } 0 \leq t < T,$$

$$(2.3) \quad u(x, 0) = 0 \quad \text{for } 0 \leq x \leq l,$$

where  $\phi$  is continuous on the interval  $[0, a)$  and has a continuous positive derivative over this interval,  $\lim_{u \rightarrow a^-} \phi(u) = \infty$ , and  $\phi(0) > 0$ . There is a close relationship between solutions of (2.1)–(2.3) and stationary solutions of (2.1) and (2.2); i.e., solutions of

$$(3.1) \quad 0 = f_{xx}(x) + \phi(f(x)) \quad \text{and}$$

$$(3.2) \quad f(0) = f(l) = 0.$$

A weak stationary solution of (2.1) and (2.2) is a once continuously differentiable function  $g$  which satisfies (3.1) and (3.2) with the possible exceptions of  $x = l/2$  and

$$(4.1) \quad g(x) = \int_0^l G(x, y)\phi(g(y)) dy,$$

where  $G(x, y)$  is the Green's function associated with the operator  $-d^2/dx^2$  on  $[0, l]$  with Dirichlet end conditions at 0 and  $l$ . That is,

$$G(x, y) \equiv \begin{cases} \frac{x}{l}(l-y) & \text{for } 0 \leq x \leq y \leq l, \\ \frac{y}{l}(l-x) & \text{for } 0 \leq y \leq x \leq l. \end{cases}$$

Note that a stationary solution is also a weak stationary solution.

Let us establish the following notation:  $\Phi(u) \equiv \int_0^u \phi(v) dv$  for  $0 \leq u < a$ , and  $R \equiv \{\Phi(u) | 0 \leq u < a\}$ , the range of  $\Phi$ . Since  $\Phi$  is monotone,  $\Phi^{-1}$  exists on  $R$ . We also let  $l_0 \equiv \sup \{\Psi(c) | c \in R\}$  where  $\Psi(c) \equiv 2\sqrt{2} \int_0^c (\Phi(\alpha) - \Phi(\eta))^{-1/2} d\eta$ ,  $(\Phi(\alpha) = c)$ . Since  $\Phi$  is positive and continuous,  $\Psi$  is bounded and  $0 < l_0 < \infty$ . To see that  $l_0$  is finite, we make use of the assumptions that  $\phi'(u) > 0$  on  $[0, a)$ , that  $0 < \phi(0) \leq \phi(u) < \infty$  on  $[0, a)$ , and that  $\phi(u) \rightarrow \infty$  as  $u \rightarrow a^-$ . Since  $\Phi$  is one to one,  $\Psi(c) = \hat{\Psi}(\alpha)$ . Thus  $l_0 = \sup \{\hat{\Psi}(\alpha) | 0 \leq \alpha < a\}$ .

From the mean value theorem, we have two numbers  $\eta_1, \eta_2 \in (\eta, \alpha)$  such that

$$\begin{aligned} \Phi(\alpha) - \Phi(\eta) &= \Phi'(\eta)(\alpha - \eta) + \frac{1}{2}\Phi''(\eta_1)(\alpha - \eta)^2 \\ &= \phi(\eta)(\alpha - \eta) + \frac{1}{2}\phi'(\eta_1)(\alpha - \eta)^2 \\ &\geq \phi(0)(\alpha - \eta) \end{aligned}$$

since  $\phi$  is increasing, and

$$\begin{aligned} \Phi(\alpha) - \Phi(\eta) &= \Phi'(\alpha)(\alpha - \eta) - \frac{1}{2}\Phi'(\eta_2)(\alpha - \eta)^2 \\ &\leq \phi(\alpha)(\alpha - \eta). \end{aligned}$$

Therefore,

$$(4.2.1) \quad \hat{\Psi}(\alpha) \leq \sqrt{2} \int_0^\alpha \frac{d\eta}{\sqrt{\phi(0)(\alpha - \eta)}} \leq 2\sqrt{2}(a/\phi(0))^{1/2},$$

while

$$(4.2.2) \quad \hat{\Psi}(\alpha) \cong \sqrt{2} \int_0^\alpha \frac{d\eta}{\sqrt{\phi(\alpha)(\alpha - \eta)}} = 2\sqrt{2}(\alpha/\phi(\alpha))^{1/2}.$$

Thus, not only is  $l_0$  finite but we also have the bounds

$$(4.3) \quad \sup_{\alpha \in [0, a]} (\alpha/\phi(\alpha))^{1/2} \leq l_0/2\sqrt{2} \leq (a/\phi(0))^{1/2}.$$

We are now ready to state the main result:

**THEOREM 3.** *The number  $l_0$  is the same as that mentioned in the introduction. Furthermore, if  $l = l_0$  in (2.1)–(2.3), then exactly one of the following hold:*

- (a) *If there exists  $c$  such that  $l_0 = \Psi(c)$ , then the solution of (2.1)–(2.3) exists for all  $t \geq 0$ . And as  $t \rightarrow \infty$ ,  $u(x, t)$  approaches monotonely from below the smallest stationary solution, which must exist and be bounded away from  $a$ .*
- (b) *If  $\Psi$  does not attain its supremum, but  $\Phi$  has bounded range  $R = [0, c_0)$ , (i.e., the integral of  $\phi$  over  $[0, a)$  is  $c_0$ ,  $u(x, t)$  exists for all  $t$  and approaches monotonely from below the smallest stationary solution  $g$ , which must be weak but not strong.)*
- (c) *If  $l = l_0$ ,  $R = [0, \infty)$ , and  $\Psi$  does not attain its supremum, then the solution  $u(x, t)$  of (2) is defined only in a finite interval  $[0, T)$ ,  $u(l/2, t) \rightarrow a$  as  $t \rightarrow T$  from below.*

We first state some preliminary results.

**LEMMA 1.** *Let  $u(x, t)$  be the solution of (2.1)–(2.3) defined on  $[0, l] \times [0, T)$ . Then the following hold:*

- (a)  *$u$  has continuous derivatives  $u_t, u_{txx}$ , and  $u_{xxx}$  on  $(0, l) \times [0, T)$ , and if  $T < \infty$  is maximal, we have  $\lim_{t \rightarrow T^-} u(x, t) = a$  for some  $x, 0 < x < l$ .<sup>1</sup>*
- (b)  *$u$  is unique and symmetric about the line  $x = l/2$ .*
- (c)  *$u_t$  is strictly positive when  $x \neq 0, x \neq l$ .*
- (d)  *$u_x$  is strictly positive for  $0 < x < l/2$ , and strictly negative for  $l/2 < x < l$ . It follows that for each  $t, u(\cdot, t)$  is strictly maximized at  $x = l/2$ .*

For the proof of (b), (c), and (d), see [1]. (a) is a more or less a standard result that follows upon formulating  $u$  as a double integral of  $\phi$  against Green’s function for the heat equation.

The proof of Theorem 3 requires two more preliminary lemmas.

**LEMMA 2.** *The solution  $u(x, t)$  of (2.1)–(2.3) exists for all  $t \geq 0$  if and only if there exists a weak stationary solution of (2.1) and (2.2). In this case,  $u(\cdot, t)$  approaches uniformly from below the smallest weak stationary solution as  $t \rightarrow \infty$ .*

*Proof.* Suppose  $f$  is a weak stationary solution of (3.1) and (3.2) and  $w = f - u$ . Then  $w$  satisfies at  $x \neq l/2$  (for some  $u_0$  between  $f$  and  $u$ ):

$$(5.1) \quad w_t = w_{xx} + \phi(f) - \phi(u) = w_{xx} + \phi'(u_0)w,$$

$$(5.2) \quad w(x, 0) = f(x), \quad w(0, t) = 0, \quad w(l, t) = 0, \quad \text{and}$$

$$(5.3) \quad w(l/2, t) \geq 0.$$

It follows from the maximum principle (Theorem 4, p. 173 of [5]) (applied for  $x \in (0, l/2)$ , and  $x \in (l/2, l)$ ) that  $w \geq 0$ .

We will first show that  $u$  must exist for all  $t \geq 0$ : Since  $w_t = -u_t$  is nonpositive and  $\phi(f) - \phi(u)$  is nonnegative, it follows from (5.1) that  $w_{xx}(x, t) \leq 0$  except possibly at  $x = l/2$ . However,  $w_x$  exists and is continuous everywhere; furthermore, it follows from Lemma 1(d) that  $w_x$  is zero at  $x = l/2$ . It is an amusing exercise in elementary calculus to

<sup>1</sup> Actually, for our choice of initial values,  $x = l/2$  if  $T < \infty$ .

show that this implies that  $w$  is maximized at  $x = l/2$ . If there is a  $T$  such that  $u$  is defined only for  $0 \leq t < T$ , then Lemma 1 implies that  $\lim_{t \rightarrow T^-} u(l/2, t) = a$  and the above argument implies that  $\lim_{t \rightarrow T^-} u(x, t) = f(x)$  uniformly in  $x$ . We will show this cannot happen unless  $T = \infty$ . Let  $x_0 < l/2$ . Then on  $[0, x_0] \times [0, T]$ ,  $w$  is nonnegative and satisfies (5.1) with  $u(x, T) \equiv f(x)$ . The maximum principle would then imply that  $w(x, T) > 0$  for  $0 < x < x_0$ . Since  $w(x, T) \equiv 0$ ,  $T = \infty$ .

The second step in the proof of this lemma is to show that if  $u$  exists for all  $t \geq 0$ , then  $u$  approaches uniformly from below in a monotone fashion a weak stationary solution  $g$ .

To see this, let  $F(x, t) \equiv \int_0^l u(y, t)G(x, y) dy$ . Then

$$(6.1) \quad \begin{aligned} F_t(x, t) &= \int_0^l u_t(y, t)G(x, y) dy \\ &= \int_0^l u_{xx}(y, t)G(x, y) dy + \int_0^l G(x, y)\phi(u(y, t)) dy, \end{aligned}$$

or

$$(6.2) \quad F_t(x, t) = -u(x, t) + \int_0^l G(x, y)\phi(u(y, t)) dy,$$

which is valid on  $[0, l]$  for any  $t$  for which  $u(x, t) < a$ . Since  $\phi' > 0$  and  $u_t > 0$ , the integrand in (6.2) is monotone in  $t$ ; thus the monotone convergence theorem implies that the right side of (6.2) approaches the limit

$$J(x) \equiv -g(x) + \int_0^l G(x, y)\phi(g(y)) dy,$$

where we have set  $g(x) = \lim_{t \rightarrow \infty} u(x, t) \leq a$ . We claim that  $J(x) = 0$  for all  $x$ . In view of (6.1) and the fact that  $u_t > 0$ , we have that  $J \geq 0$ . But if for some  $x$ , we have  $J(x) > 0$ , then it follows easily that  $F(x, t)$  would increase without bound as  $t \rightarrow \infty$ , and examination of the definition of  $F$  reveals that  $u$  would reach  $a$  in finite time, contrary to assumption. Therefore,  $J(x) = 0$ . Rewriting this, we have that  $g$  is a solution of (3.1) and (3.2). It follows that  $g$  is continuous, and from (d) of Lemma 1 and the fact that the integral in (4.1) is finite it follows that it is possible that  $g(x) = a$  only if  $x = l/2$ . Now it is a routine matter to verify that  $g$  is continuously differentiable, and at any point  $x$  where  $g(x) \neq a$ , that  $g$  is twice differentiable and satisfies (3.1) and (3.2).

LEMMA 3. *A weak solution exists if and only if there is a real number  $c$  such that either  $c \in \mathbf{R}$  and  $\psi(c) = l$ , or else  $c = \int_0^a \phi(u) du < \infty$  and  $l = \lim_{a \rightarrow c} \psi(d)$ . In the latter cases, the weak solution is not strong.*

*Proof.* Let  $f$  be a weak solution. It is easy to show that  $f$  must be symmetric about  $x = l/2$ , and that  $f_x(l/2) = 0$ . On the interval where  $f$  satisfies (3.1) and (3.2),  $f$  must lie on a level surface of the Hamiltonian ‘‘energy’’ function associated with (3.1); that is,

$$(7) \quad H(f, f_x) \equiv \frac{1}{2}f_x^2 + \Phi(f) = c,$$

where  $c$  is a constant which, since  $f(0) = \Phi(0) = 0$  and  $f_x(l/2) = 0$ , must satisfy  $c = \frac{1}{2}f_x^2(0) = \Phi(f(l/2))$ . From (4.1) it follows that  $f_x(x)$  is positive for  $x < l/2$ , so (7) can be rewritten for  $0 \leq x \leq l/2$  as

$$(8) \quad (c - \Phi(f))^{-1/2}f_x = \sqrt{2},$$

which can be integrated from 0 to  $l/2$  to get

$$(9) \quad \int_0^{l/2} (c - \Phi(f(x)))^{-1/2} f_x(x) dx = l/\sqrt{2}.$$

Letting  $y = f(x)$ , we obtain

$$(10) \quad l = \sqrt{2} \int_0^{f(l/2)} [\Phi(f(l/2)) - \Phi(y)]^{-1/2} dy,$$

and therefore  $l = \Psi(c)$ .

On the other hand, suppose there is a number  $c \in \mathbb{R}$  such that (10) holds. Let  $f$  be the unique solution of (3.1) with  $f(0) = 0$  and  $f_x(0) = (2c)^{1/2}$ . Then  $f$  is defined for all  $x$  such that  $f(x) < a$ , and  $f$  satisfies (7) and therefore (8), as long as  $f_x \geq 0$ . It is clear from (8) that there must be a point  $x_0$  such that  $f_x(x_0) = 0$  and  $\Phi(f(x_0)) = c$ . Otherwise  $f_x$  is bounded away from 0, which would imply that  $f$  increases to at least  $\Phi^{-1}(c)$ , which would in turn imply that  $f_x$  would decrease to 0, contrary to assumption. Integrating (8) from 0 to  $x_0$  yields  $\int_0^{x_0} (c - \Phi(f(x)))^{-1/2} f_x(x) dx = \sqrt{2}x_0$ . Thus,

$$x_0 = \frac{1}{2}\sqrt{2} \int_0^{f(l/2)} [\Phi(f(l/2)) - \Phi(y)]^{-1/2} dy.$$

This, with (10), implies that  $x_0 = l/2$ . Since  $f$  satisfies (3.1) on  $0 \leq x < l/2$ , and  $f(l/2) = \Phi^{-1}(c) < a$ ,  $f$  must extend to a strong stationary solution.

Now consider the case where  $c = \int_0^a \phi(u) du < \infty$ , and let  $f$  be the unique solution of (8) with  $f(0) = 0$ . Then an argument similar to the previous one shows that  $f_x(l/2) = 0$ , which implies that  $f(l/2) = a$ . Defining  $f(l/2 + x) \equiv f(l/2 - x)$ , we see that  $f$  satisfies (7) and is twice differentiable except at  $l/2$ , and therefore is a weak stationary solution of (2.1) and (2.2), but not a strong stationary solution.

*Example 1 (Kawarada).* We examine (1.1)–(1.3), the case where  $\phi(u) = (1 - u)^{-1}$ . Then  $\Phi(u) = \int_0^u (1 - r)^{-1} dr = -\ln(1 - u)$ ,  $\Phi^{-1}(y) = 1 - e^{-y}$ , and  $R = [0, \infty)$ . Thus  $\Psi(c) = 2\sqrt{2} e^{-b^2} \int_0^b e^{y^2} dy$  where  $b = \sqrt{c}$ . But this is just a multiple of Dawson’s integral  $D(b)$  [6], whose unique local maximum  $D_0$  is known to occur at a finite value of  $b$ . Consequently, if  $l < l_0 \equiv 2\sqrt{2}D_0 \approx 1.5303$ , there are two equilibrium solutions. Thus, the solution exists for all time. If  $l = l_0$ , then there is one equilibrium solution and still the solution exists for all  $t \geq 0$ . Finally, if  $l > l_0$ , there are no equilibrium solutions and  $u$  reaches 1 in finite time. This with Theorem 2 implies that quenching occurs in finite time.

Our next example shows that quenching need not occur (in Kawarada’s sense) even though  $u \rightarrow a$ .

*Example 2.* Examine the case that  $\phi(u) = (1 - u)^{-1/2}$ . One can easily compute that  $\Phi^{-1}(y) = 1 - (1 - y/2)^2$ ,  $R = [0, 1)$ , and  $\Phi(\eta) = 2(1 - (1 - \eta)^{1/2})$ . Thus,

$$\Psi(c) = 2\sqrt{2}(c^{1/2} - c^{3/2}/3),$$

which is monotone increasing in  $(0, 1)$ . Thus, when  $l = l_0 = 4\sqrt{2}/3$  we are in case (b) of Theorem 3. In this case,  $u \rightarrow 1$ , but in infinite time. Furthermore,  $u_t \rightarrow 0$  so quenching does not occur. When  $l < l_0$ , there is one equilibrium solution and there is no quenching; when  $l > l_0$ , there are no equilibria, and quenching occurs in finite time.

**3.** We would now like to indicate briefly how the foregoing can help to reveal a more global picture of the semigroup generated by (2.1) and (2.2). There is no reason to assume that  $\phi$  is defined only for positive initial data. Indeed, in Kawarada’s original equation  $\phi$  is defined on  $(-\infty, 1)$ , and it seems reasonable to ask about nonzero,

possibly even negative, initial data. Thus, in this section we assume  $\phi$  is defined on  $(-\infty, a)$ , is positive, with  $\phi' \geq 0$  and  $\lim_{u \rightarrow a^-} \phi(u) = \infty$ .

**THEOREM 4.** (a) *If (2.1) and (2.2) has no weak stationary solutions, then every solution with continuous initial data reaches  $a$  in finite time.* (b) *If  $f$  is a strong equilibrium solution, then any solution with continuous initial data which is everywhere smaller than  $f$  exists for all  $t \geq 0$  and remains smaller than  $f$ .*

*Proof.* Part (b) follows from the maximum principle. To prove (a), let  $u$  be a solution of (2.1) and (2.2) with continuous initial data  $u_0(x) = u(x, 0)$ . Then there exists a nonpositive function  $v_0(x)$  symmetric about  $x = l/2$  with  $v_0(0) = v_0(l) = 0$ ,  $v_{xx} > 0$  for  $0 < x < l$ , and  $u_0(x) \geq v_0(x)$  for all  $x$ . As before, it follows from the maximum principle's application to the equation  $w_t = w_{xx} + \phi'(v)w$  satisfied by  $w = v$ , that if  $v(x, t)$  is the solution of (2.1) and (2.2) with  $v(x, 0) = v_0(x)$ , then  $v_t \geq 0$  and  $v(\cdot, t)$  is symmetric as long as the solution exists. Examination of the proof of Lemma 2 now reveals that if  $v$  exists for all  $t \geq 0$ , it must increase to a weak stationary solution. Since there are none, it follows that  $v$  reaches  $a$  in finite time.

#### REFERENCES

- [1] A. ACKER AND W. WALTER, *The Quenching Problem for Nonlinear Parabolic Equations*, Lecture Notes in Mathematics, 564, Springer-Verlag, 1976.
- [2] ———, *On the global existence of solutions of parabolic differential equations with a singular non-linear term*, *Nonlinear Analysis*, 2 (1978), pp. 499–505.
- [3] H. KAWARADA, *On the solutions of initial boundary problem for  $u_t = u_{xx} + 1/(1-u)$* , *Publ. RIMS Kyoto Univ.*, 10 (1975), pp. 729–736.
- [4] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice Hall, Englewood Cliffs, NJ, 1967.
- [5] W. WALTER, *Parabolic differential equations with a singular nonlinear term*, *Funkcial. Ekvac.*, 19 (1976), pp. 271–277.
- [6] M. ABROMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover Publications, New York, 1965.

## THE CONNECTION PROBLEM FOR GENERAL LINEAR ORDINARY DIFFERENTIAL EQUATIONS AT TWO REGULAR SINGULAR POINTS WITH APPLICATIONS IN THE THEORY OF SPECIAL FUNCTIONS\*

REINHARD SCHÄFKE† AND DIETER SCHMIDT‡

**Abstract.** A two point connection problem for local solutions at two regular singularities of a general linear ordinary differential equation is studied. Explicit formulas for the connection coefficients are obtained which have a wide field of applications especially in the theory of special functions of mathematical physics. Applications to the ellipsoidal wave equation and to Heun's equation are considered.

**Introduction.** In the theory of complex ordinary differential equations a study of the global behavior of solutions is one of the most interesting and difficult problems. Specifically, such a global problem for linear equations consists in finding explicit connection relations between the local solutions at two different (regular or singular) points  $z_0$  and  $z_1$ . That is what one usually calls a two point connection problem. In the present paper we study the two point connection problem for two singularities  $z_0$  and  $z_1$  of the first kind (simple singularities) under the general assumption that there are no other singularities than  $z_0$  and  $z_1$  within the closed disk  $|z - z_0| \leq |z_1 - z_0|$ . In § 1 we consider the case of a general first order system of differential equations. Without loss of generality it may be assumed that the two singular points  $z_0$  and  $z_1$  are at 0 and 1 and no further singularity is within the open disk  $\mathfrak{R} := \{z \in \mathbb{C} : |z| < r\}$ , where  $1 < r \leq \infty$ . Then the equation has the form

$$(0.1) \quad y'(z) = \left( \frac{1}{z} A_0 + \frac{1}{z-1} A_1 + G(z) \right) y(z),$$

where  $A_0$  and  $A_1$  are complex  $n$  by  $n$  matrices and  $G$  is a corresponding matrix-valued function holomorphic in  $\mathfrak{R}$ .

Since 0 and 1 are simple singularities of Equation (0.1), the local behavior of the solutions at these points is completely known (see, for example, [1, Chap. 4], [3, pp. 192–198, 234–239], [12]): For each singularity there exists a characteristic fundamental set of solutions. The central problem arising here is to find a method to evaluate the connection relation between these fundamental sets explicitly. In the present note we solve this problem essentially for the special case where no logarithmic terms appear in the fundamental sets at 0 and 1. The general case is treated by R. Schäfke in an another article appearing in this issue (see pp. 863–875).

More precisely, we consider here a Floquet solution  $y$  of (0.1) at 0,

$$y(z) = z^\alpha \sum_{k=0}^{\infty} z^k d_k,$$

where  $\alpha$  is an eigenvalue of  $A_0$  and the  $d_k \in \mathbb{C}^n$ . Furthermore, we consider a fundamental set  $y_1, \dots, y_n$  of Floquet solutions of (0.1) at 1:

$$y_j(z) = (1-z)^{\alpha_j} \sum_{k=0}^{\infty} (1-z)^k d_k^j, \quad (j = 1, \dots, n),$$

\* Received by the editors February 12, 1979, and in final revised form December 12, 1979.

† Fachbereich Mathematik, Universität Essen, Gesamthochschule, Postfach 6843, 4300 Essen 1, West Germany.

where  $\alpha_j$  are eigenvalues of  $A_1$  and the  $d_k^j \in \mathbb{C}^n$ . (A sufficient condition to insure the existence of such a fundamental set is that  $A_1$  is diagonalizable and the eigenvalues of  $A_1$  do not differ by nonzero integers.) The  $y$  can be written as a linear combination of the  $y_j$ :

$$y(z) = \sum_{j=1}^n \gamma_j y_j(z)$$

with  $\gamma_j \in \mathbb{C}$  called “connection coefficients”. The main result of § 1 is an explicit limit formula for these connection coefficients in terms of the  $d_k$  as  $k \rightarrow \infty$  and some of the  $d_k^j$ . We obtain this formula—which is not only of theoretical but also of practical interest since it allows the derivation of methods for numerical computations—by a surprisingly simple idea, using only the Cauchy integral formula and making some estimates of the integrals. (As the referee remarked, a similar method is used in analytic number theory, where it is known as the “circle method.”)

In § 2 we restrict our considerations to the case of a general second order differential equation which we assume to be in the following normal form:

$$(0.2) \quad y''(z) + \left( \frac{1-\mu_0}{z} + \frac{1-\mu_1}{z-1} + a(z) \right) y'(z) + \frac{b(z)}{z(z-1)} y(z) = 0,$$

where  $\mu_0$  and  $\mu_1$  are complex numbers and  $a$  and  $b$  are holomorphic functions in  $\mathfrak{R}$ . 0 and 1 are simple singularities with exponents  $\{0, \mu_0\}$  and  $\{0, \mu_1\}$ .

In the special situation of equation (0.2) the method of § 1 and the resulting formulas become very simple. It will be shown that in this case the full connection problem can be solved by essentially considering only *one* connection coefficient and without making any restrictive assumptions on the parameters  $\mu_0$  and  $\mu_1$  or the functions  $a$  and  $b$ .

In § 3 some important applications to the theory of special functions are discussed. We consider there the special second order differential equation

$$(0.3) \quad y''(z) + \left( \frac{1-\mu_0}{z} + \frac{1-\mu_1}{z-1} + \frac{1-\mu_2}{z-a} + \alpha \right) y'(z) + \frac{\beta_0 + \beta_1 z + \beta_2 z^2}{z(z-1)(z-a)} y(z) = 0,$$

where  $a \in \mathbb{C} \setminus \{0, 1\}$  and  $\mu_0, \mu_1, \mu_2, \alpha, \beta_0, \beta_1, \beta_2$  are arbitrary complex numbers. 0, 1, and  $a$  are then simple singularities with exponents  $\{0, \mu_0\}$ ,  $\{0, \mu_1\}$  and  $\{0, \mu_2\}$  respectively, while  $\infty$  is (at most) an irregular singularity of rank 1.

It will be shown, that, with a few restrictions on  $a$ , *all* connection coefficients between the Floquet solutions at the three simple singularities can be obtained by the results of § 2, and thus the full monodromy group of the equation can be determined. Furthermore, an important property of equation (0.3) is that the connection coefficients may be computed by four-term recurrence relations.

To underline the importance of equation (0.3), it may be sufficient to remark that it contains the ellipsoidal wave equation as well as Heun’s equation and thus the Mathieu, spheroidal, Lamé, Whittaker-Hill and Ince equations as special cases.

Our main results are summarized in Theorems 1.7, 2.15 and 3.13. They seem to be new even when restricted to the special case of the Heun equation (See, e.g., [4]). Some of the results, in the case of the ellipsoidal wave equation, can be found in [11], but the methods used there are quite different and much more complicated.

For the Mathieu, spheroidal, Whittaker-Hill and Ince equations the connection problem discussed herein is not as interesting as the problem of the “characteristic exponents,” especially for methods for their numerical calculation. These problems

have been discussed in a series of papers ([9], [5], and [6]). Since the knowledge of some of the connection coefficients implies the knowledge of the characteristic exponents, our results can be applied to this problem, too. By doing so, we obtain a new, simpler theoretical foundation for the results in [9], [5], and [6].

**1. On the general first order system (0.1).** In this section we make the following general assumptions on the system of differential equations (0.1):

Let  $y$  be a Floquet solution of (0.1) at 0 given by

$$(1.1) \quad y(z) = z^\alpha h(z), \quad h(z) = \sum_{k=0}^\infty z^k d_k,$$

where  $\alpha \in \mathbb{C}$ , the  $d_k \in \mathbb{C}^n$  and especially  $d_0 \neq 0$ .  $\alpha$  is then an eigenvalue of  $A_0$  with corresponding eigenvector  $d_0$ .

Further, suppose that there exists a fundamental set  $y_1, \dots, y_n$  of Floquet solutions of (0.1) at 1 given by

$$(1.2) \quad y_j(z) = (1-z)^{\alpha_j} h_j(z), \quad h_j(z) = \sum_{k=0}^\infty (1-z)^k d_k^j,$$

where the  $\alpha_j \in \mathbb{C}$ , the  $d_k^j \in \mathbb{C}^n$  and especially the  $d_0^j \neq 0$ . In this case the  $\alpha_j$  are eigenvalues of  $A_1$  with corresponding eigenvectors  $d_0^j$ .

If the powers  $z^\alpha$  and  $(1-z)^{\alpha_j}$  are determined for  $z \in ]0, 1[$  by  $\arg z = \arg(1-z) = 0$ , a linear connection relation

$$(1.3) \quad y(z) = \sum_{j=1}^n \gamma_j y_j(z), \quad (z \in ]0, 1[),$$

with unique coefficients  $\gamma_j \in \mathbb{C}$  ( $j = 1, \dots, n$ ) is valid.

The aim of the following considerations is, as stated in the introduction, to obtain an explicit formula for the connection coefficients  $\gamma_j$ . This will be done by deriving first an asymptotic formula for the coefficients  $d_k$  as  $k \rightarrow \infty$  in terms of the  $\gamma_j$  and some of the  $d_k^j$ . The precise statement of this fundamental result is given in the following theorem, which we will prove at the end of this section.

**THEOREM 1.4.** *Suppose the general assumptions and notations in (1.1), (1.2), (1.3) to be given. Then*

$$d_k = \sum_{j=1}^n \gamma_j \left( \sum_{l=0}^{m_j} \frac{\Gamma(k + \alpha - l - \alpha_j)}{\Gamma(k + \alpha + 1)} \frac{1}{\Gamma(-l - \alpha_j)} d_l^j \right) + \mathcal{O}(k^{-\beta-1}) \quad \text{as } k \rightarrow \infty,$$

where the  $m_j$  are arbitrary nonnegative integers and  $\beta := \min \{ \operatorname{Re} \alpha_j + m_j + 1 : j = 1, \dots, n \}$ .

Since, by Stirling's series, for any  $\delta, \varepsilon \in \mathbb{C}$

$$(1.5) \quad \frac{\Gamma(z + \delta)}{\Gamma(z + \varepsilon)} \sim z^{\delta - \varepsilon} \left( 1 + \sum_{m=1}^\infty \tau_m z^{-m} \right) \quad \text{as } \operatorname{Re} z \rightarrow +\infty,$$

with coefficients  $\tau_m \in \mathbb{C}$  depending on  $\delta$  and  $\varepsilon$  (see e.g., [7, Ch. 4, § 5]), one can derive from Theorem 1.4 the asymptotic behavior of the  $d_k$  as  $k \rightarrow \infty$  in terms of powers of  $k^{-1}$ . This result immediately yields

**COROLLARY 1.6.** *Suppose the general assumptions and notations in (1.1), (1.2), (1.3) to be given. Further, let*

$$\alpha_- := \min \{ \operatorname{Re} \alpha_j : j = 1, \dots, n \}$$



and

$$I := \{j: \operatorname{Re} \alpha_j = \alpha_-\}.$$

Then

$$\lim_{k \rightarrow +\infty} k^{\alpha_- + 1} d_k = \sum_{j \in I} \frac{\gamma_j}{\Gamma(-\alpha_j)} d_0^j.$$

If, in particular,  $I = \{j_0\}$  and  $\alpha_{j_0} \notin \mathbb{N}_0$ , then Corollary 1.6 is an explicit limit formula for  $\gamma_{j_0}$ . More generally, if  $\alpha_j \notin \mathbb{N}_0$  for  $j \in I$  and  $\{d_0^j : j \in I\}$  is linearly independent, then all  $\gamma_j$  ( $j \in I$ ) can be evaluated explicitly by Corollary 1.6.

With some more effort and computations it is possible to evaluate all  $\gamma_j$  ( $j = 1, \dots, n$ ). An appropriate formula (one of our main results) is contained in the following theorem.

**THEOREM 1.7.** *Suppose the general assumptions and notations in (1.1), (1.2), (1.3) to be given. Moreover, suppose the fundamental set  $y_1, \dots, y_n$  to be normalized such that  $d_0^1, \dots, d_0^n$  are linearly independent. Further, let the  $m_j \in \mathbb{N}_0$  ( $j = 1, \dots, n$ ) be chosen such that  $\beta := \min \{\operatorname{Re} \alpha_j + m_j + 1 : j = 1, \dots, n\} > \alpha_+ := \max \{\operatorname{Re} \alpha_j : j = 1, \dots, n\}$ . If  $c$  denotes the vector with components  $\gamma_j / [\Gamma(-\alpha_j)]$  ( $j = 1, \dots, n$ ), and  $D_k$  the  $n$  by  $n$  matrix with columns*

$$d_0^j + \sum_{l=1}^{m_j} \prod_{\sigma=1}^l \left( \frac{-\alpha_j - \sigma}{k + \alpha - \alpha_j - \sigma} \right) d_0^j, \quad (j = 1, \dots, n),$$

then

$$c = \lim_{k \rightarrow \infty} \operatorname{diag} \left( \frac{\Gamma(k + \alpha + 1)}{\Gamma(k + \alpha - \alpha_j)} \right) D_k^{-1} d_k,$$

the convergence being  $\mathcal{O}(k^{\alpha_+ - \beta})$  as  $k \rightarrow \infty$ .

Before proving Theorem 1.7, let us illustrate how to construct a “normalized” fundamental set  $\tilde{y}_1, \dots, \tilde{y}_n$  with linearly independent  $\tilde{d}_0^1, \dots, \tilde{d}_0^n$  from a given fundamental set  $y_1, \dots, y_n$ . Suppose, e.g.,

$$\sum_{j \in I} \delta_j d_0^j = 0,$$

with  $I = \{j: \alpha_j = \alpha_+\}$  and  $\delta_1 \neq 0$ . Replacing  $y_1$  by

$$\tilde{y}_1 := \sum_{j \in I} \delta_j y_j,$$

we obtain a fundamental set  $\tilde{y}_1, y_2, \dots, y_n$  with

$$\tilde{y}_1(z) = (1 - z)^{\tilde{\alpha}_1} \tilde{h}_1(z), \quad \tilde{h}_1(z) = \sum_{k=0}^{\infty} (1 - z)^k \tilde{d}_k^1,$$

where  $\tilde{\alpha}_1 \in \alpha_+ + \mathbb{N}$  and  $\tilde{d}_0^1 \neq 0$ . Proceeding in this way, we ultimately get the desired result.

In order to derive Theorem 1.7, let us write Theorem 1.4 in the form

$$d_k = D_k \operatorname{diag} \left( \frac{\Gamma(k + \alpha - \alpha_j)}{\Gamma(k + \alpha + 1)} \right) c + \mathcal{O}(k^{-\beta-1}).$$

Now,  $\lim_{k \rightarrow +\infty} D_k$  exists and equals the  $n$  by  $n$  matrix with columns  $d_0^j$  ( $j = 1, \dots, n$ ). Since the  $d_0^1, \dots, d_0^n$  are linearly independent,  $D_k$  is invertible for large  $k$  and  $D_k^{-1} = \mathcal{O}(1)$  as  $k \rightarrow \infty$ . Using (1.5) we immediately get Theorem 1.7.

The convergence in Theorem 1.7 will be better, the larger the  $m_j$  chosen. In this case, the evaluation of the  $D_k$  becomes in general much more extensive, however.

Theorem 1.7 actually represents an explicit limit formula for those  $\gamma_j$  with  $\alpha_j \notin \mathbb{N}_0$ . A  $\gamma_j$  with  $\alpha_j \in \mathbb{N}_0$  can be evaluated by Theorem 1.7 or by Corollary 1.6, using a preliminary transformation

$$y(z) = (1 - z)^\nu \tilde{y}(z)$$

with some suitable  $\nu \in \mathbb{C}$ , changing  $\alpha_j$  into  $\alpha_j - \nu \notin \mathbb{N}_0$ .

A special case of Theorem 1.7 should be mentioned explicitly since in this case the limit formula becomes very simple. If

$$\max \{ \operatorname{Re}(\alpha_j - \alpha_l) : j \neq l \} < 1,$$

we may choose  $m_1 = \dots = m_n = 0$  and then obtain (using (1.5) once more,)

$$(1.8) \quad c = \lim_{k \rightarrow \infty} \operatorname{diag} (k^{\alpha_j+1})(d_0^1, \dots, d_0^n)^{-1} d_k.$$

Let us now begin with the proof of the asymptotic formula, Theorem 1.4. Assume for this purpose that  $h$  is analytically continued in  $\mathfrak{R}_0 := \mathfrak{R} \setminus [1, r[$ , and the  $h_j$  ( $j = 1, \dots, n$ ) in  $\mathfrak{R}_1 := \mathfrak{R} \setminus ]-r, 0]$ . Further let the powers  $z^\alpha$  and  $(1 - z)^{\alpha_j}$  be uniquely determined in  $\mathfrak{R}_1$  and  $\mathfrak{R}_0$  by  $\arg z, \arg(1 - z) \in ]-\pi, \pi[$ , respectively. Then the connection formula (1.3) is even valid for  $z \in \mathfrak{R}_0 \cap \mathfrak{R}_1$ .

Now we choose  $1 < \rho < r$  and let  $c_0$  and  $c_1$  be the two curves of Fig. 1.

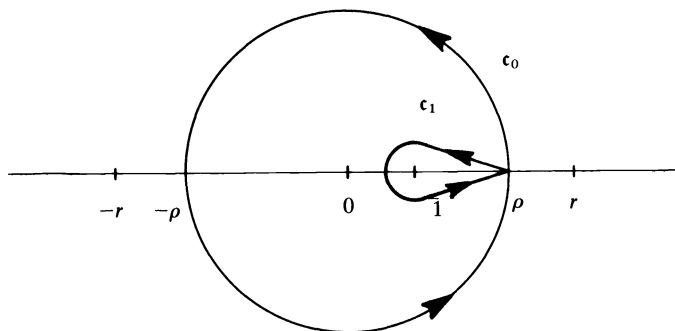


FIG. 1.

By the Cauchy formula we have for  $k \in \mathbb{N}_0$ ,

$$d_k = \frac{1}{2\pi i} \int_{c_0} z^{-k-1} h(z) dz - \frac{1}{2\pi i} \int_{c_1} z^{-k-1} h(z) dz.$$

Since  $c_1$  lies within  $\mathfrak{R}_0 \cap \mathfrak{R}_1$ , we can substitute (1.3) in the second integral. Then, for  $k \in \mathbb{N}_0$ , we get

$$(1.9) \quad d_k = \tilde{d}_k^0 + \sum_{j=1}^n \gamma_j \tilde{d}_k^j,$$

where

$$(1.10) \quad \tilde{d}_k^0 = \frac{1}{2\pi i} \int_{c_0} z^{-k-1} h(z) dz,$$

and

$$(1.11) \quad \tilde{d}_k^j = \frac{-1}{2\pi i} \int_{c_1} z^{-k-\alpha_j-1} (1 - z)^{\alpha_j} h_j(z) dz, \quad (j = 1, \dots, n).$$

In the following we shall calculate the asymptotic expansions of the  $\tilde{d}_k^j$  as  $k \rightarrow \infty$ .  
 Obviously,

$$(1.12) \quad \tilde{d}_k^0 = \mathcal{O}(\rho^{-k}) \quad \text{as } k \rightarrow \infty.$$

The asymptotic expansion of the  $\tilde{d}_k^j$  ( $j = 1, \dots, n$ ) is not so easy to derive. Nevertheless (1.11) suggests looking at Watson’s lemma in the theory of the Laplace transform. Our further calculations will follow this approach, for which we refer for example to [7, Ch. 4, § 3], or to [8, Ch. 2].

For  $j = 1, \dots, n$  and  $m \in \mathbb{N}_0$  let

$$h_j(z) = \sum_{l=0}^m (1-z)^l d_l^j + (1-z)^{m+1} \tilde{h}_j(z),$$

with  $\tilde{h}_j$  holomorphic in  $\mathfrak{R}_1$ . Substituting this in (1.11) leads to

$$(1.13) \quad \begin{aligned} \tilde{d}_k^j = \sum_{l=0}^m & \left( -\frac{1}{2\pi i} \int_{c_1} z^{-k-\alpha-1} (1-z)^{l+\alpha_j} dz \right) d_l^j \\ & - \frac{1}{2\pi i} \int_{c_1} z^{-k-\alpha-1} (1-z)^{m+1+\alpha_j} \tilde{h}_j(z) dz. \end{aligned}$$

Now we choose  $m + 1 + \operatorname{Re} \alpha_j \geq 0$ . Since  $\tilde{h}_j$  is bounded in compact subsets of  $\mathfrak{R}_1$ , there exists a constant  $M \in \mathbb{R}_+$  such that the remainder in (1.13) can be estimated for sufficiently large  $k \in \mathbb{N}$  by

$$\begin{aligned} & \left| -\frac{1}{2\pi i} \int_{c_1} z^{-k-\alpha-1} (1-z)^{m+1+\alpha_j} \tilde{h}_j(z) dz \right| \\ & = \left| \frac{\sin(\pi\alpha_j)}{\pi} \int_1^\rho z^{-k-\alpha-1} (z-1)^{m+1+\alpha_j} \tilde{h}_j(z) dz \right| \\ & \leq M \int_1^\infty z^{-k-\operatorname{Re} \alpha-1} (z-1)^{m+1+\operatorname{Re} \alpha_j} dz \\ & = M \frac{\Gamma(k + \operatorname{Re}(\alpha - \alpha_j) - m - 1) \Gamma(m + \operatorname{Re} \alpha_j + 2)}{\Gamma(k + \operatorname{Re} \alpha + 1)}. \end{aligned}$$

Here the first equality is obtained by deformation of the path of integration (see Fig. 1), and the last by the substitution  $z = t^{-1}$  which yields the first Euler integral. After use of (1.5), the remainder in (1.13) becomes, as  $k \rightarrow \infty$ ,

$$(1.14) \quad -\frac{1}{2\pi i} \int_{c_1} z^{-k-\alpha-1} (1-z)^{m+1-\alpha_j} \tilde{h}_j(z) dz = \mathcal{O}(k^{-m-\operatorname{Re} \alpha_j-2}).$$

Below it will be shown that for any  $\alpha, \beta \in \mathbb{C}$ ,

$$(1.15) \quad -\frac{1}{2\pi i} \int_{c_1} z^{-k-\alpha-1} (1-z)^\beta dz = \frac{\Gamma(k + \alpha - \beta)}{\Gamma(k + \alpha + 1)\Gamma(-\beta)} + \mathcal{O}(\rho^{-k})$$

as  $k \rightarrow \infty$ . Substituting this result and (1.14) in (1.13), we finally obtain for  $\tilde{d}_k^j$  ( $j = 1, \dots, n$ ) the asymptotic representation

$$(1.16) \quad \tilde{d}_k^j = \sum_{l=0}^m \frac{\Gamma(k + \alpha - l - \alpha_j)}{\Gamma(k + \alpha + 1)} \frac{1}{\Gamma(-l - \alpha_j)} d_l^j + \mathcal{O}(k^{-m-\operatorname{Re} \alpha_j-2}),$$

as  $k \rightarrow \infty$ ; this is now valid for all  $m \in \mathbb{N}_0$ , since we can remove the assumption  $m + 1 + \operatorname{Re} \alpha_j \geq 0$  afterwards by use of (1.5).

Together (1.9), (1.12) and (1.16) yield the asymptotic formula for  $d_k$  in Theorem 1.4.

It only remains to prove (1.15). For that purpose let  $c$  be the (infinite) curve consisting of the following three parts: First the straight line from  $\infty$  to  $\rho$  with  $\arg z = 0$  and  $\arg(1 - z) = -\pi$ , then the curve  $c_1$  of Fig. 1, and finally the straight line from  $\rho$  to  $\infty$  with  $\arg z = 0$  and  $\arg(1 - z) = \pi$ . Then for  $\operatorname{Re}(k + \alpha - \beta) > 0$ ,

$$\frac{1}{2\pi i} \left( \int_c - \int_{c_1} \right) z^{-k-\alpha-1} (1-z)^\beta dz = \frac{\sin \pi\beta}{\pi} \int_\rho^\infty z^{-k-1-\alpha} (z-1)^\beta dz.$$

Obviously,

$$\int_\rho^\infty z^{-k-1-\alpha} (z-1)^\beta dz = \mathcal{O}(\rho^{-k}) \quad \text{as } k \rightarrow \infty.$$

On the other hand, for  $\operatorname{Re}(k + \alpha - \beta) > 0$  and the temporary added condition  $\operatorname{Re} \beta > 0$ ,

$$\begin{aligned} -\frac{1}{2\pi i} \int_c z^{-k-1-\alpha} (z-1)^\beta dz &= -\frac{\sin \pi\beta}{\pi} \int_1^\infty z^{-k-1-\alpha} (z-1)^\beta dz \\ &= \frac{\sin \pi(-\beta)}{\pi} \int_0^1 t^{k+\alpha-\beta-1} (1-t)^\beta dt \\ &= \frac{\Gamma(k + \alpha - \beta)}{\Gamma(k - \alpha - 1)\Gamma(-\beta)}. \end{aligned}$$

The first equality is obtained by collapsing the curve onto the two sides of the interval  $[1, \infty[$ , the second by means of the substitution  $z = t^{-1}$ , the last by use of the beta-function integral and the reflection formula for the gamma-function. The condition  $\operatorname{Re} \beta > 0$  can be removed afterwards by analytic continuation. This completes the proof of (1.15). Thus, Theorem 1.4 is established.

We conclude this section with some remarks concerning generalizations of Theorem 1.4. These can be obtained without difficulty by modifying the proof of Theorem 1.4 in nonessential ways.

*Remarks.* Let  $y$  be an analytic function with values in any (complex) Banach space  $E$ , having a representation

$$(1.17) \quad y(z) = z^\alpha \sum_{k=-\infty}^{+\infty} z^k d_k,$$

with  $\alpha \in \mathbb{C}$  and  $d_k \in E$ , the Laurent series converging for  $r' < |z| < 1$ .

Assume further that  $y$  is analytic at all points on the circle  $|z| = 1$  except for a finite number  $s$  of singularities  $a_\sigma$  ( $\sigma = 1, \dots, s$ ), where a representation of the form

$$(1.18) \quad y(z) = \sum_{j=1}^{n_\sigma} \gamma_{\sigma j} y_{\sigma j}(z),$$

with  $n_\sigma \in \mathbb{N}$ ,  $\gamma_{\sigma j} \in \mathbb{C}$  and

$$(1.19) \quad y_{\sigma j}(z) = (a_\sigma - z)^{\alpha_{\sigma j}} \sum_{k=0}^{\infty} (a_\sigma - z)^k d_k^{\sigma j},$$

is valid. In (1.19) let  $\alpha_{\sigma j} \in \mathbb{C}$  and  $d_k^{\sigma j} \in E$ , especially  $d_0^{\sigma j} \neq 0$ . In (1.18) let the powers  $z^\alpha$

and  $(a_\sigma - z)^{\alpha_{\sigma j}}$  be determined such that  $\arg z = \arg(a_\sigma - z) \in [-\pi, \pi[$  for  $z = ta_\sigma (t \in ]0, 1[)$ .

Then we obtain the following generalization of (1.4),

$$(1.20) \quad d_k = \sum_{\sigma=1}^s \sum_{j=1}^{n_\sigma} \gamma_{\sigma j} \left( \sum_{l=0}^{m_{\sigma j}} \frac{\Gamma(k + \alpha - l - \alpha_{\sigma j})}{\Gamma(k + \alpha + 1)} \frac{a_{\sigma j}^{\alpha_{\sigma j} + l - \alpha - k}}{\Gamma(-l - \alpha_{\sigma j})} d_l^{\sigma j} \right) + \mathcal{O}(k^{-\beta-1})$$

as  $k \rightarrow +\infty$ , where the  $m_{\sigma j}$  are arbitrary nonnegative integers,  $\beta := \min \{\operatorname{Re} \alpha_{\sigma j} + m_{\sigma j} + 1 : \sigma, j\}$  and  $\arg a_\sigma \in [-\pi, \pi[$ .

This remark shows that the property of  $y$  to be a solution of a differential equation (especially an equation of the form (0.1) with two simple singularities at 0 and 1) is not at all essential for our considerations. Nevertheless, the case where  $y$  is a solution of a system of differential equations

$$y' = F(z)y,$$

where  $F$  is holomorphic in  $r' < |z| < r$  ( $r' < 1 < r$ ) except for a finite number of singularities  $a_\sigma$ , is one of the main applications of (1.20). We do not discuss further details here. It should only be mentioned that it is generally not possible to derive from (1.20) an appropriate limit formula for all  $\gamma_{\sigma j}$  corresponding to Theorem 1.7.

**2. On the second order equation (0.2).** In this section we shall study the (general) connection problem for the second order equation (0.2). For this purpose it will be useful to call special attention to the dependence of equation (0.2) on the parameters  $\mu := (\mu_0, \mu_1)$ .

First of all, it turns out to be convenient for the  $\mu$ -dependence to be symmetric with respect to certain index transformations. Therefore, let the coefficient  $b(z) = b(z, \mu)$  be of the form

$$(2.1) \quad b(z, \mu) = \frac{1}{2}(1 - \mu_0)(1 - \mu_1) + \frac{1}{2}a(z)((1 - \mu_0)(z - 1) + (1 - \mu_1)z) + b_0(z),$$

where  $b_0$  is holomorphic in  $\mathfrak{R}$ .

Further on, it turns out to be useful, especially in order to avoid additional complications for exceptional values of  $\mu$ , to consider holomorphic  $\mu$ -dependence of the solutions.

We begin by stating a proposition on the holomorphic solutions of (0.2) at the singular points  $z_0 := 0$  and  $z_1 := 1$ .

**PROPOSITION 2.2.** *For  $j = 0, 1$  there exists a unique function  $\eta_j$  holomorphic with respect to  $(z, \mu) \in \mathfrak{R}_j \times \mathbb{C}^2$  (where  $\mathfrak{R}_0 := \mathfrak{R} \setminus [1, r[$  and  $\mathfrak{R}_1 := \mathfrak{R} \setminus ]-r, 0]$ ), such that, for each  $\mu$ ,  $\eta_j(\cdot, \mu)$  is a solution of (0.2) satisfying  $\eta_j(z_j, \mu) = 1/\Gamma(1 - \mu_j)$ .  $\eta_0$  can be expanded in a power series*

$$\eta_0(z, \mu) = \sum_{k=0}^{\infty} \frac{\tau_k^0(\mu)}{\Gamma(1 - \mu_0 + k)} z^k,$$

and  $\eta_1$  correspondingly as

$$\eta_1(z, \mu) = \sum_{k=0}^{\infty} \frac{\tau_k^1(\mu)}{\Gamma(1 - \mu_1 + k)} (1 - z)^k,$$

where the (unique) coefficients  $\tau_k^j$  are holomorphic with respect to  $\mu$ . In particular,  $\tau_0^j(\mu) = 1$ .

The proof is, as usual, by the method of power series. A more detailed discussion of this and the following preliminary results can also be found in [13, §1.2].

To obtain further solutions of (0.2), we consider the substitution

$$(2.3) \quad y(z) = z^{\sigma_0}(1-z)^{\sigma_1}\tilde{y}(z), \quad \sigma_j \in \{0, \mu_j\} \quad (j = 0, 1).$$

By the special choice of the  $\mu$ -dependence in (2.1), this substitution transforms (0.2) into an equation of the same type with  $\tilde{\mu}_j = \mu_j - 2\sigma_j \in \{\mu_j, -\mu_j\}$ ,  $(j = 0, 1)$ , and  $\tilde{a} = a$ ,  $\tilde{b}_0 = b_0$ . From this it is easily seen that  $z^{\mu_0}\eta_0(z, -\mu_0, \mu_1)$  and  $(1-z)^{\mu_1}\eta_1(z, \mu_0, -\mu_1)$  are also solutions of (0.2) in  $\mathfrak{R}_0 \cap \mathfrak{R}_1$ . Thus, defining

$$(2.4) \quad (y_{01}, y_{02})(z, \mu) := (\eta_0(z, \mu_0, \mu_1), z^{\mu_0}\eta_0(z, -\mu_0, \mu_1))$$

and

$$(2.5) \quad (y_{11}, y_{12})(z, \mu) := (\eta_1(z, \mu_0, \mu_1), (1-z)^{\mu_1}\eta_1(z, \mu_0, -\mu_1)),$$

for  $z \in \mathfrak{R}_0 \cap \mathfrak{R}_1$  and  $\mu \in \mathbb{C}^2$ , we get a set of two Floquet solutions of (0.2) at  $z_0 = 0$  and at  $z_1 = 1$ , respectively.

The aim of the following discussion is to obtain explicit connection relations between the sets of Floquet solutions  $y_{j1}, y_{j2}$  for  $j = 0, 1$ . To avoid excluding the exceptional values of  $\mu_j$  where  $y_{j1}$  and  $y_{j2}$  become linearly dependent, it turns out to be necessary to write the connection relations in a homogeneous form.

For this purpose, we introduce the Wronskian

$$(2.6) \quad w[y_1, y_2](z) := y_1(z)y_2'(z) - y_2(z)y_1'(z)$$

of two arbitrary solutions  $y_1, y_2$  of (0, 2) in  $\mathfrak{R}_0 \cap \mathfrak{R}_1$ . Solving the linear first order differential equation for the Wronskian, we find that there exists a unique constant  $[y_1, y_2] \in \mathbb{C}$ , such that

$$(2.7) \quad w[y_1, y_2](z) = [y_1, y_2]z^{\mu_0-1}(1-z)^{\mu_1-1} \exp\left(-\int_1^z a(\zeta) d\zeta\right)$$

for  $z \in \mathfrak{R}_0 \cap \mathfrak{R}_1$ , the powers being determined by  $\arg z, \arg(1-z) \in ]-\pi, \pi[$ . Obviously,  $[y_1, y_2] \neq 0$  if and only if  $y_1, y_2$  constitute a fundamental set. Since  $[\cdot, \cdot]$  is bilinear and alternating, we have for any three solutions  $y, y_1, y_2$  of (0.2) in  $\mathfrak{R}_0 \cap \mathfrak{R}_1$ , the identity

$$(2.8) \quad [y_1, y_2]y(z) = [y, y_2]y_1(z) + [y_1, y]y_2(z).$$

(2.8) is the homogeneous form of the connection relation which has turned out to be appropriate for the study of the connection problem between the sets of Floquet solutions  $y_{j1}, y_{j2}$ ,  $(j = 0, 1)$ . The main advantages of using this formula are that no restrictions need be made on the parameters  $\mu_j$  involving the linear independence of  $y_{j1}, y_{j2}$ , and that the connection coefficients  $[y_{j\kappa}, y_{l\sigma}]$ ,  $(j, l \in \{0, 1\}; \kappa, \sigma \in \{1, 2\})$  are directly defined by means of the Wronskian.

The coefficients  $[y_{j1}, y_{j2}]$ ,  $(j = 0, 1)$ , can be evaluated explicitly. Substitution of the power series of  $\eta_j$  in (2.6), comparison with (2.7) and inspection of the leading terms immediately yields

$$(2.9) \quad [y_{j1}, y_{j2}] = \frac{\sin(\pi\mu_j)}{\pi} \xi_j, \quad (j = 0, 1),$$

with  $\xi_0 = \exp(-\int_0^1 a(z) dz)$  and  $\xi_1 = 1$ . This especially shows that  $y_{j1}$  and  $y_{j2}$  constitute a fundamental set of solutions if and only if  $\mu_j \notin \mathbb{Z}$ .

Next, we show that the remaining coefficients  $[y_{j\kappa}, y_{l\sigma}]$   $(j \neq l)$  can be represented by means of only one function  $q$  defined by

$$(2.10) \quad q(\mu) := [\eta_0(\cdot, \mu), \eta_1(\cdot, \mu)], \quad (\mu \in \mathbb{C}^2).$$

It can be seen from Proposition 2.2 and Equations (2.6) and (2.7) that  $q$  is an entire function of  $\mu$ .

By definition,  $[y_{01}, y_{11}] = q(\mu)$ . In order to obtain the representations of the other coefficients in terms of  $q$ , we need the following identities,

$$(2.11) \quad \eta_0(z, \mu) = (1 - z)^{\mu_1} \eta_0(z, \mu_0, -\mu_1)$$

and

$$(2.12) \quad \eta_1(z, \mu) = z^{\mu_0} \eta_1(z, -\mu_0, \mu_1),$$

which are an immediate consequence of (2.3) and the uniqueness of the functions  $\eta_j$  in Proposition 2.2. Using then (2.4), (2.5) and (2.11), (2.12) in (2.6), (2.7), with  $\mu_j$  replaced by  $-\mu_j$  if necessary, we finally obtain

$$(2.13) \quad [y_{0\kappa}, y_{1\sigma}] = q((-1)^{\kappa-1} \mu_0, (-1)^{\sigma-1} \mu_1), \quad (\kappa, \sigma \in \{1, 2\}).$$

Combining (2.8), (2.9) and (2.13) leads to

PROPOSITION 2.14. For  $\kappa = 1, 2$

$$\frac{\sin(\pi\mu_1)}{\pi} \xi_1 y_{0\kappa} = q(\pm\mu_0, -\mu_1) y_{11} - q(\pm\mu_0, \mu_1) y_{12},$$

and

$$\frac{\sin(\pi\mu_0)}{\pi} \xi_0 y_{1\kappa} = -q(-\mu_0, \pm\mu_1) y_{01} + q(\mu_0, \pm\mu_1) y_{02},$$

with  $+$  for  $\kappa = 1$  and  $-$  for  $\kappa = 2$ .

Proposition 2.14 shows that the connection problem between the sets of Floquet solutions  $y_{j1}, y_{j2}, (j = 0, 1)$  will actually be solved if the function  $q$  can be evaluated. Using the methods of § 1, we obtain the following formulas by which  $q$  may be calculated explicitly.

THEOREM 2.15. Let the notations in Proposition 2.2 be given. Then for  $\mu \in \mathbb{C}^2$ ,

$$\frac{\Gamma(k+1)}{\Gamma(k+1-\mu_0)\Gamma(k-\mu_1)} \tau_k^0(\mu) = q(\mu) \left( 1 + \sum_{l=1}^m \tau_l^1(\mu_0, -\mu_1) \prod_{\sigma=1}^l (\sigma + \mu_1 - k)^{-1} \right) + \mathcal{O}(k^{-m-1})$$

as  $k \rightarrow \infty$ , where  $m$  is an arbitrary nonnegative integer. In particular,

$$\lim_{k \rightarrow \infty} \frac{\Gamma(k+1)}{\Gamma(k+1-\mu_0)\Gamma(k-\mu_1)} \tau_k^0(\mu) = q(\mu).$$

We prove this result by proceeding exactly as in § 1. Using here the connection relation between  $\eta_0 = y_{01}$  and the  $y_{11}, y_{12}$  given in Proposition 2.14 and noting that  $y_{11} = \eta_1$  is holomorphic in  $z_1 = 1$ , we obtain for  $k \in \mathbb{N}_0$  and  $\mu \in \mathbb{C}^2$

$$(2.16) \quad \frac{\tau_k^0(\mu)}{\Gamma(1-\mu_0+k)} = \tilde{\tau}_k^0(\mu) + q(\mu) \tilde{\tau}_k^1(\mu),$$

where

$$\tilde{\tau}_k^0(\mu) = \frac{1}{2\pi i} \int_{c_0} z^{-k-1} \eta_0(z, \mu) dz$$

and

$$\tilde{\tau}_k^1(\mu) = \frac{1}{2i \sin(\pi\mu_1)} \int_{c_1} z^{-k-1}(1-z)^{\mu_1} \eta_1(z, \mu_0, -\mu_1) dz.$$

In the last formula, of course, the limits for  $\mu_1 \in \mathbb{Z}$  must be taken. These exist, since  $y_{12}(z, \mu) = (1-z)^{\mu_1} \eta_1(z, \mu_0, -\mu_1)$  is holomorphic in  $z_1 = 1$  for  $\mu_1 \in \mathbb{Z}$  and therefore the contour integral, being an entire function with respect to  $\mu_1$ , becomes zero by Cauchy’s integral theorem.

Corresponding to (1.12), we have obviously

$$\tilde{\tau}_k^0(\mu) = \mathcal{O}(\rho^{-k}) \quad \text{as } k \rightarrow \infty.$$

Corresponding to (1.16), we obtain by using (1.15) and the reflection formula for the gamma-function,

$$\tilde{\tau}_k^1(\mu) = \sum_{l=0}^m (-1)^l \frac{\Gamma(k-l-\mu_1)}{\Gamma(k+1)} \tau_l^1(\mu_0, -\mu_1) + \mathcal{O}(k^{-m-\text{Re } \mu_1 - 2})$$

as  $k \rightarrow \infty$ , where  $m$  is an arbitrary nonnegative integer. Multiplying (2.16) with  $\Gamma(k+1)/\Gamma(k-\mu_1)$  and using once more (1.5) finally leads to Theorem 2.15.

*Remarks.*

(1) If the coefficients  $a$  and  $b_0$  of equation (0.2) depend analytically on further parameters, say  $\lambda_\sigma \in \mathbb{C}$ , then the functions  $\eta_j, \tau_k^j, y_{j\kappa}$ , and  $q$  will also depend analytically on the  $\lambda_\sigma$ .

(2) In the case  $\mu_j \in \mathbb{Z}$  the solutions  $y_{j1}$  and  $y_{j2}$  become linearly dependent. Then it is always possible to obtain a fundamental system at  $z_j$  by differentiating a suitable linear combination of  $y_{j1}$  and  $y_{j2}$  with respect to  $\mu_j$ . (See [13, § 1.2]), Without going into further details it should be mentioned that the corresponding connection coefficients can be obtained by differentiating  $q(\mu)$  with respect to  $\mu_j$ .

(3) A number  $\nu \in \mathbb{C}$  is called a “characteristic exponent of (0.2) in  $1 < |z| < r$ ” if and only if there exists a nontrivial solution  $y(z) = z^\nu h(z)$  of (0.2) with  $h$  holomorphic in  $1 < |z| < r$ . In [13, § 1.3], it has been shown that the characteristic exponents  $\nu$  are then determined by

$$(2.17) \quad \cos \pi(2\nu - \mu_0 - \mu_1) = \cos \pi(\mu_0 + \mu_1) + \frac{2\pi^2}{\xi_0} q(\mu_0, \mu_1) q(-\mu_0, -\mu_1)$$

and also by

$$(2.18) \quad \cos \pi(2\nu - \mu_0 - \mu_1) = \cos \pi(\mu_0 - \mu_1) + \frac{2\pi^2}{\xi_0} q(-\mu_0, \mu_1) q(\mu_0, -\mu_1).$$

Thus, Theorem 2.15 can be used to calculate the characteristic exponents explicitly.

**3. On the generalized Heun equation (0.3).** In this section we shall apply the results of § 2 to the generalized Heun equation (0.3). For the same reasons as in that section, it will be convenient for the dependence of equation (0.3) on the parameters  $\mu := (\mu_0, \mu_1, \mu_2)$  to be symmetric with respect to certain index transformations. Thus, we let the second coefficient of (0.3) be in the form

$$(3.1) \quad \frac{\beta_0 + \beta_1 z + \beta_2 z^2}{z(z-1)(z-a)} = \sum_{\substack{\sigma, \rho=0 \\ \sigma \neq \rho}}^2 \frac{1}{2} \left( \frac{1-\mu_\sigma}{z-z_\sigma} \right) \left( \frac{1-\mu_\rho}{z-z_\rho} \right) + \sum_{\kappa=0}^2 \frac{(\alpha/2)(1-\mu_\kappa) + \lambda_\kappa}{z-z_\kappa}$$



with  $z_0 := 0, z_1 := 1, z_2 := a \in \mathbb{C} \setminus \{0, 1\}$ , and arbitrary parameters  $\lambda := (\lambda_0, \lambda_1, \lambda_2) \in \mathbb{C}^3$ .  $\lambda$  is then in a linear one-to-one relation with  $(\beta_0, \beta_1, \beta_2)$ .

As stated in Proposition 2.2 for Equation (0.2), we have a holomorphic solution of (0.3) at 0. (See also remark 1 in § 2).

PROPOSITION 3.2. *Let  $a \in \mathbb{C} \setminus \{0, 1\}$  be fixed. Then there exists a unique function  $\eta = \eta(\cdot; a)$  holomorphic with respect to  $(z, \mu, \alpha, \lambda) \in \{z \in \mathbb{C} : |z| < \min(1, |a|)\} \times \mathbb{C}^7$ , such that, for each  $(\mu, \alpha, \lambda)$ ,  $\eta(\cdot, \mu, \alpha, \lambda; a)$  is a solution of (0.3) satisfying*

$$\eta(0, \mu, \alpha, \lambda; a) = \frac{1}{\Gamma(1 - \mu_0)}. \quad \eta \text{ can be expanded in a power series}$$

$$\eta(z, \mu, \alpha, \lambda; a) = \sum_{k=0}^{\infty} \frac{\tau_k(\mu, \alpha, \lambda; a)}{\Gamma(k + 1 - \mu_0)\Gamma(k + 1)} z^k,$$

where the (unique) coefficients  $\tau_k$  are holomorphic with respect to  $(\mu, \alpha, \lambda)$ . In particular  $\tau_0(\mu, \alpha, \lambda; a) = 1$ .

Substituting the power series into equation (0.3) then leads to the following four-term recurrence relation for the  $\tau_k$ ,

$$(3.3) \quad \tau_k = \varphi_1(k - 1)\tau_{k-1} - \varphi_2(k - 2)\tau_{k-2} + \varphi_3(k - 3)\tau_{k-3}, \quad (k \in \mathbb{N}),$$

where  $\tau_{-1} = \tau_{-2} = 0$  and

$$(3.4) \quad \begin{aligned} \varphi_1(\xi) &= \xi(\xi + 1 - \mu_0 - \mu_1) + \frac{1}{a}\xi(\xi + 1 - \mu_0 - \mu_2) - \alpha\xi - \frac{1}{a}\beta_0, \\ \varphi_2(\xi) &= (\xi + 1)(\xi + 1 - \mu_0) \left( \frac{1}{a}\xi(\xi + 2 - \mu_0 - \mu_1 - \mu_2) - \left(1 + \frac{1}{a}\right)\alpha\xi + \frac{1}{a}\beta_1 \right), \\ \varphi_3(\xi) &= (\xi + 1)(\xi + 2)(\xi + 1 - \mu_0)(\xi + 2 - \mu_0) \left( -\frac{1}{a} \right) (\alpha\xi + \beta_2). \end{aligned}$$

The following transformations will demonstrate that all Floquet solutions at the (simple) singularities 0, 1 and  $a$  can be defined in terms of the function  $\eta$  investigated in (3.2).

We look first at the index transformations

$$(3.5) \quad y(z) = z^{\sigma_0}(z - 1)^{\sigma_1}(z - a)^{\sigma_2}\tilde{y}(z)$$

with  $\sigma_j \in \{0, \mu_j\}$ , ( $j = 0, 1, 2$ ). Considering the special choice of the  $\mu$ -dependence in (3.1), a straightforward calculation shows that (3.5) transforms (0.3) into an equation of the same type with  $\tilde{\mu}_j = \mu_j - 2\sigma_j \in \{\mu_j, -\mu_j\}$ , ( $j = 0, 1, 2$ ) and  $\tilde{\alpha} = \alpha, \tilde{\lambda} = \lambda, \tilde{a} = a$ .

The second class of transformations of interest are the linear transformations

$$(3.6) \quad \tilde{z} = \varepsilon z + \delta \quad (\varepsilon, \delta \in \mathbb{C}, \varepsilon \neq 0)$$

of the independent variable, which map the simple singularities  $\{0, 1, a\}$  into the simple singularities  $\{0, 1, \tilde{a}\}$  and keep the irregular singularity  $\infty$  fixed. Table 1 contains the six possible substitutions and yields all information about them.

Using (3.5) and Table 1, we can now define for each  $j = 0, 1, 2$  a set of two Floquet solutions  $y_{j1}, y_{j2}$  at  $z_j$  in terms of the function  $\eta$  by

$$(3.7) \quad \begin{aligned} y_{01}(z, \mu, \alpha, \lambda) &:= \eta(z, \mu_0, \mu_1, \mu_2, \alpha, \lambda; a), \\ y_{02}(z, \mu, \alpha, \lambda) &:= z^{\mu_0}\eta(z, -\mu_0, \mu_1, \mu_2, \alpha, \lambda; a), \end{aligned}$$

TABLE 1

	$z$	$\mu$	$a$	$\alpha$	$\lambda$
(a)	$z$	$(\mu_0, \mu_1, \mu_2)$	$a$	$\alpha$	$\lambda$
(b)	$1-z$	$(\mu_1, \mu_0, \mu_2)$	$1-a$	$-\alpha$	$-\lambda$
(c)	$\frac{z}{a}$	$(\mu_0, \mu_2, \mu_1)$	$\frac{1}{a}$	$a\alpha$	$a\lambda$
(d)	$\frac{1-z}{1-a}$	$(\mu_1, \mu_2, \mu_0)$	$\frac{1}{1-a}$	$(a-1)\alpha$	$(a-1)\lambda$
(e)	$1-\frac{z}{a}$	$(\mu_2, \mu_0, \mu_1)$	$1-\frac{1}{a}$	$-a\alpha$	$-a\lambda$
(f)	$\frac{a-z}{a-1}$	$(\mu_2, \mu_1, \mu_0)$	$\frac{a}{a-1}$	$(1-a)\alpha$	$(1-a)\lambda$

for  $|z| < \min(1, |a|)$ ,

$$(3.8) \quad \begin{aligned} y_{11}(z, \mu, \alpha, \lambda) &:= \eta(1-z, \mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a), \\ y_{12}(z, \mu, \alpha, \lambda) &:= (1-z)^{\mu_1} \eta(1-z, -\mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a), \end{aligned}$$

for  $|z-1| < \min(1, |a-1|)$ , and

$$(3.9) \quad \begin{aligned} y_{21}(z, \mu, \alpha, \lambda) &:= \eta\left(1-\frac{z}{a}, \mu_2, \mu_0, \mu_1, -a\alpha, -a\lambda; 1-\frac{1}{a}\right), \\ y_{22}(z, \mu, \alpha, \lambda) &:= \left(1-\frac{z}{a}\right)^{\mu_2} \eta\left(1-\frac{z}{a}, -\mu_2, \mu_0, \mu_1, -a\alpha, -a\lambda; 1-\frac{1}{a}\right) \end{aligned}$$

for  $|z-a| < \min(|a|, |1-a|)$ .

Furthermore, using (3.5) and Table 1, we can derive exactly 8 different representations for each function  $y_{j\kappa}$  in terms of  $\eta$ . We shall present here only the basic identities

$$(3.10) \quad \begin{aligned} \eta(z, \mu, \alpha, \lambda; a) &= (1-z)^{\mu_1} \eta(z, \mu_0, -\mu_1, \mu_2, \alpha, \lambda; a) \\ &= \left(1-\frac{z}{a}\right)^{\mu_2} \eta(z, \mu_0, \mu_1, -\mu_2, \alpha, \lambda; a) \\ &= (1-z)^{\mu_1} \left(1-\frac{z}{a}\right)^{\mu_2} \eta(z, \mu_0, -\mu_1, -\mu_2, \alpha, \lambda; a), \end{aligned}$$

valid for  $|z| < \min(1, |a|)$  and  $\arg(1-z), \arg(1-z/a) \in ]-\pi, \pi[$ , and

$$(3.11) \quad \eta(z, \mu, \alpha, \lambda; a) = \eta\left(\frac{z}{a}, \mu_0, \mu_2, \mu_1, a\alpha, a\lambda; \frac{1}{a}\right),$$

valid for  $|z| < \min(1, |a|)$ . These are immediate consequences respectively of (3.5), Table 1(c), and the uniqueness of the function  $\eta$  in Proposition 3.2.

Our further considerations deal with the connection relations between the different sets of Floquet solutions  $y_{j1}, y_{j2}$ , ( $j = 0, 1, 2$ ). We derive a basic connection formula by carrying over the results of § 2, and describe how this can be used to obtain all other connection relations between the  $y_{j1}, y_{j2}$  (except for a few values of  $a$ ).

From Proposition 2.14 we obtain, by considering (3.7), (3.8) and remark (1) in § 2,

**PROPOSITION 3.12.** *Let  $|a| > 1$ . Then there exists a unique function  $q = q(\cdot; a)$  holomorphic with respect to  $(\mu, \alpha, \lambda) \in \mathbb{C}^7$ , such that the connection formula*

$$\frac{\sin(\pi\mu_1)}{\pi} \eta(z, \mu, \alpha, \lambda; a) = q(\mu_0, -\mu_1, \mu_2, \alpha, \lambda; a) \eta(1-z, \mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a) \\ - q(\mu_0, \mu_1, \mu_2, \alpha, \lambda; a) (1-z)^{\mu_1} \\ \cdot \eta(1-z, -\mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a)$$

is valid for  $|z| < 1$ ,  $|z-1| < \min(1, |a-1|)$ ,  $\arg(1-z) \in ]-\pi, \pi[$  and  $(\mu, \alpha, \lambda) \in \mathbb{C}^7$ .

Applying Theorem 2.15 to this special case leads to

**THEOREM 3.13.** *Let  $|a| > 1$ . Then for  $(\mu, \alpha, \lambda) \in \mathbb{C}^7$ ,*

$$q(\mu, \alpha, \lambda; a) = \lim_{k \rightarrow \infty} \frac{\tau_k(\mu, \alpha, \lambda; a)}{\Gamma(k+1-\mu_0)\Gamma(k-\mu_1)} \\ \cdot \left( 1 + \sum_{l=1}^m \frac{\tau_l(-\mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a)}{l!} \prod_{\sigma=1}^l (\sigma + \mu_1 - k)^{-1} \right)^{-1},$$

the convergence being  $\mathcal{O}(k^{-m-1})$  as  $k \rightarrow \infty$ , where  $m$  is an arbitrary nonnegative integer.

The proof of Theorem 3.13 follows from (3.7), (3.8) and Propositions 2.2 and 3.2, which in particular yield

$$(3.14) \quad \tau_k^0(\mu) = \frac{\tau_k(\mu, \alpha, \lambda; a)}{k!}, \quad \tau_k^1(\mu) = \frac{\tau_k(\mu_1, \mu_0, \mu_2, -\alpha, -\lambda; 1-a)}{k!}, \quad (k \in \mathbb{N}_0)$$

Theorem 3.13 and (3.3) show that  $q$  can actually be computed by four-term recurrence relations. Furthermore, Theorem 3.13 shows that a satisfactory convergence should be attained by choosing  $m$  large enough.

Applying now the 6 substitutions of Table 1 to the connection formula Proposition 3.12, and using the identities (3.10), (3.11), we immediately obtain 6 connection relations between the Floquet solutions  $y_{j1}, y_{j2}$ , ( $j = 0, 1, 2$ ) in terms of  $q$ . In fact, substituting Table 1(a) for  $|a| > 1$  and Table 1(b) for  $|a-1| > 1$ , yields the connection relations between  $y_{01}, y_{02}$  and  $y_{11}, y_{12}$ . Correspondingly, substituting Table 1(c) for  $|a| < 1$  and Table 1(e) for  $\text{Re } a < \frac{1}{2}$  yields the connection relations between  $y_{01}, y_{02}$  and  $y_{21}, y_{22}$ . Finally, substituting Table 1(d) for  $|a-1| < 1$  and Table 1(f) for  $\text{Re } a > \frac{1}{2}$  yields the connection relations between  $y_{11}, y_{12}$  and  $y_{21}, y_{22}$ . We shall not list these formulas explicitly.

The preceding discussion shows that whenever  $a$  is within one of the regions  $\mathfrak{R}_1 := \{a \in \mathbb{C} : |a| > 1 \wedge \text{Re } a > \frac{1}{2}\}$  or  $\mathfrak{R}_2 := \{a \in \mathbb{C} : |a| < 1 \wedge |a-1| < 1\}$  or  $\mathfrak{R}_3 := \{a \in \mathbb{C} : |a-1| > 1 \wedge \text{Re } a < \frac{1}{2}\}$ , (see also Fig. 2), we always obtain two different connection relations in terms of  $q$ , one between the sets of Floquet solutions  $y_{j1}, y_{j2}$  and  $y_{k1}, y_{k2}$  and the other between the sets of Floquet solutions  $y_{k1}, y_{k2}$  and  $y_{l1}, y_{l2}$ , where  $\{j, k, l\} = \{0, 1, 2\}$ . Combining these, we also get a connection relation between the sets  $y_{j1}, y_{j2}$  and  $y_{l1}, y_{l2}$  in terms of  $q$ . Thus in these cases all connection coefficients between the Floquet solutions at the three simple singularities  $z_0 = 0, z_1 = 1$  and  $z_2 = a$  can be represented in terms of the function  $q$  and, therefore, the full monodromy group of Equation (0.3) can be computed by Theorem 3.13.

We conclude our discussion by stating some final points.

*Remarks:*

(1) When  $\alpha = 0$  and  $\beta_2 (= \lambda_0 + \lambda_1 + \lambda_2) = 0, \infty$  is also a simple singularity and (0.3) is Heun's equation. The exponents at  $\infty$  are then

$$(3.15) \quad \nu + 1 - \left(\frac{1}{2}\right)(\mu_0 + \mu_1 + \mu_2), \quad -\nu + 1 - \left(\frac{1}{2}\right)(\mu_0 + \mu_1 + \mu_2),$$

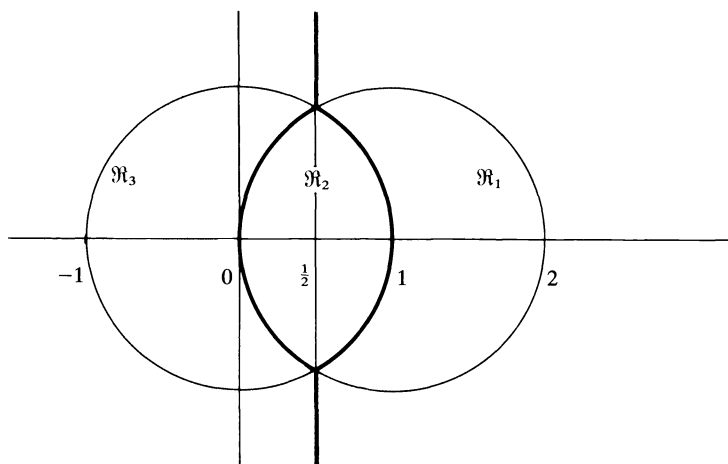


FIG. 2.

where  $\nu^2 = (\frac{1}{4})(\mu_0^2 + \mu_1^2 + \mu_2^2) - \frac{1}{2} - \lambda_1 - a\lambda_2$ . In this case, we can also define a set of Floquet solutions  $y_{31}, y_{32}$  at  $\infty$  in terms of the function  $\eta$  of Proposition 3.2. The transformations to be used here can be found in the original paper of K. Heun (see [2, p. 162–168]). By these transformations exactly 24 different representations for each of the 8 Floquet solutions  $y_{j\kappa}$ , ( $j = 0, 1, 2, 3; \kappa = 1, 2$ ), can be obtained in terms of  $\eta$ . Once more using the results of § 2, it is easy to verify that the connection coefficients between the solutions  $y_{31}, y_{32}$  at  $\infty$  and the solutions  $y_{j1}, y_{j2}$  ( $j = 0, 1, 2$ ) at the finite singularities can also be represented in terms of the function  $q$  of Proposition 3.12, whenever  $a$  is within one of the regions  $\mathfrak{R}_1, \mathfrak{R}_2, \mathfrak{R}_3$  of Fig. 2.

(2) When  $\alpha = 0$  and  $\mu_0 = \mu_1 = \mu_2 = \frac{1}{2}$ , equation (0.3) is the ellipsoidal wave equation. In this case our results agree completely with those of [11].

## REFERENCES

- [1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [2] K. HEUN, *Zur Theorie der Riemann'schen Funktionen zweiter Ordnung mit vier Verzweigungspunkten*, Math. Ann., 33 (1889), pp. 161–179.
- [3] E. HILLE, *Lectures on differential equations*, Addison-Wesley, Reading, MA, 1969.
- [4] M. KOHNO, *A two point connection problem for general linear ordinary differential equations*, Hiroshima Math. J., 4 (1974), pp. 293–338.
- [5] R. MENNICKEN, *On Ince's equation*, Arch. Rational Mech. Anal., 29 (1968), pp. 144–160.
- [6] R. MENNICKEN AND D. SCHMIDT, *Untersuchung über lineare Differentialgleichungen mit sinusförmigen Koeffizienten*, Arch. Rational Mech. Anal., 31 (1968), pp. 304–321.
- [7] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
- [8] F. W. SCHÄFKE, *Einführung in die Theorie der speziellen Funktionen der Mathematischen Physik*, Berlin-Göttingen-Heidelberg, 1963.
- [9] F. W. SCHÄFKE, *Ein Verfahren zur Berechnung des charakteristischen Exponenten der Mathieschen Differentialgleichung I*, Numer. Math., 3 (1960), pp. 30–38.
- [10] R. SCHÄFKE, *The connection problem for two neighboring regular singular points of general linear complex ordinary differential equations*, this Journal, this issue, pp. 863–875.
- [11] D. SCHMIDT, *Zur Differentialgleichung der Ellipsoidfunktionen*, Arch. Rational Mech. Anal., 53 (1973), pp. 40–55.
- [12] D. SCHMIDT, *Spektraleigenschaften und kanonische Fundamentallösungen linearer Differentialgleichungen bei einfachen Singularitäten*, Arch. Math., 31 (1978), pp. 302–309.
- [13] D. SCHMIDT, *Die Lösung der linearen Differentialgleichung 2. Ordnung um zwei einfache Singularitäten durch Reihen nach hypergeometrischen Funktionen*, J. Reine Angew. Math., 309 (1979), pp. 127–148.

## THE CONNECTION PROBLEM FOR TWO NEIGHBORING REGULAR SINGULAR POINTS OF GENERAL LINEAR COMPLEX ORDINARY DIFFERENTIAL EQUATIONS\*

REINHARD SCHÄFKE†

**Abstract.** A two point connection problem for local solutions at two regular singular points of a general linear ordinary differential equation is studied. Generalizing the results of a recent paper of D. Schmidt and the author [SIAM J. Math. Anal. 11 (1980), pp. 848–862] explicit formulas for the connection matrix of the fundamental solutions are derived without restrictive assumptions.

Applications to the hypergeometric equation in a Banach algebra yield new formulas for their connection factors.

**Introduction.** In a recent paper [6] D. Schmidt and the author considered the following differential equation,

$$(0.1) \quad y'(z) = \left( \frac{1}{z} A_0 + \frac{1}{z-1} A_1 + G(z) \right) y(z),$$

where  $A_0, A_1$  are complex matrices and  $G(z)$  is a matrix valued function holomorphic in a disk  $\mathfrak{R} = \{z \in \mathbb{C} \mid |z| < r\}$  with  $1 < r \leq \infty$ . (0.1) is the general system of linear complex ordinary differential equations with two “neighboring” singular points of the first kind, which are located at 0 and 1.

In this paper we discuss the connection formula

$$(0.2) \quad Y_0(z) = Y_1(z)C,$$

between characteristic fundamental solutions of (0.1) at 0 and 1, respectively, of the form,

$$(0.3) \quad Y_0(z) = H_0(z)z^{C_0}, \quad H_0(z) = \sum_{k=0}^{\infty} z^k D_k^0,$$

$$(0.4) \quad Y_1(z) = H_1(z)(1-z)^{C_1}, \quad H_1(z) = \sum_{k=0}^{\infty} (1-z)^k D_k^1.$$

Fundamental solutions of (0.1) of the form (0.3), (0.4) can always be established (see, e.g., [7] or [2, p. 120]).  $C_0, C_1$  and the coefficients  $D_k$  can be computed from  $A_0, A_1$  and from the coefficients of  $G(z)$ . In the discussion below they will be assumed as known.

The case which can be completely treated by the results in [6] is the case of diagonalizable matrices  $C_0, C_1$ .

In § 1 of this work we allow  $C_1$  to be any matrix and discuss the connection relation

$$(0.5) \quad y(z) = Y_1(z)c,$$

between (0.4) and a vector solution of (0.1) of the form

$$(0.6) \quad y(z) = z^\alpha \sum_{k=0}^{\infty} z^k d_k.$$

The results of [6, § 1] are completely generalized.

\* Received by the editors February 12, 1979, and in final revised form December 12, 1979.

† Fachbereich Mathematik, Universität Essen, Gesamthochschule, Postfach 6843, 4300 Essen 1, West Germany.

In § 2 these results will be applied to the general connection formula (0.2) with (0.3), (0.4) and arbitrary  $C_0, C_1$ . Instead of fundamental systems of logarithmic solutions we use fundamental solutions containing the exponential function of matrices  $z^A$ ; otherwise the computations would be much more complicated or impossible. Thus the gamma function of matrices also appears in a natural way. The properties of the gamma function needed in the text are proved in an appendix.

In § 3 we derive connection formulas for the hypergeometric equation in a Banach algebra,

$$y'(z) = \left( \frac{1}{z} a + \frac{1}{1-z} b \right) y(z),$$

with fewer restrictions than made by Burmann [1].

**1. The connection relation for a vector solution.** (a) Let  $E$  be a finite dimensional normed linear space,  $\mathcal{L} = \mathcal{L}(E)$  the Banach algebra of (bounded) endomorphisms of  $E$ . Then we regard (0.1) as a differential equation in  $E$  with  $A_0, A_1 \in \mathcal{L}$  and a holomorphic  $G: \mathfrak{R} \rightarrow \mathcal{L}$ . In this first section we consider the connection relation (0.5) with (0.6), (0.4). Then [6, § 1, (1.4)] can be generalized to the following theorem:

**THEOREM 1.1.** *Let*

$$y(z) = z^\alpha \sum_{k=0}^\infty z^k d_k \quad \text{with } \alpha \in \mathbb{C}, d_k \in E$$

be a solution of (0.1) at 0, and let

$$Y_1(z) = \sum_{k=0}^\infty (1-z)^k D_k^1 (1-z)^{C_1},$$

where  $C_1, D_k^1 \in \mathcal{L}$ , be a fundamental solution of (0.1) at 1. If the connection vector  $c$  is defined by

$$y(z) = Y_1(z)c \quad (z \in ]0, 1[),$$

where the powers are determined by  $\arg z = \arg(1-z) = 0$ , then for arbitrary  $m \in \mathbb{N}$ ,  $\delta > 0$ , we have

$$d_k = \sum_{i=0}^m D_i^1 \frac{1}{\Gamma}(-l - C_1) \frac{1}{\Gamma}(k + \alpha + 1) \Gamma(k + \alpha - l - C_1) c + \mathcal{O}(k^{-\gamma_- - m - 2 + \delta})$$

for  $\mathbb{N} \ni k \rightarrow \infty$ , where  $\gamma_- = \min \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(C_1) \}$ .

Here  $\sigma(C_1)$  denotes the spectrum of  $C_1$ . For the extension of the gamma function see the appendix. Here and in the sequel, e.g.,  $-l - C_1$  means  $-II - C_1$ .

*Proof.* The proof of the corresponding theorem (1.4) in [6] can be used with slight modifications. We only need two further statements. First we need an estimate for powers containing matrices,

$$|(1-z)^C| \leq M_\delta |z-1|^{\gamma_- - \delta}, \quad \gamma_- = \min \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(C) \},$$

which holds for any  $C \in \mathcal{L}$ ,  $\delta > 0$  and  $1 < z < r < \infty$ ,  $\arg(1-z) = -\pi$  or  $\pi$ , with  $M_\delta$  depending only on  $C$ ,  $\delta$  and  $r$ . Second in the proof of the analogue of [6, (1.15)], we need the formula

$$-\frac{1}{2\pi i} \int_c z^{-k-\alpha-1} (1-z)^{C_1} dz = \frac{1}{\Gamma}(k + \alpha + 1) \Gamma(k + \alpha - C_1) \frac{1}{\Gamma}(-C_1),$$

where  $\alpha \in \mathbb{C}$ ,  $C_1 \in \mathcal{L}$ ,  $k \in \mathbb{N}$  sufficiently large, and  $c$  is an (infinite) curve surrounding  $[1, \infty[$  in positive sense, but not 0. This formula will be proved in the appendix (A.5).

In the special case that the elements of  $\sigma(A_1)$  do not differ by nonzero integers, it is known ([2, p. 119]) that in (1.1)  $C_1 = A_1$  and  $D_0^1 = I$  may be chosen. In this case the limit formula [6, (1.7)] can be applied without any difficulty:

**THEOREM 1.2.** *Suppose that the elements of  $\sigma(A_1)$  do not differ by nonzero integers, and that the assumptions of (1.1) are satisfied with  $C_1 = A_1$ ,  $D_0^1 = I$ . Define*

$$\gamma_+ = \max \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(A_1) \}, \quad \gamma_- = \min \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(A_1) \},$$

and choose an integer  $m > \gamma_+ - \gamma_- - 1$ . If we define

$$D_k = D_0^1 + \sum_{l=1}^m D_l^1 \prod_{\sigma=1}^l [(-\sigma - A_1)(k + \alpha - \sigma - A_1)^{-1}] \in \mathcal{L}$$

for sufficiently large  $k \in \mathbb{N}$ , then

$$\frac{1}{\Gamma}(-A_1)c = \lim_{k \rightarrow \infty} \Gamma(k + \alpha + 1) \frac{1}{\Gamma}(k + \alpha - A_1) D_k^{-1} d_k,$$

the convergence being as  $\mathcal{O}(k^{\gamma_+ - \gamma_- - m - 1 + \delta})$ ,  $\delta > 0$  arbitrary.

*Proof.* Using the functional equation of the gamma function, Theorem A.2, Theorem 1.1 can be written

$$(1.3) \quad d_k = D_k \frac{1}{\Gamma}(k + \alpha + 1) \Gamma(k + \alpha - A_1) \frac{1}{\Gamma}(-A_1)c + \mathcal{O}(k^{-\gamma_- - m - 2 + \delta}),$$

for  $\mathbb{N} \ni k \rightarrow \infty$ ,  $\delta > 0$  arbitrary. Now  $\lim_{k \rightarrow \infty} D_k = I$ ; thus  $D_k$  is invertible for sufficiently large  $k$  and  $D_k^{-1}$  is bounded as  $k \rightarrow \infty$ . From the asymptotic formula (A.6) for quotients of gamma functions we know that

$$\Gamma(k + \alpha + 1) \frac{1}{\Gamma}(k + \alpha - A_1)$$

behaves like  $k^{A_1+1}$  for  $k \rightarrow \infty$ . Since

$$|k^{A_1+1}| \leq M_\delta k^{\gamma_+ + 1 + \delta}$$

for  $k \rightarrow \infty$  with a  $M_\delta$  depending only on  $A_1$  and  $\delta > 0$ , multiplication of (1.3) by  $D_k^{-1}$  and  $\Gamma(k + \alpha + 1)(1/\Gamma)(k + \alpha - A_1)$  yields the desired formula for  $c$ .

The convergence in Theorem 1.2 will be better, the larger  $m$  is chosen. The evaluation of  $D_k$  becomes in general more extensive, however.

From Lemma A.3 we see that  $(1/\Gamma)(-A_1)$  is invertible iff  $A_1$  has no nonnegative integer eigenvalue. Then by Theorem 1.2 the connection vector  $c$  itself can be calculated. If  $A_1$  has a nonnegative integer eigenvalue,  $c$  can be computed after a preliminary transformation of (0.1),  $y(z) = (1 - z)^\nu \tilde{y}(z)$ , with some suitable  $\nu \in \mathbb{C}$ , changing  $A_1$  into  $A_1 - \nu$ .

In the special case that  $\gamma_+ - \gamma_- < 1$ , i.e., if

$$\max \{ \operatorname{Re} (\gamma_1 - \gamma_2) \mid \gamma_1, \gamma_2 \in \sigma(A_1) \} < 1,$$

the limit formula of (1.2) reduces to

$$\frac{1}{\Gamma}(-A_1)c = \lim_{k \rightarrow \infty} k^{A_1+1} d_k.$$

In Theorems 1.1 and 1.2 it is not necessary that  $E$  be finite dimensional. The proofs remain valid if  $E$  is an arbitrary Banach space.

PROPOSITION 1.4. *Theorems 1.1 and 1.2 are valid if  $E$  is an arbitrary Banach space.*

For singular differential equations of the first kind in Banach spaces, it is known that a solution  $y(z)$  of (0.1) at 0 of the form assumed in Theorems 1.1 and 1.2 can be constructed from any  $\alpha \in \mathbb{C}$  and nonzero  $d_0 \in E$ , such that  $(A_0 - \alpha)d_0 = 0$  and  $\alpha + 1, \alpha + 2, \dots$  are not elements of  $\sigma(A_0)$ . Then the  $d_k$  may be determined recursively.

A fundamental solution  $Y_1(z)$  of (0.1) at 1 of the form assumed in Theorems 1.1 and 1.2 exists if the elements of  $\sigma(A_1)$  do not differ by positive integers. Then, as in the matrix case,  $C_1 = A_1$  and  $D_0^1 = I$  may be chosen (see [4, § 6.5]).

(b) If we drop the assumption  $D_0^1 = I$  (i.e., the assumption on  $A_1$ ) in Theorem 1.2,  $D_k^{-1}$  may be unbounded or may not exist, and it is more difficult to derive a limit formula for  $c$  from Theorem 1.1. Using a result of D. Schmidt [7] concerning the precise structure of the fundamental solution at a singular point of the first kind, the following generalization of Theorem 1.2 can be proved:

THEOREM 1.5. *Suppose the assumptions of Theorem 1.1 are satisfied with a  $C_1 \in \mathcal{L}$  such that*

$$\sigma(C_1) = \{\lambda \in \sigma(A_1) \mid (\lambda - \mathbb{N}) \cap \sigma(A_1) = \emptyset\}.$$

Define

$$\hat{\gamma} = \max \{\operatorname{Re}(\gamma_1 - \gamma_2) \mid \gamma_1, \gamma_2 \in \sigma(C_1)\}$$

$$d = \max \{l \in \mathbb{Z} \mid l = \alpha_1 - \alpha_2 \text{ where } \alpha_j \in \sigma(A_1)\}.$$

Choose an integer  $m \geq d$  such that  $m > \hat{\gamma} + d - 1$ , and define

$$D_k = D_0^1 = \sum_{l=1}^m D_l^1 \prod_{\sigma=1}^l [(-\sigma - C_1)(k + \alpha - \sigma - C_1)^{-1}] \in \mathcal{L}.$$

If we assume that  $C_1 + \sigma I$  is nonsingular for  $\sigma = 1, \dots, d$ , then

$$\frac{1}{\Gamma}(-C_1)c = \lim_{k \rightarrow \infty} \Gamma(k + \alpha + 1) \frac{1}{\Gamma}(k + \alpha - C_1) D_k^{-1} d_k,$$

the convergence being  $\mathcal{O}(k^{\hat{\gamma} + d - 1 - m + \delta})$ ,  $\delta > 0$  arbitrary.

Before we prove Theorem 1.5 let us discuss the condition on  $C_1$  (i.e., on  $Y_1(z)$ ). A fundamental solution  $Y_1$  of (0.1) at 1 with  $C_1$  satisfying the hypothesis of Theorem 1.5 can always be established by the procedure of [2, pp. 119ff], or as in [7] or [3, XIV, § 10]. Suppose now that  $E = \mathbb{C}^n$ , and that a fundamental solution

$$Y_1(z) = \sum D_k^1 (1-z)^k (1-z)^{C_1}$$

of (0.1) at 1 is given where  $C_1$  does not satisfy the condition in Theorem 1.5. Then we can apply Theorem 1.5 if we proceed as follows (proof omitted):

If  $J$  is a Jordan canonical form of  $C_1$ , and  $C_1 = TJT^{-1}$  with some regular  $T$ , we can write  $J = \tilde{J} + N$ , where

$$\sigma(\tilde{J}) = \{\lambda \in \sigma(A_1) \mid (\lambda - \mathbb{N}) \cap \sigma(A_1) = \emptyset\},$$

and  $N$  is a diagonal matrix commuting with  $J$  whose entries are integers. Then we can write

$$Y_1(z)T = H(z)(1-z)^{\tilde{J}}, \quad H(z) = \sum_{k=0}^{\infty} D_k^1 T (1-z)^k (1-z)^N.$$



By insertion into (0.1) we see that  $H(z)$  must be holomorphic at  $z = 1$ . The  $c$  in  $y(z) = Y_1(z)c$  can be determined by Theorem 1.5 applied on

$$y(z) = H(z)(1 - z)^J(T^{-1}c).$$

Sometimes instead of a fundamental solution containing the exponential function of matrices, a fundamental system of logarithmic vector solutions is given. In this case Theorem 1.5 can be applied in the following way (proof omitted):

Suppose for a moment that

$$y(z) = (1 - z)^\lambda \sum_{j=0}^r \frac{1}{j!} [\log(1 - z)]^j h_j(z),$$

where  $r \in \mathbb{N}$  and  $h_0(z), \dots, h_r(z)$  are vector functions holomorphic at 1, is a solution of (0.1). By insertion in (0.1) we see that  $\lambda + k_0$  is an eigenvalue of  $A_1$  if  $k = k_0$  denotes the smallest integer such that  $h_j^{(k)}(1) \neq 0$  for some  $j \in \{0, \dots, r\}$ , and that

$$\tilde{y}(z) = (1 - z)^\lambda \sum_{j=0}^{r-1} \frac{1}{j!} [\log(1 - z)]^j h_{j+1}(z)$$

is a second solution of (0.1).

Then clearly any given fundamental system of logarithmic solutions can be normalized such that  $\tilde{y}(z)$  appears in it whenever  $y(z)$  does. Then it has the form  $\{y_{jl}\}_{l=0, \dots, r_j-1}^{j=1, \dots, m}$ , where the  $y_{jl}$  can be written

$$y_{jl}(z) = (1 - z)^{\lambda_j} \sum_{\nu=0}^l \frac{1}{\nu!} [\log(1 - z)]^\nu h_{j,l-\nu}(z),$$

with  $\lambda_j \in \sigma(A_1)$  such that  $(\lambda_j - \mathbb{N}) \cap \sigma(A_1) = \emptyset$  and  $h_{jl}(z)$  is holomorphic at 1.

If now  $Y_1(z)$  denotes the fundamental solution of (0.1) whose columns are the  $y_{jl}(z)$ ,  $H(z)$  is the matrix whose columns are the  $h_{jl}(z)$ , and  $J$  is the matrix in upper Jordan canonical form with  $m$  Jordan blocks of length  $r_1, \dots, r_m$  and diagonal entries  $\lambda_1, \dots, \lambda_m$ , then

$$Y_1(z) = H(z)(1 - z)^J$$

satisfies the assumption in Theorem 1.5. Thus the connection coefficients  $\gamma_{jl}$  in a formula

$$y(z) = \sum \gamma_{jl} y_{jl}(z)$$

can be calculated by Theorem 1.5 applied on

$$y(z) = Y_1(z)c,$$

where  $c$  is the vector of the  $\gamma_{jl}$ .

We close this section with the proof of Theorem 1.5.

*Proof of Theorem 1.5.* As in the proof of Theorem 1.2 we can write Theorem 1.1 in the form

$$(1.6) \quad d_k = D_k \Gamma(k + \alpha - C_1) \frac{1}{\Gamma}(k + \alpha + 1)c + \mathcal{O}(k^{-\gamma_- - 2 - m + \delta}),$$

where  $\gamma_- = \min \{\operatorname{Re} \gamma \mid \gamma \in \sigma(C_1)\}$ . Later it will be shown that under the conditions of the theorem,

$$(1.7) \quad \begin{aligned} D_k &\text{ is nonsingular for sufficiently large } k \in \mathbb{N}, \\ D_k^{-1} &= \mathcal{O}(k^d) \quad \text{as } k \rightarrow \infty. \end{aligned}$$

This is in fact the only difficult part of the proof. Now (again as in the proof of Theorem 1.2),

$$\Gamma(k + \alpha + 1) \frac{1}{\Gamma}(k + \alpha - C_1) = \mathcal{O}(k^{\gamma_+ + 1 + \delta}),$$

where  $\gamma_+ = \max \{\operatorname{Re} \gamma \mid \gamma \in \sigma(C_1)\}$ . Now multiplication of (1.6) by  $D_k^{-1}$  and  $\Gamma(k + \alpha + 1)(1/\Gamma)(k + \alpha - C_1)$  yields the desired result.

For the proof of (1.7) we assume that  $E = \mathbb{C}^n$  for some  $n \in \mathbb{N}$ ; this can be achieved by an inessential linear isomorphism. We assume further that  $A_1$  is in Jordan canonical form. This can be achieved by a preliminary transformation  $y(z) = T\check{y}(z)$  of (0.1) with a nonsingular  $T \in \mathcal{L}$ , such that  $T^{-1}A_1T$  is in Jordan canonical form. Such a transformation is inessential because only the  $D_k^1, D_k$  and  $c_k$  are multiplied by  $T^{-1}$  from the left.

If  $A_1$  is in upper Jordan canonical form and if the Jordan blocks are appropriately ordered, we can state the main result of [7] as follows (see also [3, vol. II, pp. 163ff]).

RESULT 1.8. Equation (0.1) has a fundamental solution at 1 of the form

$$\check{Y}_1(1 - z) = \sum_{k=0}^{\infty} H_k z^k z^D z^B$$

where  $H_0 = I, D = \operatorname{diag}(l_1, \dots, l_n)$  with integers  $d = l_1 \geq l_2 \geq \dots \geq l_n = 0$ , and  $B$  is an upper triangular matrix satisfying

$$\sigma(B) = \{\lambda \in \sigma(A_1) \mid (\lambda - \mathbb{N}) \cap \sigma(A_1) = \emptyset\}.$$

Now with a nonsingular  $T \in \mathcal{L}$  we have

$$(1.9) \quad \check{Y}_1(1 - z) = Y_1(1 - z)T.$$

If we replace  $z$  by  $z e^{2\pi i}$  (i.e., if  $1 - z$  surrounds 1 once in positive direction) we get

$$\begin{aligned} \check{Y}_1(1 - z e^{2\pi i}) &= \check{Y}_1(1 - z) \exp(2\pi i B), \\ Y_1(1 - z e^{2\pi i}) &= Y_1(1 - z) \exp(2\pi i C_1), \end{aligned}$$

and conclude that

$$\exp(2\pi i B) = T^{-1} \exp(2\pi i C_1) T = \exp(2\pi i T^{-1} C_1 T).$$

Using the fact that  $\sigma(B) = \sigma(T^{-1} C_1 T)$  and that  $\lambda - \mu \notin \mathbb{N}$  for  $\lambda, \mu \in \sigma(B)$ , we obtain (see [3, vol. I, pp. 239ff, pp. 220ff])

$$(1.10) \quad B = T^{-1} C_1 T.$$

Now from (1.9) we see that the power series parts  $Y_1(1 - z)z^{-B}$  and  $Y_1(1 - z)z^{-C_1}T$  are equal, and thus

$$(1.11) \quad \sum_{\substack{j=1 \\ l_j \leq l}}^n H_{l-l_j} E_j = D_l^1 T \quad (l = 0, 1, \dots),$$

where  $E_j = \operatorname{diag}(0, \dots, 0, 1, 0, \dots, 0)$  with 1 in  $j$ th position. By (1.10) and (1.11) we can now express the  $D_k$ :

$$D_k T = \sum_{j=1}^n \sum_{s=0}^{m-l_j} H_s E_j \prod_{\sigma=1}^{s+l_j} [(-\sigma - B)(k + \alpha - \sigma - B)^{-1}].$$

This can be written as

$$D_k T = \sum_{j=1}^n \sum_{s=0}^{m-l_j} H_s E_j k^{-s-l_j} L_{sj}(k),$$

where the  $L_{sj}(k)$  are upper triangular matrices such that  $L_{sj}(\infty) = \lim_{k \rightarrow \infty} L_{sj}(k)$  exists and is nonsingular iff all  $\sigma + C_1$  ( $\sigma = 1, \dots, s + l_j$ ) are nonsingular. By assumption this is always true if  $s + l_j \leq d$ . Now we see that

$$D_k T k^D = \sum_{j=1}^n \sum_{s=0}^{m-l_j} H_s E_j k^{-s} \tilde{L}_{sj}(k),$$

where the  $\tilde{L}_{sj}(k) = k^{-D} L_{sj}(k) k^D$  have the same properties as the  $L_{sj}(k)$ . From this we claim that the limit

$$(1.12) \quad \lim_{k \rightarrow \infty} D_k T k^D = \sum_{j=1}^n E_j \tilde{L}_{0j}(\infty)$$

is nonsingular since all  $\tilde{L}_{0j}(\infty)$  are lower triangular and nonsingular. (1.12) immediately yields (1.7).

**2. The connection formula for two fundamental solutions.** Let  $E$  be a finite dimensional normed linear space and  $\mathcal{L}_1 = \mathcal{L}(E)$ ,  $\mathcal{L}_2 = \mathcal{L}(\mathcal{L}_1)$  the Banach algebras of endomorphisms of  $E$  and  $\mathcal{L}_1$ , respectively. We regard (0.1) as a differential equation in  $E$  with  $A_0, A_1 \in \mathcal{L}_1$  and a holomorphic  $G: \mathfrak{R} \rightarrow \mathcal{L}_1$ . Then we would like to have formulas for the connection matrix  $C$  in

$$Y_0(z) = Y_1(z)C, \quad z \in ]0, 1[,$$

between fundamental solutions of (0.1) of the form

$$Y_0(z) = H_0(z)z^{C_0}, \quad H_0(z) = \sum_{k=0}^{\infty} z^k D_k^0,$$

$$Y_1(z) = H_1(z)(1-z)^{C_1}, \quad H_1(z) = \sum_{k=0}^{\infty} (1-z)^k D_k^1,$$

at 0 and 1, respectively.

Here we cannot proceed in the same way as in [6, § 1] because we would need a formula for

$$\int_c (1-z)^{C_1} C z^{-C_0-1} dz,$$

where  $c$  is an infinite curve surrounding  $[1, \infty[$  in positive sense, but not 0, and  $C, C_0, C_1$  are arbitrary elements of  $\mathcal{L}_1$ . Such a formula does not seem to exist, if  $C, C_0$  and  $C_1$  do not commute. Thus we shall transform the problem so that the conditions of § 1 hold and all theorems can then be carried over.

$Y_0$  and  $Y_1$  are solutions of the matrix differential equation corresponding to (0, 1),

$$Y'(z) = \left( \frac{1}{z} A_0 + \frac{1}{z-1} A_1 + G(z) \right) Y(z).$$

This linear ordinary differential equation in  $\mathcal{L}_1$  is transformed by

$$Y(z) = H(z)z^{C_0}$$

into a new linear differential equation in  $\mathcal{L}_1$ ,

$$(2.1) \quad H'(z) = \left( \frac{1}{z} \hat{A}_0 + \frac{1}{z-1} \hat{A}_1 + \hat{G}(z) \right) H(z),$$

where  $\hat{A}_0, \hat{A}_1 \in \mathcal{L}_2$  and  $\hat{G}: \mathfrak{R} \rightarrow \mathcal{L}_2$  are defined by

$$\begin{aligned} \hat{A}_0 X &= A_0 X - X C_0 \\ \hat{A}_1 X &= A_1 X & (X \in \mathcal{L}_1). \\ \hat{G}(z) X &= G(z) X \end{aligned}$$

Here and in the following  $\hat{A}, \hat{B}, \dots$  denote elements of  $\mathcal{L}_2$ . With  $E$  and  $\mathcal{L}$  replaced by  $\mathcal{L}_1$  and  $\mathcal{L}_2$ , we have the situation of Theorem 1.1 applied for (2.1):

$$H_0(z) = \sum_{k=0}^{\infty} z^k D_k^0$$

is a solution of (2.1) at 0 (with  $\alpha = 0$ ). A fundamental solution of (2.1) at 1 is given by

$$\hat{Y}_1(z) X = H_1(z) (1-z)^{C_1} X z^{-C_0} \quad (X \in \mathcal{L}_1),$$

because we can write  $\hat{Y}_1$  in the form

$$\hat{Y}_1(z) = \hat{H}_1(z) (1-z)^{\hat{C}_1},$$

where

$$\begin{aligned} \hat{C}_1 X &= C_1 X \\ \hat{H}_1(z) X &= H_1(z) X z^{-C_0} & (X \in \mathcal{L}_1), \end{aligned}$$

and  $\hat{H}_1(z)$  is holomorphic. From

$$z^{-C_0} = \sum_{l=0}^{\infty} \binom{C_0+l-1}{l} (1-z)^l,$$

we get the power series for  $\hat{H}_1$  at 1,

$$\hat{H}_1(z) = \sum_{k=0}^{\infty} (1-z)^k \hat{D}_k^1,$$

where

$$\hat{D}_k^1 X = \sum_{l=0}^k D_{k-l}^1 X \binom{C_0+l-1}{l} \quad (X \in \mathcal{L}_1).$$

Now the connection relation  $Y_0(z) = Y_1(z)C$  reduces to

$$H_0(z) = \hat{Y}_1(z)C.$$

Thus we have in fact the situation of Theorem 1.1 with (0.1),  $E, \mathcal{L}, y, \alpha, d_k, Y_1, D_k^1, C_1, c$  replaced by (2.1),  $\mathcal{L}_1, \mathcal{L}_2, H_0, 0, D_k^0, \hat{Y}_1, \hat{D}_k^1, \hat{C}_1, C$  respectively.

Now not only Theorem 1.1, but also Theorems 1.2 and 1.5, can be translated into theorems for the matrix connection formula. Thus the following theorem is proved.

**THEOREM 2.2.** *Suppose that fundamental solutions  $Y_0(z)$  of (0.1) at 0 and  $Y_1(z)$  of (0.1) at 1 are given in the form*

$$Y_0(z) = \sum_{k=0}^{\infty} z^k D_k^0 z^{C_0},$$

$$Y_1(z) = \sum_{k=0}^{\infty} (1-z)^k D_k^1 (1-z)^{C_1},$$

where the  $C_j$  and  $D_k^j$  are complex matrices. Let the connection matrix  $C$  between them be defined by

$$Y_0(z) = Y_1(z)C, \quad z \in ]0, 1[,$$

where the powers are determined by  $\arg z = \arg(1-z) = 0$ . Finally define  $\gamma_+$  and  $\gamma_-$  by

$$\gamma_+ = \max \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(C_1) \}, \quad \gamma_- = \min \{ \operatorname{Re} \gamma \mid \gamma \in \sigma(C_1) \}.$$

(i) For arbitrary  $m \in \mathbb{N}$ ,  $\delta > 0$  we have as  $k \rightarrow \infty$ ,

$$D_k^0 = \sum_{l=0}^m \sum_{\nu=0}^l D_{l-\nu}^1 \frac{1}{\Gamma}(-l-C_1) \frac{1}{k!} \Gamma(k-l-C_1) C \binom{C_0+\nu-1}{\nu} + \mathcal{O}(k^{-\gamma_- - m - 2 + \delta}).$$

Now define  $\hat{D}_m(k) \in \mathcal{L}(\mathcal{L}_1)$  by

$$\hat{D}_m(k)X = D_0^1 X + \sum_{l=1}^m \sum_{\nu=0}^l D_{l-\nu}^1 \prod_{\sigma=1}^l [(-\sigma-C_1)(k-\sigma-C_1)^{-1}] X \binom{C_0+\nu-1}{\nu}.$$

(ii) If under the above conditions the elements of  $\sigma(A_1)$  do not differ by nonzero integers and  $C_1 = A_1$ ,  $D_0^1 = I$ , we have for integers  $m > \gamma_+ - \gamma_- - 1$ ,

$$\frac{1}{\Gamma}(-A_1)C = \lim_{k \rightarrow \infty} k! \frac{1}{\Gamma}(k-A_1) \hat{D}_m(k)^{-1} D_k^0,$$

the convergence being as  $\mathcal{O}(k^{\gamma_+ - \gamma_- - m - 1 + \delta})$ ,  $\delta > 0$  arbitrary.

(iii) Suppose  $C_1$  satisfies

$$\sigma(C_1) = \{ \lambda \in \sigma(A_1) \mid (\lambda - \mathbb{N}) \cap \sigma(A_1) = \emptyset \},$$

and define the integer  $d$  by

$$d = \max \{ \lambda_1 - \lambda_2 \mid \lambda_1, \lambda_2 \in \sigma(A_1), \lambda_1 - \lambda_2 \in \mathbb{Z} \}.$$

If  $\sigma + C_1$  is nonsingular for  $\sigma = 1, \dots, d$  then we have, for integers  $m \geq d$  such that  $m > \gamma_+ - \gamma_- + d - 1$  and real  $\delta > 0$ ,

$$\frac{1}{\Gamma}(-C_1)C = \lim_{k \rightarrow \infty} k! \frac{1}{\Gamma}(k-C_1) \hat{D}_m(k)^{-1} D_k^0,$$

the convergence being as  $\mathcal{O}(k^{\gamma_+ - \gamma_- + d - m - 1 + \delta})$ .

The remarks of § 1 concerning better convergence versus more computation, the possibility of obtaining  $C$  and not only  $(1/\Gamma)(-C_1)C$ , and methods of finding fundamental solutions of the required form, remain valid here. We state explicitly the following two remarks.

(1) In the special case  $\gamma_+ - \gamma_- < 1$  we choose  $m = 0$  in Theorem 2.2 (ii) and obtain using Theorem A.6

$$\frac{1}{\Gamma}(-A_1)C = \lim_{k \rightarrow \infty} k^{A_1+1} D_k^0.$$

(2) In Theorem 2.2, (i), (ii) it is not necessary that we have a system (0, 1) of differential equations. Proposition 1.4 and the proof of Theorem 2.2 show that we may replace (0.1) by a differential equation in any Banach algebra  $\mathcal{B}$  with unit element  $e$ ,

$$Y'(z) = \left( \frac{1}{z} A_0 + \frac{1}{z-1} A_1 + G(z) \right) Y(z),$$

where now  $A_0, A_1 \in \mathcal{B}$  and  $G: \mathbb{K} \rightarrow \mathcal{B}$  is holomorphic. Then the  $C_j$  and  $D_k^j$  are elements of  $\mathcal{B}$ .

We recall that such solutions at  $\lambda = 0$  or  $1$  exist if the elements of  $\sigma(A_\lambda)$  do not differ by nonzero integers. Then  $C_\lambda = A_\lambda$  and  $D_0^\lambda = e$  may be chosen and the  $D_k^\lambda$  are determined recursively (see [4, § 6.5]).

**3. The hypergeometric equation in a Banach algebra.** Now consider the hypergeometric equation in a Banach algebra  $\mathcal{B}$  with unit element  $e$ ,

$$(3.1) \quad y'(z) = \left( \frac{1}{z} a + \frac{1}{1-z} b \right) y(z).$$

The basic properties of (3.1) can be found in Hille [4, pp. 240–245]. If we assume that

$$(3.2) \quad (\sigma(a) - \sigma(b)) \cap \mathbb{N} = \emptyset$$

then (3.1) has a solution  $w(a, b; z)$  defined for  $|z| < 1$ :

$$(3.3) \quad w(a, b; z) = \sum_{k=0}^{\infty} z^k d_k(a, b) \cdot z^a.$$

The coefficients  $d_k(a, b)$  can be computed recursively:

$$(3.4) \quad \begin{aligned} d_0(a, b) &= e, \\ (k - C_a) d_k(a, b) &= (L_b + k - 1 - C_a) d_{k-1}(a, b). \end{aligned}$$

Here  $L_b \in \mathcal{L}(\mathcal{B})$  denotes left multiplication by  $b$ , and  $C_a \in \mathcal{L}(\mathcal{B})$  the commutator operator  $C_a x = ax - xa$ . If we furthermore assume that

$$(3.5) \quad (\sigma(b) - \sigma(b)) \cap \mathbb{N} = \emptyset,$$

then the transformation  $y(z) = \tilde{y}(1-z)$  shows that  $w(-b, -a; 1-z)$  is a solution of (3.1) as well.  $w(-b, -a; 1-z)$  is defined for  $|z-1| < 1$ . Hence there is a connection relation

$$(3.6) \quad w(a, b; z) = w(-b, -a; 1-z) c(a, b)$$

for  $z \in \mathbb{C}$  satisfying  $|z|, |1-z| < 1$  and  $\arg z, \arg(1-z) \in ]-\pi, \pi[$ , with a unique  $c(a, b) \in \mathcal{B}$  called the connection factor. Now (3.1) has the form required in Remark 2 below Theorem 2.2, with  $A_0 = a$  and  $A_1 = -b$ . Thus Theorem 2.2 (ii) can be applied and we obtain

**THEOREM 3.7.** For  $m \in \mathbb{N}$ ,  $m+1 > \mu(b) := \sup \{ \operatorname{Re}(\beta_1 - \beta_2) \mid \beta_1, \beta_2 \in \sigma(b) \}$ , and for sufficiently large  $k \in \mathbb{N}$  let

$$\hat{D}_m(a, b, k) X = \sum_{l=0}^m \sum_{\kappa=0}^l d_{l-\kappa}(-b, -a) \prod_{\rho=1}^l [(b-\rho)(b+k+\rho)^{-1}] X \binom{a+\kappa-1}{\kappa}.$$

Then for any  $\delta > 0$  and for  $\mathbb{N} \ni k \rightarrow \infty$ ,

$$\Gamma(k+1) \frac{1}{\Gamma(k+b)} \hat{D}_m(a, b, k)^{-1} d_k(a, b) = \frac{1}{\Gamma(b)} c(a, b) + \mathcal{O}(k^{\mu(b)-m-1+\delta}).$$

If  $\mu(b) < 1$  we can choose  $m = 0$  in the above theorem and get

$$\lim_{k \rightarrow \infty} k^{-b+1} d_k(a, b) = \frac{1}{\Gamma}(b)c(a, b).$$

If  $\sigma(-b)$  does not contain any nonnegative integer,  $c(a, b)$  can be determined from Theorem 3.7. If  $\sigma(a)$  does not contain any nonnegative integer,  $c(a, b)^{-1} = c(-b, -a)$  can be computed by means of Theorem 3.7. In some cases  $c(a, b)$  can be determined by using the property

$$c(a, b) = c(a + \alpha e, b + \beta e) \quad (\alpha, \beta \in \mathbb{C})$$

and by application of Theorem 3.7 to the right side.

By using transformations

$$y(z) = \tilde{y}(\varphi(z)), \quad \varphi(z) = z, \frac{1}{z}, 1 - z, \dots,$$

the connection factors not computed in Hille [4, pp. 244ff] can be expressed by some  $c(\tilde{a}, \tilde{b})$ , and hence can be determined by Theorem 3.7. This will be done here only for the connection factor which was determined by Burmann [1] in a different way.

Burmann assumes

$$(\sigma(a) - \sigma(a)) \cap \mathbb{N}_1 = \emptyset \quad \text{and} \quad (\sigma(b - a) - \sigma(b - a)) \cap \mathbb{N}_1 = \emptyset.$$

Then (3.1) has the solutions  $w(a, b; z)$  and  $w(b - a, b; 1/z)$ , which are defined and single valued in the regions  $|z| < 1, \arg z \in ]0, 2\pi[$ , and  $|z| > 1, \arg z \in ]0, 2\pi[$  respectively. These solutions can be analytically continued to solutions  $y_0$  and  $y_\infty$  of (3.1) in  $\mathbb{C}_s = \mathbb{C} - \mathbb{R}_+$ . Burmann considers the connection formula

$$y_0(z) = y_\infty(z)q(a, b) \quad (z \in \mathbb{C}_s).$$

In order to express  $q(a, b)$  in terms of  $c(\cdot, \cdot)$ , we transform  $y(z) = \tilde{y}(z/(z - 1))$  and obtain from (3.6)

$$(3.8) \quad w\left(a, a - b; \frac{z}{z - 1}\right) = w\left(b - a, -a; \frac{1}{1 - z}\right)c(a, a - b)$$

for  $z \in \mathbb{C}$  satisfying  $\operatorname{Re} z < \frac{1}{2}, |z - 1| > 1$  and

$$\arg\left(\frac{z}{z - 1}\right) \in ]-\pi, 0[, \quad \arg\left(\frac{1}{1 - z}\right) \in ]0, \pi[.$$

Since in (3.8) solutions at 0 and  $\infty$  are connected, we get, together with connection factors determined by the method of Hille [4, pp. 244ff],

$$q(a, b) = \exp(\pi i(b - a))c(a, a - b) \exp(\pi ia).$$

Now from (3.7) we have a formula for  $q(a, b)$  which is valid if weaker conditions than those of Burmann [1, Thm. 9] are satisfied.

**Appendix. Extension of the reciprocal gamma function.** Let  $\mathcal{B}$  be a Banach algebra with unit element  $e$ . (The reader not interested in Banach algebras may replace  $\mathcal{B}$  by the algebra of  $n \times n$  matrices.)

If  $f(z)$  is an entire function,  $f(z) = \sum_{l=0}^{\infty} \alpha_l z^l$  with complex  $\alpha_l$ , we define  $f: \mathcal{B} \rightarrow \mathcal{B}$  by

$$f(a) = \alpha_0 e + \sum_{l=1}^{\infty} \alpha_l a^l \quad (a \in \mathcal{B}).$$

Then  $f \rightarrow f(a)$ ,  $a \in \mathcal{B}$  fixed, is linear and multiplicative; it is continuous in the following sense.

**LEMMA A.1.** *Let  $f_n(z), f(z)$  be entire functions and suppose  $f_n(z)$  converges to  $f(z)$  uniformly on  $\{z \in \mathbb{C} \mid |z| \leq K\}$  for arbitrary  $K > 0$ . Then  $f_n(a)$  converges to  $f(a)$  uniformly on  $\{a \in \mathcal{B} \mid |a| \leq K\}$  for arbitrary  $K > 0$ .*

*Proof.* Estimate the difference between the coefficients of  $f_n$  and  $f$  by Cauchy's formula.

Now we study the extension of the entire function  $(1/\Gamma)(z)$  on  $\mathcal{B}$  and state the results needed in the text.

**THEOREM A.2.**

$$(i) \quad \frac{1}{\Gamma}(a) = \exp(\gamma a) a \prod_{n=1}^{\infty} \left[ \left( e + \frac{1}{n} a \right) \exp\left(-\frac{1}{n} a\right) \right] \quad (a \in \mathcal{B}),$$

where

$$\gamma = \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{2} + \dots + \frac{1}{n} - \log n \right).$$

(ii) If  $(1/\Gamma)^{(m)}$  denotes the  $m$ -th derivative of the (complex)  $1/\Gamma$ ,

$$\frac{1}{\Gamma}(a) = \sum_{m=0}^{\infty} \frac{1}{m!} \left( \frac{1}{\Gamma} \right)^{(m)}(\lambda) (a - \lambda e)^m \quad (a \in \mathcal{B}, \lambda \in \mathbb{C}).$$

(iii) 
$$\frac{1}{\Gamma}(a) = a \frac{1}{\Gamma}(a + e) \quad (a \in \mathcal{B}).$$

(iv) If  $a, b \in \mathcal{B}$  and  $b$  is invertible then

$$\frac{1}{\Gamma}(b^{-1}ab) = b^{-1} \frac{1}{\Gamma}(a)b.$$

*Proof.* (iv) follows from the definition of  $1/\Gamma(a)$  by a power series with complex coefficients. (i) and (ii) are applications of Lemma A.1. The compact convergence in  $\mathbb{C}$  is trivial for (ii), and is shown in [5, ch. 2, § 1.4] for (i). (iii) is an application of Lemma A.1, too, because

$$\frac{1}{\Gamma}(a) = a \lim_{n \rightarrow \infty} \sum_{m=0}^n \frac{1}{m!} \left( \frac{1}{\Gamma} \right)^{(m)}(0) (a + e)^m \quad (a \in \mathcal{B})$$

is to be shown.

From Theorem A.2, (i) we conclude immediately

**LEMMA A.3.**  $1/\Gamma(a)$  is invertible in  $\mathcal{B}$  iff  $\sigma(a) \cap (-\mathbb{N}_0) = \emptyset$ . Then define  $\Gamma(a) = [1/\Gamma(a)]^{-1}$ .

If  $\mathcal{B}$  is a matrix algebra the reciprocal gamma function can be expressed by the scalar one and its derivatives. By Theorem A.2(iv) we can limit ourselves to matrices in Jordan canonical form.

**THEOREM A.4.** *Let  $a$  be a matrix in Jordan canonical form,*

$$a = \begin{bmatrix} J_1 & & 0 \\ & \ddots & \\ & & J_s \end{bmatrix}, \quad J_i = \begin{bmatrix} \lambda_i & 1 & 0 \\ & \ddots & \vdots \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{bmatrix} \text{ (an } r_i \times r_i \text{ matrix).}$$



Then

$$\frac{1}{\Gamma}(a) = \begin{bmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_s \end{bmatrix}, \quad H_i = \begin{bmatrix} \alpha_{i0} & \alpha_{i1} & \alpha_{i r_i} \\ & \ddots & \vdots \\ & & \alpha_{i1} \\ & & & \ddots \\ & & & & \alpha_{i0} \\ & & & & & 0 \end{bmatrix},$$

with

$$\alpha_{ij} = \frac{1}{j!} \left( \frac{1}{\Gamma} \right)^{(j)} (\lambda_i).$$

*Proof.* From the definition of  $1/\Gamma$  we see that it suffices to prove  $H_i = 1/\Gamma(J_i)$ . This follows from Theorem A.2(ii) with  $\lambda = \lambda_i$ , because  $J_i - \lambda_i$  is a (nilpotent) shift matrix.

Once more by Lemma A.1 we can carry over the formula used to prove [6, Thm. 1.15].

LEMMA A.5. For  $a \in \mathcal{B}$ ,  $\alpha \in \mathbb{C}$  and  $k \in \mathbb{N}$  sufficiently large,

$$-\frac{1}{2\pi i} \int_c z^{-k-\alpha-1} (z-1)^a dz = \frac{1}{\Gamma}(k+\alpha+1) \Gamma((k+\alpha)e-a) \frac{1}{\Gamma}(-a)$$

where  $c$  is an (infinite) curve surrounding  $[1, \infty[$  in positive sense but not 0.

Somewhat more difficult is the extension of [6, Thm. 1.5] (asymptotic series for quotients of gamma functions):

THEOREM A.6. Let  $a \in \mathcal{B}$ ,  $z \in \mathbb{C}$  with  $|\arg z| < \pi/2$ ,  $\operatorname{Re} z > |a|$ . Then for arbitrary  $m \in \mathbb{N}_1$ ,

$$\frac{1}{\Gamma}(a) \Gamma(ze+a) z^{-a} = e + \sum_{l=1}^m z^{-l} p_l(a) + z^{-m} r_m(a, z),$$

where  $r_m(a, z)$  tends to 0 for  $\operatorname{Re} z \rightarrow \infty$  and the  $p_l$  are polynomials.

*Proof.* In Stirling's series for  $\Gamma(z)$  (see [5, ch. 8, § 4.2]), we insert  $z$  and  $z + \alpha$  with  $\alpha \in \mathbb{C}$  and divide. Thus we obtain for  $m \in \mathbb{N}$ ,

$$\frac{\Gamma(z) z^\alpha}{\Gamma(z+\alpha)} = 1 + \sum_{s=1}^m \tilde{p}_s(\alpha) z^{-s} + z^{-m} \tilde{r}_m(\alpha, z)$$

where the  $\tilde{p}_l(\alpha)$  are polynomials, the  $\tilde{r}_m(\alpha, z)$  are entire functions with respect to  $\alpha$  for any fixed  $z$ ,  $|\arg z| < \pi/2$ , and  $\lim_{\operatorname{Re} z \rightarrow \infty} \tilde{r}_m(\alpha, z) = 0$  uniformly on  $|\alpha| \leq K$ ,  $K > 0$  arbitrary. Now Lemma A.1 yields the desired result.

REFERENCES

[1] H. W. BURMANN, *Zur hypergeometrischen Differentialgleichung in Banachalgebren*, Math. Z., 125 (1972), pp. 139-176.  
 [2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.  
 [3] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1971.  
 [4] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading MA, 1969.  
 [5] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.  
 [6] R. SCHÄFKE AND D. SCHMIDT, *The connection problem for general linear ordinary differential equations at two regular singular points with applications in the theory of special functions*, this Journal, this issue, pp. 848-862.  
 [7] D. SCHMIDT, *Spektraleigenschaften und kanonische Fundamentallösungen linearer Differentialgleichungen bei einfachen Singularitäten*, Arch. Math., 31 (1978), pp. 302-309.

## INEQUALITIES FOR ULTRASPHERICAL AND LAGUERRE POLYNOMIALS. II\*

J. BUSTOZ† AND N. SAVAGE†

**Abstract.** The main result proved here is the inequality  $(n+1)F_n^\alpha(x)F_n^\beta(x) - nF_{n+1}^\alpha(x)F_{n-1}^\beta(x) > 0$  for  $-1 < x < 1$  and  $\frac{1}{2} \leq \alpha \leq \beta \leq \alpha + 1$ , where  $F_n^\alpha(x) = P_n^\alpha(x)/P_n^\alpha(1)$  and  $P_n^\alpha(x)$  is the ultraspherical polynomial. We discuss some other similar inequalities for ultraspherical and Laguerre polynomials.

**1. Introduction.** Let  $P_n^\lambda(x)$  denote the ultraspherical polynomial defined by the generating identity:

$$(1 - 2xz + z^2)^{-\lambda} = \sum_{n=0}^{\infty} P_n^\lambda(x)z^n.$$

In [3] we proved the inequality

$$(1.1) \quad \frac{P_n^\alpha(x)P_{n+1}^\beta(x) - P_{n+1}^\alpha(x)P_n^\beta(x)}{\beta - \alpha} > 0, \quad 0 < x < 1$$

for various values of  $\alpha$  and  $\beta$  including  $\frac{1}{2} < \beta < \alpha \leq \beta + 2$  and  $\frac{1}{2} < \alpha < \beta \leq \alpha + 2$ . We then used (1.1) with  $\beta = 1$  to prove

$$(1.2) \quad \sum_{k=0}^n \frac{(\lambda)_k (\lambda)_{n-k} \sin[(k+1)\theta]}{k! (n-k)! (k+1) \sin \theta} > 0, \quad 0 < \theta < \pi, \quad 2 < \lambda < 3.$$

R. Askey and G. Gasper [2] had proved (1.2) for  $1 < \lambda < 2$ , and Askey [1] proved (1.2) for  $\lambda = 3$ , also showing it is false for  $\lambda > 3$ . P. Turan proved the interesting inequality for Legendre polynomials

$$(1.3) \quad [P_n(x)]^2 - P_{n+1}(x)P_{n-1}(x) > 0, \quad -1 < x < 1.$$

G. Szego [9] gave four elegant proofs of (1.3) and showed that similar inequalities hold for the polynomials  $F_n^\lambda(x) = P_n^\lambda(x)/P_n^\lambda(1)$  and for Laguerre and Hermite polynomials. The generalization of (1.3) to the  $F_n^\lambda$  is

$$(1.4) \quad [F_n^\lambda(x)]^2 - F_{n+1}^\lambda(x)F_{n-1}^\lambda(x) > 0, \quad -1 < x < 1, \quad \lambda > -\frac{1}{2}.$$

Various other authors have studied inequalities for Turan type. In particular G. Gasper [5] proved such an inequality for Jacobi polynomials, and S. Karlin and G. Szego [7] generalized Turan inequalities to  $n \times n$  determinants.

These inequalities suggest the existence of inequalities of the Turan type involving two parameters. In this paper we will prove the inequality

$$(1.5) \quad (n+1)F_n^\alpha F_n^\beta - nF_{n+1}^\alpha F_{n-1}^\beta > 0, \quad -1 < x < 1, \quad \frac{1}{2} \leq \alpha \leq \beta \leq \alpha + 1.$$

When  $\beta = \alpha$ , (1.5) is weaker than (1.4) since in this case (1.5) becomes

$$n[(F_n^\alpha)^2 - F_{n+1}^\alpha F_{n-1}^\alpha] + (F_n^\alpha)^2 > 0.$$

We will also prove an inequality similar to (1.5) for Laguerre polynomials. We let  $\Delta_n(x; \alpha, \beta) = P_n^\alpha(x)P_{n+1}^\beta(x) - P_{n+1}^\alpha(x)P_n^\beta(x)$ . G. Gasper [6] proved that

$$(1.6) \quad \frac{d}{d\theta} \sum_{k=0}^n \frac{(\mu+1)_k (\mu+1)_{n-k} \sin[(k+1)\theta]}{k! (n-k)! (k+1) \sin \theta/2} < 0, \quad 0 < \theta < \pi, \quad 0 \leq \mu \leq 1.$$

\* Received by the editors February 20, 1979, and in revised form October 15, 1979.

† Department of Mathematics, Arizona State University, Tempe, Arizona, 85281.

This inequality can be rewritten as

$$\frac{d}{dx} \Delta_n(x; 1, \lambda) > 0, \quad 0 < x < 1, \quad 1 \leq \lambda \leq 2,$$

and this suggests the possibility that the determinants  $\Delta_n(x; \alpha, \beta)$  may be monotonic in  $0 < x < 1$  for certain values of  $\alpha$  and  $\beta$ . We will prove this monotonicity when  $\beta = \alpha + 1$ .

**2. Monotonicity of  $\Delta_n(x; \alpha, \alpha + 1)$ .** Set  $D_n(x; \alpha, \beta) = F_n^\alpha(x)F_{n+1}^\beta(x) - F_{n+1}^\alpha(x)F_n^\beta(x)$ . In proving that  $(d/dx) \Delta_n(x; \alpha, \alpha + 1) > 0$  and that  $(n + 1)F_n^\alpha F_n^\beta - nF_{n+1}^\alpha F_{n-1}^\beta > 0$  we will need the fact that  $D_n(x; \alpha, \beta) > 0$  for  $0 < x < 1, -\frac{1}{2} < \alpha < \beta \leq \alpha + 1$ . The proof that  $D_n(x; \alpha, \beta) > 0$  is very similar to the proof of (1.1) given in [3]; however, there are important differences and thus we will give a complete proof here. Note that in [3] we proved the inequality (1.1) under the condition  $\alpha > \frac{1}{2}$  whereas in the following we will have  $\alpha > -\frac{1}{2}$ .

By using the standard identities satisfied by ultraspherical polynomials, see [3] and [10], we can prove the differential identity

$$(2.1) \quad \frac{d}{dx} [(1 - x^2)^{\alpha-1/2} D_n(x; \alpha, \beta)] = 2(\beta - \alpha)(1 - x^2)^{\alpha-3/2} F_{n+1}^\alpha (F_{n+1}^\beta - xF_n^\beta).$$

LEMMA 2.1.  $D_n(x; \alpha, \alpha + 1) > 0$  for  $0 < x < 1, \alpha > -\frac{1}{2}$ .

*Proof.* Setting  $\beta = \alpha + 1$  in (2.1), we have

$$\frac{d}{dx} [(1 - x^2)^{\alpha-1/2} D_n] = 2(1 - x^2)^{\alpha-3/2} F_{n+1}^\alpha (F_{n+1}^{\alpha+1} - xF_n^{\alpha+1}).$$

*Case 1.* Suppose  $F_{n+1}^\alpha = 0$ . Then  $D_n = F_n^\alpha F_{n+1}^{\alpha+1}$ . But  $(n + 2\alpha + 2)(n + 2\alpha + 1)(1 - x^2)F_{n+1}^{\alpha+1} = (2\alpha + 1)\{[n + 2\alpha + 1 - 2(n + \alpha + 1)x^2]F_{n+1}^\alpha + (n + 1)nF_n^\alpha\}$ . Hence if  $F_{n+1}^\alpha = 0$  then  $D_n = F_n^\alpha F_{n+1}^{\alpha+1} > 0$  for  $0 < x < 1, \alpha > -\frac{1}{2}$ .

*Case 2.* Suppose  $F_{n+1}^{\alpha+1} - xF_n^{\alpha+1} = 0$ . Then  $D_n = F_{n+1}^{\alpha+1} (xF_n^\alpha - F_{n+1}^\alpha)$ . Using the identities

$$xF_n^\alpha - F_{n+1}^\alpha = \frac{(1 - x^2)(F_n^\alpha)'}{n + 2\alpha} \quad \text{and} \quad (F_n^\alpha)' = \frac{n + 2\alpha}{2\alpha + 1} F_{n-1}^{\alpha+1},$$

we get

$$D_n = \frac{(1 - x^2)}{2\alpha + 1} F_n^{\alpha+1} F_{n-1}^{\alpha+1}.$$

Now, use the identity  $nF_{n-1}^{\alpha+1} = 2(n + \alpha + 1)xF_n^{\alpha+1} - (n + 2\alpha + 2)F_{n+1}^{\alpha+1}$  which, since  $F_{n+1}^{\alpha+1} = xF_n^{\alpha+1}$ , becomes  $F_{n-1}^{\alpha+1} = xF_n^{\alpha+1}$ . Hence, if  $F_{n+1}^{\alpha+1} = xF_n^{\alpha+1}$  then

$$D_n = \frac{(1 - x^2)x}{2\alpha + 1} (F_n^{\alpha+1})^2 > 0 \quad \text{for} \quad 0 < x < 1, \quad \alpha > -\frac{1}{2}.$$

LEMMA 2.2. If  $-\frac{1}{2} < \alpha < \beta < \gamma \leq \alpha + 1$  and  $D_n(x; \alpha, \gamma) > 0$  for  $0 < x < 1$ , then  $D_n(x; \beta, \gamma) > 0$  for  $0 < x < 1$ .

*Proof.* From (2.1) we have

$$[(1 - x^2)^{\beta-1/2} D_n(x; \beta, \gamma)]' = \frac{-2(\gamma - \beta)}{n + 2\beta} (1 - x^2)^{\beta-1/2} F_{n+1}^\beta (F_n^\gamma)'$$

We consider two cases for critical points.

Case 1. Suppose  $F_{n+1}^\beta = 0$  at the points  $x_j$ ,  $1 > x_1 > x_2 > \dots$ ; then,  $D_n(x_j; \beta, \gamma) = F_n^\beta(x_j)F_{n+1}^\gamma(x_j)$ . Since  $\text{sgn } F_n^\beta(x_j) = (-1)^{j+1}$  we have  $\text{sgn } D_n(x_j; \beta, \gamma) = (-1)^{j+1} \text{sgn } F_{n+1}^\gamma(x_j)$ . Since  $D_n(x; \alpha, \gamma) > 0$ , it follows that the roots of  $F_{n+1}^\alpha$  and  $F_{n+1}^\gamma$  interlace and hence so do the roots of  $F_{n+1}^\beta$  and  $F_{n+1}^\gamma$ . Consequently  $\text{sgn } F_{n+1}^\gamma(x_j) = (-1)^{j+1}$  and  $D_n(x_j; \beta, \gamma) > 0$ .

Case 2. Suppose  $(F_n^\gamma)' = 0$  at the points  $x_j$  with  $1 > x_1 > x_2 > \dots$ . Since

$$xF_n^\gamma - F_{n+1}^\gamma = \frac{(1-x^2)(F_n^\gamma)'}{n+2\alpha} \quad \text{and} \quad (F_n^\gamma)' = \frac{n+2\gamma}{2\gamma+1} F_{n-1}^{\gamma+1},$$

then  $F_{n-1}^{\gamma+1}(x_j) = 0$  and  $x_j F_n^\gamma(x_j) = F_{n+1}^\gamma(x_j)$ . Also,

$$D_n(x_j; \beta, \gamma) = F_n^\gamma(x_j)[x_j F_n^\beta(x_j) - F_{n+1}^\beta(x_j)] = \frac{(1-x_j^2)}{n+2\beta} F_n^\gamma(x_j)(F_n^\beta)'(x_j).$$

Now  $\text{sgn } F_n^\gamma(x_j) = (-1)^j$  and we need to show that  $\text{sgn } (F_n^\beta)'(x_j) = (-1)^j$ , or equivalently that  $\text{sgn } F_{n-1}^{\beta+1}(x_j) = (-1)^j$ . The zeros of  $F_{n-1}^{\gamma+1}$  and  $F_{n-1}^{\beta+1}$  interlace because  $0 < \gamma - \beta < 1$  and this implies  $\text{sgn } F_{n-1}^{\beta+1}(x_j) = (-1)^j$  and consequently  $D_n(x_j; \beta, \gamma) > 0$ .

By Lemma 2.1 and Lemma 2.2 we obtain the following theorem.

**THEOREM 2.1.** *If  $-\frac{1}{2} < \alpha \leq \beta \leq \alpha + 1$ , then  $D_n(x; \alpha, \beta) > 0$  for  $0 < x < 1$ .*

When  $\alpha = 1$ , Theorem 2.1 yields the trigonometric inequality

$$(2.2) \quad \sum_{k=0}^n (n-k+1) \frac{(\beta)_k (\beta)_{n-k}}{k! (n-k)!} \frac{\sin [(k+1)\theta]}{(k+1) \sin \theta} > 0, \quad 0 < \theta < \pi, \quad 1 < \beta \leq 2.$$

We will briefly outline the steps that lead from Theorem 2.1 to (2.2). First set  $x = \cos(\theta/2)$  for  $0 < \theta < \pi$ , and write

$$(n+2\beta)(n+2)P_n^\beta(1)D_n(x; 1, \beta) = 2(1-\beta) \frac{\sin [(n+2)\theta/2]}{\sin \theta/2} P_n^\beta(\cos \theta/2) + (n+2) \left[ \frac{\sin [(n+1)\theta/2]}{\sin \theta/2} P_{n+1}^\beta(\cos \theta/2) - \frac{\sin [(n+2)\theta/2]}{\sin \theta/2} P_n^\beta(\cos \theta/2) \right].$$

The second term on the right in this last expression can be written as (see [3])

$$2(\beta-1)(n+2) \cos \frac{\theta}{2} \sum_{k=0}^n \frac{(\beta)_k (\beta)_{n-k}}{k! (n-k)!} \frac{\sin [(k+1)\theta]}{(k+1) \sin \theta},$$

and the first term on the right can be rewritten as

$$\frac{2(1-\beta)}{\sin \theta/2} \sum_{k=0}^n \frac{(\beta)_k (\beta)_{n-k}}{k! (n-k)!} \sin [(k+1)\theta].$$

Combining these trigonometric sums gives (2.2).

**THEOREM 2.2.** *Write  $\Delta_n(x; \alpha, \beta) = P_n^\alpha(x)P_{n+1}^\beta(x) - P_{n+1}^\alpha(x)P_n^\beta(x)$ . If  $\alpha > 0$ , then  $(d/dx) \Delta_n(x; \alpha, \alpha + 1) > 0$  for  $0 < x < 1$ .*

*Proof.* Using the fact that

$$P_n(1) = \binom{n+2\alpha-1}{n},$$

we find that

$$(2.3) \quad A_n D_n(x; \alpha, \beta) = (n+2\alpha) \Delta_n(x; \alpha, \beta) - 2(\beta-\alpha) P_{n+1}^\alpha(x) P_n^\beta(x),$$

where  $A_n = (n+2\alpha)(n+2\beta)P_n^\alpha(1)P_n^\beta(1)/(n+1)$ . Since  $A_n > 0$  for  $\alpha > 0$  and  $\beta > 0$ , we

have from Theorem 2.1 that

$$(2.4) \quad (n + 2\alpha) \Delta_n(x; \alpha, \beta) > 2(\beta - \alpha)P_{n+1}^\alpha(x)P_n^\beta(x) \quad \text{for } 0 < x < 1, 0 < \alpha < \beta \leq \alpha + 1.$$

Next we need the identity

$$(2.5) \quad (1 - x^2) \frac{d}{dx} \Delta_n(x; \alpha, \beta) = (2\alpha - 1)x \Delta_n(x; \alpha, \beta) + 2(\beta - \alpha)P_n^\beta(x)[P_n^\alpha(x) - xP_{n+1}^\alpha(x)].$$

This identity was proved in [3]. Using (2.5) in (2.4) and combining terms we get

$$(2.6) \quad \frac{(n + 2\alpha)(1 - x^2)}{2(\beta - \alpha)} \frac{d}{dx} \Delta_n(x; \alpha, \beta) > P_n^\beta(x)[(n + 2\alpha)P_n^\alpha(x) - (n + 1)xP_{n+1}^\alpha(x)] = 2\alpha(1 - x^2)P_n^\beta(x)P_n^{\alpha+1}(x).$$

The Theorem follows upon setting  $\beta = \alpha + 1$  in (2.5). Specifically, the result is

$$(2.7) \quad \frac{d}{dx} \Delta_n(x; \alpha, \alpha + 1) > \frac{4\alpha[P_n^{\alpha+1}(x)]^2}{n + 2\alpha}, \quad \alpha > 0.$$

Integration of (2.7) yields the following positive lower bound for  $\Delta_n(x; \alpha, \alpha + 1)$ :

$$(2.8) \quad \Delta_n(x; \alpha, \alpha + 1) > \frac{4\alpha}{n + 2\alpha} \int_0^x [P_n^{\alpha+1}(t)]^2 dt, \quad 0 < x < 1, \quad \alpha > 0.$$

In view of Gaspers' inequality (1.9) and Theorem 2.2 we venture the following conjecture:

$$\frac{d}{dx} \Delta_n(x; \alpha, \beta) > 0 \quad \text{for } 0 < x < 1 \quad \text{and} \quad 0 < \alpha < \beta \leq \alpha + 1.$$

**3. Two parameter inequalities for ultraspherical polynomials.** Set  $B_n(x; \alpha, \beta) = (n + 1)F_n^\alpha F_n^\beta - nF_{n+1}^\alpha F_{n+1}^\beta$ . In this section we will prove that  $B_n(x; \alpha, \beta) > 0$  for  $-1 < x < 1$  and  $\frac{1}{2} < \alpha \leq \beta \leq \alpha + 1$ . The proof uses a differential identity and essentially the same idea used both in [3] and in the proof of Theorem 2.1 in the present paper. We will prove that a positive multiple of  $B_n(x; \alpha, \beta)$  has positive extrema in  $-1 < x < 1$ . This accounts for the presence of the factors  $n + 1$  and  $n$  in  $B_n(x; \alpha, \beta)$ . With these factors we get a tractable differential identity for  $B_n(x; \alpha, \beta)$ . From the identities

$$(1 - x^2)(F_n^\lambda)' = n(F_{n-1}^\lambda - xF_n^\lambda)$$

and

$$(n + 2\lambda)F_{n+1}^\lambda = 2(n + \lambda)x F_n^\lambda - nF_{n-1}^\lambda,$$

we can prove

$$(3.1) \quad (1 - x^2)(B_n)' = 2\alpha x B_n + 2[n(\beta - \alpha - 1) - \alpha]F_{n+1}^\alpha [F_n^\beta - \phi_n(\alpha, \beta)x F_{n-1}^\beta],$$

where  $\phi_n(\alpha, \beta) = n(\beta - \alpha - 1)[n(\beta - \alpha - 1) - \alpha]^{-1}$ .

Now, (3.1) can be rewritten as

$$(3.2) \quad [(1 - x^2)^\alpha B_n(x; \alpha, \beta)]' = 2(1 - x^2)^{\alpha-1} [n(\beta - \alpha - 1) - \alpha] F_{n+1}^\alpha [F_n^\beta - \phi_n x F_{n-1}^\beta].$$

From (3.2) it follows that  $B_n(x; \alpha, \beta)$  is positive in  $-1 < x < 1$  if  $B_n(x; \alpha, \beta)$  is positive at those points in  $-1 < x < 1$  where  $F_{n+1}^\alpha = 0$  and  $F_n^\beta = \phi_n x F_{n-1}^\beta$ . First we prove that  $B_n(x; \alpha, \beta) > 0$  when  $\beta = \alpha$  and when  $\beta = \alpha + 1$ .

LEMMA 3.1.  $B_n(x; \alpha, \alpha) > 0$  and  $B_n(x; \alpha, \alpha + 1) > 0$  for  $-1 < x < 1$  and  $\alpha > -\frac{1}{2}$ .

*Proof.*  $B_n(x; \alpha, \alpha) > 0$  in  $-1 < x < 1$  for  $\alpha > -\frac{1}{2}$  by (1.4) since  $B_n(x; \alpha, \alpha) = n[(F_n^\alpha)^2 - F_{n-1}^\alpha F_{n+1}^\alpha] + (F_n^\alpha)^2$ . When  $\beta = \alpha + 1$  we use (3.2) which reduces to

$$(3.3) \quad [(1-x^2)^\alpha B_n]^\prime = -2\alpha(1-x^2)^{\alpha-1} F_{n+1}^\alpha F_n^{\alpha+1}.$$

The right side of (3.3) vanishes when  $F_{n+1}^\alpha = 0$  and when  $F_n^{\alpha+1} = 0$ . We consider these two cases separately.

*Case 1.* If  $F_{n+1}^\alpha = 0$ , then  $B_n(x; \alpha, \alpha + 1) = (n + 1)F_n^\alpha F_n^{\alpha+1}$ . Setting  $F_{n+1}^\alpha = 0$  in the identity  $(n + 2\alpha + 1)(1 - x^2)F_n^{\alpha+1} = (2\alpha + 1)(F_n^\alpha - xF_{n+1}^\alpha)$  yields  $F_n^{\alpha+1} = (2\alpha + 1)(n + 2\alpha + 1)^{-1}(1 - x^2)^{-1}F_n^\alpha$ . Hence when  $F_{n+1}^\alpha = 0$ , then  $B_n(x; \alpha, \alpha + 1) = (2\alpha + 1)(n + 1)(n + 2\alpha + 1)^{-1}(1 - x^2)^{-1}(F_n^\alpha)^2$  and this is positive for  $-1 < x < 1, \alpha > -\frac{1}{2}$ .

*Case 2.* If  $F_n^{\alpha+1} = 0$ , then  $B_n(x; \alpha, \alpha + 1) = -nF_{n+1}^\alpha F_n^{\alpha+1}$ . From the identity  $(2\alpha + 1)F_n^\alpha = (n + 2\alpha + 1)F_n^{\alpha+1} - nxF_{n-1}^{\alpha+1}$  we get  $(2\alpha + 1)F_n^\alpha = -nxF_{n-1}^{\alpha+1}$ , and from the identity  $(1 - x^2)(n + 2\alpha + 1)F_n^{\alpha+1} = (2\alpha + 1)(F_n^\alpha - xF_{n+1}^\alpha)$  we get  $F_n^\alpha = xF_{n+1}^\alpha$ . Thus when  $F_n^{\alpha+1} = 0$ , then  $B_n(x; \alpha, \alpha + 1) = (2\alpha + 1)x^{-2}(F_n^\alpha)^2$ . This completes the proof of the lemma.

As in [3] we say that the zeros of two polynomials  $P(x)$  and  $Q(x)$  interlace if between every two consecutive zeros of  $P(x)$  there is precisely one zero of  $Q(x)$  and vice versa. The proof of the next lemma is omitted since it is almost exactly like the proof of Lemma 2.1 in [3].

LEMMA 3.2. If  $B_n(x; \alpha, \beta) \neq 0$  in  $-1 < x < 1$ , then the zeros of  $F_n^\alpha$  and  $F_{n-1}^\beta$  interlace.

THEOREM 3.1.  $B_n(x; \alpha, \beta) > 0$  for  $-1 < x < 1, \frac{1}{2} \leq \alpha \leq \beta \leq \alpha + 1$ .

*Proof.* After Lemma 3.1 we need only consider  $\frac{1}{2} < \alpha < \beta < \alpha + 1$ , and since  $B_n(-x) = B_n(x)$  and  $B_n(0) > 0$  we may assume  $0 < x < 1$ . From (3.2) it suffices to prove  $B_n > 0$  at points where  $F_{n+1}^\alpha = 0$  and at points where  $F_n^\beta = \phi_n x F_{n-1}^\beta$ .

*Case 1.* Suppose  $F_{n+1}^\alpha = 0$ . Note that the zeros of  $F_n^\alpha$  and  $F_{n+1}^\alpha$  interlace. The positive zeros of  $F_n^\beta$  are monotone decreasing functions of  $\beta$ . Since  $F_n^\beta(-x) = (-1)^n F_n^\beta(x)$ , it follows for  $\alpha < \beta < \alpha + 1$  that the zeros of  $F_n^\beta$  and  $F_{n+1}^\alpha$  interlace. Hence at a zero of  $F_{n+1}^\alpha$  the polynomials  $F_n^\beta$  and  $F_n^\alpha$  have the same sign for  $\alpha < \beta < \alpha + 1$ . We conclude that  $B_n = (n + 1)F_n^\alpha F_n^\beta > 0$  at each zero of  $F_{n+1}^\alpha$  where  $\alpha < \beta < \alpha + 1$ .

*Case 2.* In this case we have  $F_n^\beta - \phi_n x F_{n-1}^\beta = 0$ , and  $B_n$  becomes  $B_n = F_{n-1}^\beta [(n + 1)\phi_n x F_n^\alpha - nF_{n+1}^\alpha]$ . Write  $\psi_n^\beta = F_n^\beta - \phi_n x F_{n-1}^\beta$ . If  $0 < \alpha < \beta < \alpha + 1$ , then  $0 < \phi_n < 1$ . From  $0 < \phi_n < 1$  and the fact that the zeros of  $F_n^\beta, F_{n-1}^\beta$  interlace we deduce that the equation  $\psi_n^\beta = 0$  has  $n$  distinct roots in  $(-1, 1)$  and that if  $z_1 > z_2 > \dots$  are the positive roots of  $\psi_n^\beta = 0$ , then  $\text{sgn } F_{n-1}^\beta(z_j) = (-1)^{j+1}$ . Note also that the roots of  $\psi_n^\beta = 0$  are symmetric about zero. In order to prove then that  $B_n(z_j) > 0$  we need to prove that if  $\frac{1}{2} \leq \alpha < \beta < \alpha + 1$ , then

$$(3.4) \quad \text{sgn} [(n + 1)\phi_n z_j F_n^\alpha(z_j) - nF_{n+1}^\alpha(z_j)] = (-1)^{j+1}.$$

It is convenient to introduce a third parameter  $\lambda$  in addition to  $\alpha$  and  $\beta$  and define  $g_n^\lambda(x; \alpha, \beta)$  by

$$g_n^\lambda(x; \alpha, \beta) = (n + 1)\phi_n(\alpha, \beta)x F_n^\lambda(x) - nF_{n+1}^\lambda(x).$$

In this notation (3.4) becomes  $\text{sgn } g_n^\alpha(z_j) = (-1)^{j+1}$ .

First we will establish that if  $\lambda = \beta$  or  $\lambda = \beta - 1$ , then  $\text{sgn } g_n^\lambda(z_j) = (-1)^{j+1}$ . Note that  $\psi_n^\beta(z_j) = 0$  implies  $\phi_n z_j = F_n^\beta(z_j)/F_{n-1}^\beta(z_j)$ , so that

$$g_n^\lambda(z_j) = \frac{(n + 1)F_n^\beta(z_j)F_n^\lambda(z_j) - nF_{n-1}^\beta(z_j)F_{n+1}^\lambda(z_j)}{F_{n-1}^\beta(z_j)}.$$

The numerator in this last expression is positive for  $\lambda = \beta$  and  $\lambda = \beta - 1$  by Lemma 3.1. Hence if  $\lambda = \beta$  or  $\lambda = \beta - 1$ , then  $\text{sgn } g_n^\lambda(z_j) = \text{sgn } F_{n-1}^\beta(z_j) = (-1)^{j+1}$ . Next we will prove that if  $\beta - 1 < \lambda < \beta$ , then

$$(3.5) \quad \text{sgn } g_n^\lambda(z_j) = (-1)^{j+1}.$$

This will imply (3.4) by setting  $\lambda = \alpha$ . Since  $\text{sgn } g_n^\beta(z_j) = (-1)^{j+1}$  and  $\text{sgn } g_n^{\beta-1}(z_j) = (-1)^{j+1}$ , it follows that both  $g_n^{\beta-1}$  and  $g_n^\beta$  vanish at least once between consecutive zeros of  $\psi_n^\beta$ . We need to prove that  $g_n^\beta$  and  $g_n^{\beta-1}$  vanish exactly once between consecutive roots of  $\psi_n^\beta$ . Let  $y_1 > y_2 > \dots$  and  $x_1 > x_2 > \dots$  be the positive solutions of  $g_n^\beta = 0$  and  $g_n^{\beta-1} = 0$  respectively. Note that the roots of  $g_n^\lambda = 0$  are symmetric about zero, and that the degree of  $g_n^\lambda$  is  $n + 1$  while the degree of  $\psi_n^\beta$  is  $n$ . Thus, since  $g_n^\beta(z_1) > 0$  and  $g_n^{\beta-1}(z_1) > 0$ , we will have

$$(3.6) \quad y_1 > z_1 > y_2 > z_2 \dots \quad \text{and} \quad x_1 > z_1 > x_2 > z_2 \dots$$

if we can prove that  $g_n^\beta$  and  $g_n^{\beta-1}$  are negative at a point to the right of  $z_1$ . Since  $F_n^\lambda(1) = 1$ , we have  $g_n^\lambda(1) = n(\beta - 1) / [n(\beta - \alpha - 1) - \alpha]$  and hence  $g_n^\lambda(1) < 0$  for  $\beta > 1$ ,  $\frac{1}{2} \leq \alpha < \beta < \alpha + 1$ . This gives (3.6) for  $\beta > 1$ . When  $\beta = 1$  then  $g_n^\lambda(1) = 0$  and again (3.6) holds. By using the explicit representation of  $F_n^\lambda$  as a polynomial we find that

$$g_n^\beta(x) = \frac{n 2^n (\beta)_n (2\beta + n - n^2)(\beta - \alpha - 1) + 2\alpha(\beta + n)}{(2\beta)_n (2\beta + n)[n(\beta - \alpha - 1) - \alpha]} x^{n+1} + \dots,$$

$$g_n^{\beta-1}(x) = \frac{n 2^n (\beta - 1)_n (2\beta - 2 + n - n^2)(\beta - \alpha - 1) + 2\alpha(\beta - 1 + n)}{(2\beta - 2)_n (2\beta - 2 + n)[n(\beta - \alpha - 1) - \alpha]} x^{n+1} + \dots.$$

It is easy to check that the leading coefficients in these last two expressions are negative for  $n > 1$  if  $\frac{1}{2} \leq \alpha < \beta < \alpha + 1$  and  $\beta < 1$ . Thus, (3.6) holds for  $\frac{1}{2} \leq \alpha < \beta < \alpha + 1$  and  $n > 1$ . When  $n = 1$  a simple calculation shows that  $B_n > 0$ .

Next we will prove that if  $\beta - 1 < \lambda < \beta$ , then  $\text{sgn } g_n^\lambda(y_j) = (-1)^{j+1}$  and  $\text{sgn } g_n^\lambda(x_j) = (-1)^{j+2}$ . This and (3.6) will establish (3.5). Since  $g_n^\beta(y_j) = 0$  we have  $(n + 1)\phi_n y_j = nF_{n+1}^\beta(y_j) / F_n^\beta(y_j)$  and hence

$$g_n^\lambda(y_j) = n \frac{F_n^\lambda(y_j)F_{n+1}^\beta(y_j) - F_{n+1}^\lambda(y_j)F_n^\beta(y_j)}{F_n^\beta(y_j)}.$$

By Theorem 2.1 the numerator in this last expression is positive for  $\beta - 1 < \lambda < \beta$  and hence  $\text{sgn } g_n^\lambda(y_j) = \text{sgn } F_n^\beta(y_j)$ . But by writing  $\phi_n y_j = nF_{n+1}^\beta(y_j) / (n + 1)F_n^\beta(y_j)$  in  $\psi_n^\beta$  we find

$$\psi_n^\beta(y_j) = \frac{(n + 1)[F_n^\beta(y_j)]^2 - nF_{n+1}^\beta(y_j)F_{n-1}^\beta(y_j)}{(n + 1)F_n^\beta(y_j)}.$$

Since the numerator is positive,  $\text{sgn } \psi_n^\beta(y_j) = \text{sgn } F_n^\beta(y_j)$ . However  $\text{sgn } F_n^\beta(y_j) = (-1)^{j+1}$  and we have  $\text{sgn } g_n^\lambda(y_j) = (-1)^{j+1}$ . Similarly

$$g_n^\lambda(x_j) = n \frac{F_n^\lambda(x_j)F_{n+1}^{\beta-1}(x_j) - F_{n+1}^\lambda(x_j)F_n^{\beta-1}(x_j)}{F_n^{\beta-1}(x_j)}$$

and the numerator is negative by Theorem 2.1. Hence,  $\text{sgn } g_n^\lambda(x_j) = (-1)^{j+2}$ . This implies that  $g_n^\lambda$  vanishes between  $x_j$  and  $y_j$ . Since  $x_1 > z_1 > x_2 > z_2 > \dots$  and  $y_1 > z_1 > y_2 > z_2 > \dots$  it follows that  $\text{sgn } g_n^\lambda(z_j) = (-1)^{j+1}$ . This completes the proof of the Theorem.

Set  $T_n(x; \alpha, \beta) = P_n^\alpha(x)P_n^\beta(x) - P_{n+1}^\alpha(x)P_{n-1}^\beta(x)$ . We will next prove that  $T_n(x; \alpha, \beta) > 0$  when  $\beta = \alpha + 1$  and  $\beta = \alpha + 2$ .

**THEOREM 3.2.**  $T_n(x; \alpha, \alpha + 1) > 0$  and  $T_n(x; \alpha, \alpha + 2) > 0$  for  $-1 < x < 1$  and  $\alpha > \frac{1}{2}$ .

*Proof.* The Christoffel–Darboux formula [10, p. 433] gives

$$(3.7) \quad \sum_{k=0}^n (P_k^\alpha)^2 = c_n [(P_{n+1}^\alpha)' P_n^\alpha - (P_n^\alpha)' P_{n+1}^\alpha] \quad \text{where } c_n > 0.$$

Now using the identity

$$(3.8) \quad (P_n^\alpha)' = 2\lambda P_{n-1}^{\alpha+1}$$

we have from (3.7) that

$$2\alpha c_n T_n(x; \alpha, \alpha + 1) = \sum_{k=0}^n (P_k^\alpha)^2$$

and hence  $T_n(x; \alpha, \alpha + 1) > 0$  for all real  $x$  if  $\alpha > 0$ . To prove that  $T_n(x; \alpha, \alpha + 2) > 0$  we will use Theorem 2.2. From (3.8) we have

$$\begin{aligned} \frac{d}{dx} \Delta_n(x; \alpha, \beta) &= 2\beta (P_n^\alpha P_n^{\beta+1} - P_{n+1}^\alpha P_{n-1}^{\beta+1}) - 2\alpha (P_n^{\alpha+1} P_n^\beta - P_{n-1}^{\alpha+1} P_{n+1}^\beta) \\ &= 2\beta T_n(x; \alpha, \beta + 1) - 2\alpha T_n(x; \alpha + 1, \beta). \end{aligned}$$

Setting  $\beta = \alpha + 1$  above and invoking Theorem 2.2 gives

$$(\alpha + 1) T_n(x; \alpha, \alpha + 2) > \alpha T_n(x; \alpha + 1, \alpha + 1).$$

This proves the theorem since  $\alpha T_n(x; \alpha + 1, \alpha + 1) > 0$  for  $-1 < x < 1$  and  $\alpha > \frac{1}{2}$  by (1.5).

**4. Two parameter inequalities for generalized Laguerre polynomials.** In this section we prove theorems similar to those of § 3 but for generalized Laguerre polynomials.

Let  $L_n^\alpha(x)$  denote the generalized Laguerre polynomial defined by the generating identity

$$(1 - z)^{-\alpha-1} \exp\left(\frac{-xz}{1-z}\right) = \sum_{n=0}^\infty L_n^\alpha(x) z^n.$$

In [3] we proved the inequality

$$(4.1) \quad \frac{L_n^\alpha(x)L_{n+1}^\beta(x) - L_{n+1}^\alpha(x)L_n^\beta(x)}{\beta - \alpha} > 0 \quad \text{for } x > 0$$

if  $(\alpha, \beta)$  lies in either of the following regions:

- (i)  $0 < \alpha < \beta \leq \alpha + 2$
- (ii)  $0 < \beta < \alpha \leq \beta + 2$ .

Set  $l_n^\alpha(x) = L_n^\alpha(x)/L_n^\alpha(0)$ . The following theorem can be proved using arguments similar to those in [3] and consequently we omit the proof.

**THEOREM 4.1.**  $[l_n^\alpha(x)l_{n+1}^\beta(x) - l_{n+1}^\alpha(x)l_n^\beta(x)]/(\beta - \alpha) > 0$  for  $x > 0$  if  $(\alpha, \beta)$  lies in either of the regions:

- (i)  $0 < \alpha < \beta \leq \alpha + 1$
- (ii)  $0 < \beta < \alpha \leq \beta + 1$ .

Set  $\delta_n(x; \alpha, \beta) = (n + 1)l_n^\alpha l_n^\beta - n l_{n+1}^\alpha l_{n-1}^\beta$ . We will prove the analogue of Theorem 3.1.



**THEOREM 4.2.**  $\delta_n(x; \alpha, \beta) > 0$  for  $x > 0$  and  $0 < \alpha < \beta < \alpha + 1$ .

Writing  $\delta_n$  for  $\delta_n(x; \alpha, \beta)$  we can derive the differential identity

$$(4.2) \quad x(\delta_n)' = (x - \alpha - 1)\delta_n + [n(\alpha - \beta + 2) + \alpha + 1]l_{n+1}^\alpha [l_n^\beta - \phi_n(\alpha, \beta)l_{n-1}^\beta]$$

where  $\phi_n(\alpha, \beta) = n(\alpha - \beta + 2)[n(\alpha - \beta + 2) + \alpha + 1]^{-1}$ . The derivation makes use of the identities  $x(l_n^\alpha)' = n(l_n^\alpha - l_{n-1}^\alpha)$  and  $(n + \alpha + 1)l_{n+1}^\alpha = (2n + \alpha + 1 - x)l_n^\alpha - nl_{n-1}^\alpha$ .

Equation (4.2) can be rewritten as

$$(4.3) \quad [x^{\alpha+1} e^{-x} \delta_n]' = [n(\alpha - \beta + 2) + \alpha + 1] e^{-x} x^\alpha l_{n+1}^\alpha [l_n^\beta - \phi_n l_{n-1}^\beta].$$

We prove that  $x^{\alpha+1} e^{-x} \delta_n$  has positive extrema for  $x > 0$  which implies  $\delta_n$  is positive. According to (4.3) the extrema occur when  $l_{n+1}^\alpha = 0$  or  $l_n^\beta = \phi_n l_{n-1}^\beta$ . First we prove  $\delta_n(x; \alpha, \beta) > 0$  when  $\beta = \alpha$  or when  $\beta = \alpha + 1$ .

**LEMMA 4.1.**  $\delta_n(x; \alpha, \alpha) > 0$  and  $\delta_n(x; \alpha, \alpha + 1) > 0$  for  $x > 0$  and  $\alpha > -1$ .

*Proof.*  $\delta_n(x; \alpha, \alpha) = n[(l_n^\alpha)^2 - l_{n-1}^\alpha l_{n+1}^\alpha] + (l_n^\alpha)^2$ . But  $(l_n^\alpha)^2 - l_{n-1}^\alpha l_{n+1}^\alpha > 0$  for  $x > 0$  and  $\alpha > -1$  is a well known result (see [9]). Hence  $\delta_n(x; \alpha, \alpha) > 0$ .

When  $\beta = \alpha + 1$  equation (4.3) reduces to

$$(4.4) \quad [x^{\alpha+1} e^{-x} \delta_n]' = (n + \alpha + 1) e^{-x} x^\alpha l_{n+1}^\alpha [l_n^{\alpha+1} - n(n + \alpha + 1)^{-1} l_{n-1}^{\alpha+1}].$$

The right side of (4.4) vanishes when  $l_{n+1}^\alpha = 0$  or when  $l_n^{\alpha+1} = n(n + \alpha + 1)^{-1} l_{n-1}^{\alpha+1}$ .

*Case 1.* If  $l_{n+1}^\alpha = 0$ , then  $\delta_n(x; \alpha, \alpha + 1) = (n + 1)l_n^\alpha l_n^{\alpha+1}$ . Setting  $l_{n+1}^\alpha = 0$  in the identity

$$(4.5) \quad x l_n^{\alpha+1} = (\alpha + 1)(l_n^\alpha - l_{n+1}^\alpha)$$

gives  $l_n^{\alpha+1} = (\alpha + 1)x^{-1} l_n^\alpha$ . Hence, when  $l_{n+1}^\alpha = 0$  then  $\delta_n(x; \alpha, \alpha + 1) = (n + 1)(\alpha + 1)x^{-1} (l_n^\alpha)^2$  and this is positive for  $x > 0$  and  $\alpha > -1$ .

*Case 2.* If  $l_n^{\alpha+1} = n(n + \alpha + 1)^{-1} l_{n-1}^{\alpha+1}$  then from the identity

$$(4.6) \quad (\alpha + 1)l_n^\alpha = (n + \alpha + 1)l_n^{\alpha+1} - nl_{n-1}^{\alpha+1}$$

we obtain  $l_n^\alpha = 0$ ; hence,  $\delta_n(x; \alpha, \alpha + 1) = -nl_{n+1}^\alpha l_{n-1}^{\alpha+1}$ . From (4.5) we get  $l_{n+1}^\alpha = -x(\alpha + 1)^{-1} l_n^{\alpha+1}$  when  $l_n^\alpha = 0$ . From (4.6) we get  $nl_{n-1}^{\alpha+1} = (n + \alpha + 1)l_n^{\alpha+1}$  when  $l_n^\alpha = 0$ . Hence  $\delta_n(x; \alpha, \alpha + 1) = (n + \alpha + 1)(\alpha + 1)^{-1} x (l_n^{\alpha+1})^2$ . Thus when  $l_n^{\alpha+1} = n(n + \alpha + 1)^{-1} l_{n-1}^{\alpha+1}$ , then  $\delta_n(x; \alpha, \alpha + 1)$  is positive for  $x > 0$  and  $\alpha > -1$ . Lemma 3.2 holds for Laguerre polynomials so we have

**LEMMA 4.2.** If  $\delta_n(x; \alpha, \beta) \neq 0$  for  $x > 0$  then the zeros of  $l_n^\alpha$  and  $l_{n-1}^\beta$  interlace.

Returning to the proof of Theorem 4.2, we see from (4.3) that it suffices to prove  $\delta_n > 0$  at the points where  $l_{n+1}^\alpha = 0$  and at the points where  $l_n^\beta = \phi_n l_{n-1}^\beta$ . The condition  $0 < \alpha < \beta < \alpha + 1$  implies  $0 < \phi_n < n/(n + 1)$  and the argument goes through just as in the proof of Theorem 3.1. An important difference is that the zeros of  $l_n^\alpha$  are monotone increasing functions of  $\alpha$ .

**Note added in proof.** In a paper titled *Two Parameter Turan Inequalities for Ultraspherical and Laguerre Polynomials* which will appear in J. Math. Anal. Appl., J. Bustoz has proved the inequalities

$$(1) \quad (n + 1)F_n^\alpha(x)F_n^\beta(x) - nF_{n+1}^\alpha(x)F_{n-1}^\beta(x) > A_n \circ [F_n^\alpha(x)]^2,$$

$-1 < x < 1, -\frac{1}{2} < \alpha \leq \beta \leq \alpha + 1$ , where  $A_n$  is a positive constant involving  $\alpha$  and  $\beta$ ;

$$(2) \quad F_n^\alpha(x)F_n^\beta(x) - F_{n+1}^\alpha(x)F_{n-1}^\beta(x) > 0, \quad -1 < x < 1, -\frac{1}{2} < \alpha \leq \beta \leq \alpha + 1;$$

$$(3) \quad G_n^\alpha(x)G_n^\beta(x) - G_{n+1}^\alpha(x)G_{n-1}^\beta(x) > 0, \quad x > 0, \quad 0 < \alpha \leq \beta \leq \alpha + 1.$$

The proofs depend on writing the quantities in question as explicit positive sums.

## REFERENCES

- [1] R. ASKEY, *Some absolutely monotonic functions*, *Studia Sci. Math. Hungar.*, 9 (1974), pp. 51–56.
- [2] R. ASKEY AND G. GASPER, *Positive Jacobi polynomial sums II*, *Amer. J. Math.*, 98 (1976), pp. 709–738.
- [3] J. BUSTOZ AND N. SAVAGE, *Inequalities for ultraspherical and Laguerre polynomials*, *SIAM J. Math. Anal.*, 10 (1979), pp. 902–912.
- [4] A. E. DANESE, *Explicit evaluations of Turan expressions*, *Ann. Mat. Pura Appl.*, (4) 38 (1955), pp. 339–348.
- [5] G. GASPER, *On the extension of Turan's inequality to Jacobi polynomials*, *Duke Math. J.*, 38 (1971), pp. 415–428.
- [6] ———, *Positive sums of the classical orthogonal polynomials*, *SIAM J. Math. Anal.*, 8 (1977), pp. 423–447.
- [7] S. KARLIN AND G. SZEGO, *On certain determinants whose elements are orthogonal polynomials*, *J. Analyse Math.*, 8 (1961), pp. 1–157.
- [8] O. SZASZ, *Identities and inequalities concerning orthogonal polynomials and Bessel functions*, *J. Analyse Math.*, 1 (1951), pp. 116–134.
- [9] G. SZEGO, *On an inequality of P. Turan concerning Legendre polynomials*, *Bull. Amer. Math. Soc.*, 54 (1948), pp. 401–405.
- [10] ———, *Orthogonal Polynomials*, rev. ed., *Amer. Math. Soc. Colloq. Publ.*, 23, Amer. Math. Soc., Providence, RI, 1959.

## WEIGHTED $L^1$ -REMAINDER THEOREMS FOR RESOLVENTS OF VOLTERRA EQUATIONS\*

G. S. JORDAN<sup>†</sup> AND ROBERT L. WHEELER<sup>‡</sup>

**Abstract.** Conditions are found that guarantee that the resolvent of a linear Volterra integral or integro-differential equation may be written as a finite sum of products of polynomials and exponentials, plus a remainder term which belongs to a weighted  $L^1$ -space. The kernel has the form  $a(t) = c + b(t)$ ; here  $c$  is a constant and  $b(t)$  belongs to the same weighted  $L^1$ -space. It is assumed that  $b(t)$  satisfies a combination of moment and monotonicity hypotheses that is determined by the maximum of the orders of the zeros on  $\text{Re } z = 0$  of certain Laplace transform equations. The results extend to weighted  $L^1$ -spaces some recent  $L^1$ -remainder theorems due to K. B. Hannsgen (Indiana Univ. Math. J., 29 (1980), pp. 103–120). The results for resolvents are deduced from more general results for linear Volterra-Stieltjes equations. The proofs employ extensions of Banach algebra techniques used by the authors in an earlier related paper, where the hypotheses involve only moment conditions.

**1. Introduction.** We study the asymptotic structure of the integral and integro-differential resolvent kernels  $r_1$  and  $r_2$  defined by

$$(1.1) \quad r_1(t) = a(t) - \int_0^t r_1(t-s) a(s) ds, \quad t \in R^+ \equiv [0, \infty),$$

$$(1.2) \quad r_2'(t) = - \int_0^t r_2(t-s) a(s) ds, \quad r_2(0) = 1, \quad t \in R^+,$$

respectively. Here  $a$  is a complex-valued function of the form  $a(t) = c + b(t)$  where  $c$  is a constant and  $b(t)$  belongs to a weighted space  $L^1(R^+; \rho)$  (see § 2). The importance of these resolvents derives from the fact that, under mild conditions, the Volterra equations

$$(1.3) \quad x(t) = f(t) - \int_0^t x(t-s) a(s) ds, \quad t \in R^+,$$

$$(1.4) \quad x'(t) = f(t) - \int_0^t x(t-s) a(s) ds, \quad x(0) = x_0, \quad t \in R^+,$$

are solved by

$$(1.5) \quad x(t) = f(t) - \int_0^t f(t-s) r_1(s) ds, \quad t \in R^+,$$

$$(1.6) \quad x(t) = x_0 r_2(t) + \int_0^t f(t-s) r_2(s) ds, \quad t \in R^+,$$

respectively. Moreover, the resolvents  $r_k$  ( $k = 1, 2$ ) occur in variation of constants formulas associated with certain nonlinear perturbed forms of (1.3) and (1.4) (see [12, Chapt. 4] and [4]).

\* Received by the editors December 10, 1979.

<sup>†</sup> Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916. The work of this author was partially supported by the National Science Foundation under Grant MCS79-03358.

<sup>‡</sup> Department of Mathematics, University of Missouri, Columbia, Missouri 65211. The work of this author was partially supported by the National Science Foundation under Grant MCS78-01330 A01. This author is presently on leave visiting the Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

Associate with the resolvents  $r_1$  and  $r_2$  the zero sets

$$(1.7) \quad Z_1 = \{z \mid 1 + \hat{a}(z) = 0, \operatorname{Re} z \geq 0\},$$

$$(1.8) \quad Z_2 = \{z \mid z + \hat{a}(z) = 0, \operatorname{Re} z \geq 0\},$$

respectively. Here  $\hat{a}$  denotes the Laplace transform

$$\hat{a}(z) \equiv \int_0^\infty e^{-zt} a(t) dt.$$

(We have assumed that the weight  $\rho(t)$  is such that the abscissa of convergence of  $\hat{a}(z)$  is  $\operatorname{Re} z = 0$ ; this is not an essential restriction, see § 2, Remark 2.2.)

If  $a \in L^1(R^+; \rho)$  and  $Z_1 = \emptyset$ , Gelfand, Raikov and Shilov [1, p. 116] have shown that  $r_1 \in L^1(R^+; \rho)$ ; this generalizes the classical result due to Paley and Wiener [14] for the usual  $L^1$  space, that is,  $L^1(R^+) \equiv L^1(R^+; \rho)$  with  $\rho(t) \equiv 1$ . The corresponding result that  $r_2 \in L^1(R^+; \rho)$  when  $Z_2 = \emptyset$  and  $a \in L^1(R^+; \rho)$  is due to Shea and Wainger [15, Theorem 3] (and Grossman and Miller [4] in the unweighted  $L^1(R^+)$  case).

For recent results which yield that  $r_k(t)$  belongs to a weighted space  $L^1(R^+; \rho)$  when  $Z_k = \emptyset$  and  $a \notin L^1(R^+; \rho)$ ; see Gripenberg [2] and Jordan and Wheeler [11]. Also, the asymptotic rate of decay of the resolvents  $r_k(t)$  for classes of nonintegrable kernels  $a(t)$  satisfying  $Z_k = \emptyset$  is studied by Hannsgen [6], Wong and Wong [16], Gripenberg [3], and others.

We assume, for  $k = 1, 2$ , that

$$(1.9) \quad Z_k = \{z_1, \dots, z_N\} \quad \text{with } z_j \neq 0, \quad 1, \leq j \leq N < \infty,$$

and give conditions that ensure that the corresponding resolvent  $r_k$  has the form

$$(1.10) \quad r_k(t) = \sum_{j=1}^N p_j(t) e^{z_j t} + u_k(t), \quad t \in R^+,$$

where, for  $1 \leq j \leq N$ ,  $p_j(t)$  is a polynomial, and  $u_k(t) \in L^1(R^+; \rho)$ . Of particular importance are the zeros  $z_j \in Z_k$  that satisfy  $\operatorname{Re} z_j = 0$ .

Recently, the authors [10] deduced that the representation (1.10) holds with  $u_k \in L^1(R^+; \rho)$  provided that the moment hypothesis  $(1 + t)^{2M} a(t) \in L^1(R^+; \rho)$  is satisfied. Here  $M$  is the maximum of the orders of the zeros  $z_j \in Z_k$  satisfying  $\operatorname{Re} z_j = 0$ . The proof in [10] relies on Banach algebra methods, and the results for the resolvents  $r_k$  are deduced as corollaries of more general results obtained in [10] for the Volterra-Stieltjes equation

$$x(t) + \int_0^t x(t - s) dA(s) = f(t), \quad t \in R^+.$$

On the other hand, Hannsgen [8] has recently obtained a Wiener-Lévy Theorem for quotients which, when applied to the study of the resolvents  $r_k$ , also yields conditions under which (1.10) holds (see also Hannsgen [7]). The proof in [8] relies on sophisticated techniques from the theory of Fourier transforms; compare Shea and Wainger [15]. The result in [8] as applied to the study of the resolvents  $r_k(t)$  is more general than that obtained in [10], in the sense that the moment hypothesis on  $a(t)$  in [10] is generalized to a combination of moment and monotonicity assumptions; see § 2. (Also, the result in [8] may be applied to study  $r_k(t)$  for a certain class of nonintegrable kernels  $a(t)$ ; see § 2, Remark 2.1.) On the other hand, the conclusion obtained in [8] is that the remainder  $u_k(t)$  in expression (1.10) belongs to the usual Lebesgue space  $L^1(R^+)$ ; this is less sharp than the weighted remainder estimate  $u_k \in L^1(R^+; \rho)$  obtained in [10].

The purpose of this paper is to obtain the representation (1.10) with the weighted remainder estimate  $u_k \in L^1(R^+; \rho)$  when  $a(t) = c + b(t)$  with  $b \in L^1(R^+; \rho)$ , and with  $b(t)$  satisfying a combination of moment and monotonicity conditions analogous to those assumed in [8]. The proof uses an extension of the Banach algebra techniques employed in [10].

In § 2 we describe the weighted spaces and the classes of kernels that we consider. We then state and discuss our Theorems 2.1 and 2.2 for the resolvents  $r_1$  and  $r_2$ . In § 3 and § 4, results for general linear Volterra-Stieltjes equations are formulated and proved, and in § 5 we deduce Theorems 2.1 and 2.2 from the theorems of § 3.

**2. Resolvents.** By a weight  $\rho(t)$  we mean a positive continuous function on  $R^+$  such that  $\rho(0) = 1$ ,

$$(2.1) \quad \rho(t + s) \leq \rho(t) \rho(s), \quad 0 \leq t, \quad s < \infty,$$

and

$$(2.2) \quad \lim_{t \rightarrow \infty} \frac{\log \rho(t)}{-t} = \sup_{t > 0} \frac{\log \rho(t)}{-t} = 0.$$

In addition, we assume that  $\rho(t)$  satisfies the regularity condition

$$(2.3) \quad \rho(t) \text{ is nondecreasing on } R^+.$$

We remark that the first equality in (2.2) is guaranteed by (2.1) and the continuity of  $\rho(t)$ , and that

$$\rho_* \equiv \lim_{t \rightarrow \infty} \frac{\log \rho(t)}{-t}$$

satisfies  $-\infty < \rho_* \leq \infty$ ; see [1, p. 113]. Our assumption that  $\rho_* = 0$  is not an essential restriction. It is made merely to simplify the statement of our results; see Remark 2.2 at the end of this section.

Some interesting and important weights  $\rho(t)$  that satisfy (2.1)–(2.3) are

$$\begin{aligned} \rho_1(t) &= (1 + t)^\delta, \quad t \in R^+, \quad \delta \geq 0, \\ \rho_2(t) &= (1 + \log(1 + t))^\gamma \rho_1(t), \quad t \in R^+, \quad \gamma \geq 0, \\ \rho_3(t) &= \exp(t^\alpha) \rho_2(t), \quad t \in R^+, \quad 0 \leq \alpha < 1. \end{aligned}$$

The space  $L^1(R^+; \rho)$  consists of all measurable functions  $a(t)$  for which

$$\int_0^\infty \rho(t)|a(t)| dt < \infty.$$

We let  $L^1(R^+) \equiv L^1(R^+; \rho)$  when  $\rho(t) \equiv 1$ .

Let  $\rho(t)$  be a weight satisfying (2.1)–(2.3). Analogous to Hannsgen [8] we define, for  $a \in L^1(0, T)$  for each  $T > 0$ , the hypotheses  $H(M, D; \rho)$  where  $D \leq M$  are non-negative integers.

$$H(M, 0; \rho): \int_0^\infty t^M \rho(t)|a(t)| dt < \infty.$$

$$H(M, 1; \rho): a \text{ has bounded variation on } [1, \infty),$$

$$a(\infty) = 0, \quad \int_1^\infty t^M \rho(t)|da(t)| < \infty.$$

$H(M, D; \rho) (2 \leqq D)$ :  $a^{(D-2)}$  is locally absolutely continuous on  $(1, \infty)$ ,

$$a^{(\infty)} = 0, \quad \int_1^\infty t^M \rho(t) |da^{(D-1)}(t)| < \infty.$$

(Here  $a^{(D-1)}$  is normalized to be left-continuous.)

For a nonnegative integer  $M$ , we let  $S(M; \rho)$  consist of the linear combinations

$$S(M; \rho) = \left\{ a = \sum_{D=0}^M a_D \mid a_D \in H(M, D; \rho) \right\}.$$

Clearly if  $a \in S(M; \rho)$  we may, if necessary, redefine the components  $a_D(t)$  near  $t = 0$  and assume that, for  $D \geqq 1$ ,  $a_D(t) \equiv 0$  on  $0 \leqq t \leqq 1$ . Throughout this paper we will always assume that any  $a \in S(M; \rho)$  has been expressed in this manner.

The following lemma lists some elementary consequences of  $S(M; \rho)$ .

LEMMA 2.1. *Let  $M \geqq 1$  and  $a = \sum_0^M a_D \in S(M; \rho)$ . Then*

(i) *For  $0 \leqq d < D$ ,*

$$t^{M-D+1+d} \rho(t) a_D^{(d)}(t) \rightarrow 0, \quad (t \rightarrow \infty),$$

$$\int_0^\infty t^{M-D} \rho(t) |a_D(t)| dt < \infty, \quad \int_0^\infty t^{M-D+1+d} \rho(t) |da_D^{(d)}(t)| < \infty.$$

(ii) *For  $0 \leqq d < D$ ,*

$$\hat{a}_D(z) = z^{-(d+1)} \int_0^\infty e^{-zt} da_D^{(d)}(t), \quad \text{Re } z \geqq 0, \quad z \neq 0.$$

*Proof.* If  $D = 0$ , (i) is vacuous. When  $0 \leqq d \leqq D - 1 \leqq M - 1$ , note that  $a_D^{(d)}(\infty) = 0$  follows from  $a_D(\infty) = 0$ ; hence,

$$a_D^{(d)}(t) = - \int_t^\infty da_D^{(d)}(s).$$

The last equality and (2.3) yield

$$(2.4) \quad t^k \rho(t) |a_D^{(d)}(t)| \leqq \int_t^\infty s^k \rho(s) |da_D^{(d)}(s)|,$$

for  $0 \leqq d < D$  and  $0 \leqq k \leqq M - D + 1 + d$ . In particular, it follows from  $H(M, D; \rho)$  and (2.4) with  $d = D - 1$  and  $k = M$ , that  $t^M \rho(t) a_D^{(D-1)}(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Also,  $da_D^{(d-1)}(t) = a_D^{(d)}(t) dt$  for  $0 \leqq d < D$  (by convention  $da_D^{-1}(t) \equiv a_D(t) dt$ ). Thus,  $H(M, D; \rho)$  together with inequality (2.4) with  $d = D - 1$ ,  $k = M - 1$ , and an interchange of the order of integration yield the integral inequality in (i) when  $d = D - 2$ . The proof of (i) is completed by successively applying the above arguments with  $d = D - 2$  and  $d = D - 3, \dots, d = 0$  and  $d = -1$ , using at each stage the integral inequality in (i) that was obtained at the previous stage.

Part (ii) follows from the definition of  $\hat{a}_D(z)$  after  $d + 1$  successive integrations by parts. The boundary terms vanish, since  $a_D(t) = 0$  on  $0 \leqq t \leqq 1$  and  $a_D^{(k)}(t) \rightarrow 0$  as  $t \rightarrow \infty$  for  $k = 0, 1, \dots, d$ . This completes the proof of Lemma 2.1.

From Lemma 2.1 (i) it is clear that  $H(M, D; \rho) \subseteq H(M - d, D - d; \rho)$  for  $0 \leqq d \leqq D$ . In particular  $H(M + 1, M + 1; \rho) \subseteq H(M, M; \rho)$ ; hence,  $S(M + 1; \rho) \subseteq S(M; \rho)$  for each nonnegative integer  $M$ .

Also, from Lemma 2.1 (ii) with  $d = D - 1$ ,  $0 < D \leqq M$ , we see that  $\hat{b}(z)$  is  $M$  times continuously differentiable for  $\text{Re } z \geqq 0, z \neq 0$ , whenever  $b \in S(M; \rho)$ . Thus,

if  $a(t) = c + b(t)$  with  $b \in S(M; \rho)$ , we say that  $z_0 \in Z_1$  satisfying  $\operatorname{Re} z_0 = 0, z_0 \neq 0$ , is a zero of order  $m \leq M$  if the first  $m - 1$  derivatives of  $1 + \hat{a}(z)$  vanish at  $z = z_0$ , but  $\hat{a}^{(m)}(z_0) \neq 0$ . The order of a zero  $z_0 \in Z_2$  satisfying  $\operatorname{Re} z_0 = 0, z_0 \neq 0$ , is defined in the analogous manner.

**THEOREM 2.1.** *Let  $M$  be a nonnegative integer and  $a(t) = c + b(t)$  with  $b \in S(2M; \rho)$ . Suppose that  $Z_1 = \{z_1, \dots, z_N\}$  with  $z_j \neq 0$  for  $1 \leq j \leq N < \infty$ . Let  $m_j$  be the order of the zero  $z_j$ , and if  $\operatorname{Re} z_j = 0$ , assume that  $m_j \leq M$ . Then the solution  $r_1(t)$  of (1.1) has the form (1.10) where, for each  $j$ ,  $p_j(t)$  is a polynomial of degree at most  $m_j - 1$  which depends only on  $a(t)$ , and  $u_1(t) \in L^1(R^+; \rho)$ .*

For the differential resolvent  $r_2$  we have

**THEOREM 2.2.** *Let the hypotheses of Theorem 2.1 hold with the exception that the zero set  $Z_1$  in the statement of Theorem 2.1 is now replaced by the zero set  $Z_2$ . Then the solution  $r_2(t)$  of (1.2) has the form (1.10) with both  $u_1(t)$  and  $u'_1(t)$  in  $L^1(R^+; \rho)$ .*

As an application of Theorem 2.2 we state the following weighted  $L^1$ -generalization of the  $L^1(R^+)$ -remainder theorem for  $r_2(t)$  obtained by Hannsgen [7], in the case where  $a(t)$  has a special piecewise linear form.

**COROLLARY 2.1.** *Let  $\rho(t)$  be a weight satisfying (2.1)–(2.3). Let  $a(t) = c + b(t)$ , where  $c \geq 0$ , and*

$$(2.5) \quad b(t) = \sum_{n=1}^{\infty} \delta_n [1 - (nt_0)^{-1} \min \{t, nt_0\}],$$

with  $t_0 > 0$  and

$$\delta_n \geq 0, \quad 0 < b(0) = \sum_{n=1}^{\infty} \delta_n \equiv \delta < \infty.$$

Suppose that  $\omega \equiv \sqrt{\delta + c} = 2\pi k/t_0$  for some integer  $k$ , and that

$$(2.6) \quad \sum_{n=1}^{\infty} n\delta_n \rho(nt_0) < \infty.$$

Then

$$r_2(t) = 2\gamma^{-1} \cos \omega t + u_2(t),$$

where  $\gamma = (3\delta + 2c)/(\delta + c)$ , and both  $u_2$  and  $u'_2$  belong to  $L^1(R^+; \rho)$ .

Corollary 2.1 is an immediate consequence of Theorem 2.2 with  $M = 1$  since  $Z_2$  consists of a pair of simple zeros at  $\pm i\omega$  (see [7]), and, as one easily checks,

$$(2.7) \quad \int_0^{\infty} \rho(t) b(t) dt \leq 2^{-1} \int_0^{\infty} t^2 \rho(t) db'(t) \\ = t_0 2^{-1} \sum_{n=1}^{\infty} n\delta_n \rho(nt_0) < \infty,$$

so that  $b \in S(2; \rho)$ . Corollary 2.1 reduces to Hannsgen's result when  $\rho(t) \equiv 1$  in which case the first inequality in (2.7) becomes an equality.

*Remark 2.1.* Hannsgen's result in [8] also permits a certain class of nonintegrable kernels  $a(t)$ . Specifically, it follows from his result that, for  $k = 1, 2$ , (1.10) holds with  $u_k \in L^1(R^+)$  when 1) the kernel  $a(t)$  is such that  $(-1)^d a^{(d)}(t) \geq 0$  ( $0 < t < \infty, 0 \leq d \leq 2M - 1$ ),  $|a^{(2M-1)}(t)|$  is nonincreasing and convex, and  $a^{(\infty)} = 0$ ; and 2)  $Z_k = \{z_1, \dots, z_N\}$  with the zeros on  $\operatorname{Re} z = 0$  having orders  $\leq M$ . (Of course,  $z_j \neq 0, 1 \leq j \leq N$ .) The techniques of this paper do not appear to yield a weighted analogue of this result.

*Remark 2.2.* As noted above, the assumption that the weight  $\rho(t)$  is such that  $\rho_* = 0$  causes no loss of generality in Theorems 2.1 and 2.2. To state the analogous results in the case where  $\rho_* \neq 0$ , we must modify condition (2.3) to “ $\exp(\rho_* t)\rho(t)$  is nondecreasing on  $R^+$ ”. Also, the integral inequalities in the definition of  $H(M, D; \rho)$  ( $D \geq 1$ ) must be modified to

$$\int_1^\infty t^M \exp(\rho_* t)\rho(t) |d((\exp(-\rho_* t)a(t))^{(D-1)})| < \infty.$$

(Condition  $H(M, 0; \rho)$  is unchanged.)

The appropriate form of the kernel becomes  $a(t) = c \exp(\rho_* t) + b(t)$ , where  $b$  belongs to the class  $S(M; \rho)$  which now is defined using the modified  $H(M, D; \rho)$  classes. Finally,  $\operatorname{Re} z \geq \rho_*$  replaces  $\operatorname{Re} z \geq 0$  in the definitions of the zero sets  $Z_k$ . The more general result is easily obtained from the corresponding Theorem 2.1 or 2.2 by setting  $\rho_1(t) = \exp(\rho_* t)\rho(t)$ ,  $a_1(t) = \exp(-\rho_* t)a(t)$ ,  $r_{k1}(t) = \exp(-\rho_* t)r_k(t)$ , and using the operational calculus.

**3. Volterra-Stieltjes equations.** In this section we consider the scalar linear Volterra-Stieltjes equations

$$(3.1) \quad x \circledast A(t) \equiv \int_0^t x(t-s) dA(s) = f(t), \quad t \in R^+,$$

$$(3.2) \quad x'(t) + x \circledast A(t) = f(t), \quad x(0) = x_0, \quad t \in R^+.$$

For a weight  $\rho(t)$  satisfying (2.1)–(2.3), let  $V_+[\rho]$  denote the weighted space consisting of functions  $A(t)$  that are of bounded variation on  $R^+$ , normalized to be left-continuous and vanish at 0, and that satisfy

$$\|A\| \equiv \int_0^\infty \rho(t) |dA(t)| < \infty.$$

Recall (see [1, p. 166] or [9]) that  $A \in V_+[\rho]$  may be uniquely decomposed as

$$(3.3) \quad A(t) = h_A(t) + s_A(t) + g_A(t),$$

where  $h_A$  is a discrete function,  $s_A$  is a singular function, and  $g_A$  is absolutely continuous on each finite subinterval of  $R^+$ .

For  $M$  a nonnegative integer, we consider the class  $\mathcal{S}(M; \rho)$  consisting of those  $A \in V_+[\rho]$  for which

$$(3.4) \quad \int_0^\infty t^M \rho(t) |dh_A(t)| < \infty, \quad \int_0^\infty t^M \rho(t) |ds_A(t)| < \infty,$$

and

$$(3.5) \quad g_A(t) = \int_0^t a(s) ds \quad \text{with } a \in S(M; \rho).$$

Here  $S(M; \rho)$  is the class defined in § 2. Clearly  $\mathcal{S}(M + 1; \rho) \subseteq \mathcal{S}(M; \rho)$  for each nonnegative integer  $M$ .

It follows from the definition of  $\mathcal{S}(M; \rho)$  and Lemma 2.1 that the Laplace-Stieltjes transform

$$\tilde{A}(z) \equiv \int_0^\infty e^{-zt} dA(t)$$



is analytic in  $\text{Re } z \geq 0, z \neq 0$ , whenever  $A \in \mathcal{S}(M; \rho)$ . Hence, the meaning of a zero of order  $m \leq M$  of  $\tilde{A}(z)$  or  $z + \tilde{A}(z)$  in  $\text{Re } z \geq 0, z \neq 0$ , is clear.

We now state our fundamental result which concerns (3.1).

**THEOREM 3.1.** *Let  $M$  be a nonnegative integer, and assume that  $A \in \mathcal{S}(2M; \rho)$  and  $f \in S(M; \rho)$ . Suppose that the only zeros of  $\tilde{A}(z)$  in  $\text{Re } z \geq 0$  occur at  $z = z_j, 1 \leq j \leq N < \infty$ , and that no  $z_j = 0$ . Let  $m_j$  be the order of the zero  $z_j$ , and if  $\text{Re } z_j = 0$ , assume that  $m_j \leq M$ . In addition, assume that*

$$(3.6) \quad \frac{1}{\tilde{A}(z)} \text{ is bounded in } \text{Re } z \geq 0 \text{ except near the points } z_j, \quad 1 \leq j \leq N,$$

and

$$(3.7) \quad \inf_{-\infty < y < \infty} |\tilde{h}_A(iy)| > \|s_A\|.$$

If  $x$  is a Borel measurable solution of (3.1) which is integrable on each finite interval, then  $x$  has the form

$$(3.8) \quad x(t) = \sum_{j=1}^N p_j(t) e^{z_j t} + x_1(t),$$

where, for each  $j, p_j(t)$  is a polynomial of degree at most  $m_j - 1$  which depends only on  $A$  and  $f$ , and  $x_1(t) \in L^1(\mathbb{R}^+; \rho)$ .

We now turn to (3.2). By a solution of (3.2) we mean a function  $x(t)$  absolutely continuous on bounded intervals  $[0, T]$ , and such that  $x(0) = x_0$  and (3.2) holds a.e. on  $\mathbb{R}^+$ .

**THEOREM 3.2.** *Let  $M$  be a nonnegative integer, and assume that  $A \in \mathcal{S}(2M; \rho)$  and  $f \in S(M; \rho)$ . Assume that the only zeros of  $z + \tilde{A}(z)$  in  $\text{Re } z \geq 0$  occur at  $z = z_j, 1 \leq j \leq N < \infty$ , and that no  $z_j = 0$ . Let  $m_j$  be the order of the zero  $z_j$ , and if  $\text{Re } z_j = 0$ , assume that  $m_j \leq M$ .*

If  $x(t)$  is a solution of (3.2), then  $x(t)$  has the form (3.8) with both  $x_1(t)$  and  $x'_1(t)$  in  $L^1(\mathbb{R}^+; \rho)$ .

Theorem 3.2 is obtained from Theorem 3.1 by a method used in [10]. Observe that a technical hypothesis such as (3.7) is not required in Theorem 3.2.

**4. Proofs of Theorems 3.1 and 3.2.** The proof of Theorem 3.1 employs a technique similar to that used to prove the scalar case of Theorem 2.1 of [10]. The principal technical difference is treated in Lemma 4.1, which gives some consequences of  $\mathcal{S}(M; \rho)$ .

**LEMMA 4.1.** *Let  $P$  and  $Q$  be nonnegative integers such that  $Q \geq P + 1$ , and let  $z$  satisfy  $\text{Re } z = 0, z \neq 0$ . If  $A \in \mathcal{S}(Q; \rho)$ , then the function  $R(t) = R(t; A, P, z), t > 0$ , defined by*

$$(4.1) \quad R(t) = \int_0^t (t - s)^P e^{z(t-s)} dA(s) - e^{zt} \sum_{p=0}^P \binom{P}{p} t^p \tilde{A}^{(P-p)}(z),$$

satisfies  $R(t) \in S(Q - P - 1; \rho)$ .

*Proof.* For simplicity we first assume that  $P = 0$ . To establish Lemma 4.1 in this case, observe that if  $a_D \in H(Q, D; \rho), 1 \leq D \leq Q$ , then successive integrations by

parts (recall  $a_D(t) = 0$  on  $0 \leqq t \leqq 1$ ) and Lemma 2.1 (ii) yield

$$\begin{aligned}
 \int_0^t e^{z(t-s)} a_D(s) ds &= z^{-D} \int_0^t e^{z(t-s)} da_B^{(D-1)}(s) - \alpha(t) \\
 &= e^{zt} \hat{a}_D(z) - z^{-D} \int_t^\infty e^{z(t-s)} da_B^{(D-1)}(s) - \alpha(t),
 \end{aligned}
 \tag{4.2}$$

where

$$\alpha(t) \equiv \alpha(t; D, 0) \equiv \sum_{k=0}^{D-1} z^{-(k+1)} a_B^{(k)}(t).$$

Here  $a_B^{(0)} \equiv a_D$ . From Lemma 2.1 (i) we have  $a_B^{(k)} \in H(Q, D - k; \rho)$  for  $0 \leqq k \leqq D - 1$ ; hence,  $\alpha(t; D, 0) \in S(Q; \rho)$ . Moreover, since  $a_B^{(D-1)} \in H(Q, 1; \rho)$ , an interchange of the order of integration and (2.3) yields

$$\begin{aligned}
 \int_0^\infty t^{Q-1} \rho(t) |z^{-D} \int_t^\infty e^{z(t-s)} da_B^{(D-1)}(s)| dt \\
 \leqq |z|^{-D} \int_0^\infty s^Q \rho(s) |da_B^{(D-1)}(s)| < \infty.
 \end{aligned}
 \tag{4.3}$$

Combining (4.2) and (4.3) we see that

$$R(t; a_D, 0, z) = -z^{-D} \int_t^\infty e^{z(t-s)} da_B^{(D-1)}(s) - \alpha(t; D, 0)$$

belongs to  $S(Q - 1; \rho)$ .

Now, decompose  $A \in \mathcal{S}(Q; \rho)$  as in (3.3)–(3.5) with  $a = \sum_{D=0}^Q a_D$ . Since  $a_0 \in H(Q, 0; \rho)$ , the calculation in (4.3) shows that

$$\int_0^\infty t^{Q-1} \rho(t) \left| \int_t^\infty e^{z(t-s)} a_0(s) ds \right| dt < \infty;
 \tag{4.4}$$

hence

$$R(t; a_0, 0, z) = - \int_t^\infty e^{z(t-s)} a_0(s) ds \in H(Q - 1, 0; \rho).$$

In the same manner  $R(t; h_A, 0, z)$  and  $R(t; s_A, 0, z)$  both belong to  $H(Q - 1, 0; \rho)$ . Lemma 4.1 with  $P = 0$  now follows from the linearity of the Laplace-Stieltjes transform.

The proof of Lemma 4.1 for  $P > 0$  is analogous to the proof when  $P = 0$ , except for the fact that the calculations become more involved.

As in the  $P = 0$  case, we first show that  $R(t; a_D, P, z) \in S(Q - P - 1; \rho)$  when  $a_D \in H(Q, D; \rho)$ ,  $1 \leqq D \leqq Q$ . In order to prove this, we establish the following formulas ( $\text{Re } z = 0, z \neq 0$ ):

$$\begin{aligned}
 e^{zt} \sum_{p=0}^P \binom{P}{p} t^p \hat{a}_B^{(P-p)}(z) \\
 = \int_0^\infty K(t - s, z; D, P) e^{z(t-s)} da_B^{(D-1)}(s),
 \end{aligned}
 \tag{4.5}$$

and

$$(4.6) \quad \int_0^t (t-s)^P e^{z(t-s)} a_D(s) ds = \int_0^t K(t-s, z; d, P) e^{z(t-s)} da_D^{(D-1)}(s) - \alpha(t; d, P),$$

for  $1 \leq d \leq D$ , where

$$(4.7) \quad K(t, z; d, P) = \sum_{i=0}^P (-1)^i \frac{P!}{(P-i)!} \binom{d-1+i}{d-1} t^{P-i} z^{-(d+i)},$$

$$(4.8) \quad \alpha(t; d, P) = (-1)^P P! \sum_{k=0}^{d-1} \binom{P+k}{k} z^{-(P+k+1)} a_D^{(k)}(t).$$

To verify (4.5), observe that Leibniz’s rule for differentiating products may be applied to the expression for  $\hat{a}_D(z)$  with  $d = D - 1$  in Lemma 2.1 (ii), to yield

$$(4.9) \quad \hat{a}_D^{(k)}(z) = \int_0^\infty T(s, z) e^{-zs} da_D^{(D-1)}(s),$$

where  $T(s, z) = T(s, z; D, k)$ ,  $0 \leq k \leq P$ , is defined by

$$T(s, z) = \sum_{i=0}^k (-1)^i \binom{k}{i} \frac{(D-1+i)!}{(D-1)!} (-s)^{k-i} z^{-(D+i)}.$$

Then, substituting the right side of (4.9) for each  $\hat{a}_D^{(p)}(z)$  on the left side of (4.5), we obtain (after an interchange of the order of summation, some algebra, and an application of the binomial theorem to  $(t-s)^{P-i}$ ) formula (4.5) with  $K(t, z; D, P)$  defined as in (4.7).

The proof of formula (4.6) is by induction on  $d$ . To obtain (4.6) when  $d = 1$ , observe that an integration by parts using the functions  $e^{z(t-s)}$  and  $(t-s)^k a_D(s)$  yields

$$(4.10) \quad \int_0^t e^{z(t-s)} a_D(s) ds = z^{-1} \int_0^t e^{z(t-s)} da_D(s) - z^{-1} a_D(t),$$

and

$$(4.11) \quad \int_0^t (t-s)^k e^{z(t-s)} a_D(s) ds = z^{-1} \int_0^t (t-s)^k e^{z(t-s)} da_D(s) - kz^{-1} \int_0^t (t-s)^{k-1} e^{z(t-s)} a_D(s) ds,$$

for  $k \geq 1$ . Then,  $P$  successive applications of (4.11) followed by an application of (4.10), yields (4.6) when  $d = 1$  with  $K(t, z; 1, P)$  and  $\alpha(t; 1, P)$  defined in (4.7) and (4.8), respectively.

Next, assume formula (4.6) holds for some  $d$ ,  $1 \leq d < D$ , and deduce (4.6) for  $d + 1$ . To do this, recall that  $a_D^{(d)}(s) ds = da_D^{(d-1)}(s)$ , and use the method just employed to deduce (4.6) when  $d = 1$  to replace each term in the integral

$$\int_0^t K(t-s, z; d, P) e^{z(t-s)} a_D^{(d)}(s) ds$$

by the corresponding sum of integrals with measure  $da_D^{(d)}(s)$ , and with a boundary

term involving  $a_D^{(d)}(t)$ . The resulting boundary term becomes

$$-(-1)^P P! \sum_{i=0}^P z^{-(P+d+1)} a_D^{(d)}(t) \binom{d-1+i}{d-1} - \alpha(t; d, P) = -\alpha(t; d+1, P),$$

where the equality follows from the identity

$$(4.12) \quad \sum_{k=m}^n \binom{k}{m} = \binom{n+1}{m+1}.$$

Also, the double sum in the resulting integral term may be reduced to  $K(t-s, z; d+1, P)$  after some elementary algebra, a change of the variable of summation in the inner sum, an interchange of the order of summation, and use of the identity (4.12). Thus (4.6) holds for  $d+1$  with  $K(t, z; d+1, P)$  and  $\alpha(t; d+1, P)$  defined in (4.7) and (4.8), respectively, and our inductive proof of formula (4.6) is complete.

Combining (4.5), (4.6) (with  $d = D$ ), and the definition (4.1) of  $R$ , we see that when  $a_D \in H(Q, D; \rho)$ ,  $1 \leq D \leq Q$ ,

$$(4.13) \quad R(t; a_D, P, z) = - \int_t^\infty K(t-s, z; D, P) e^{z(t-s)} da_D^{(D-1)}(s) - \alpha(t; D, P).$$

Observe that  $a_D^{(D-1)} \in H(Q, 1; \rho)$ ; hence, an interchange of order of integration and (2.3) yield

$$(4.14) \quad \begin{aligned} & \int_0^\infty t^{Q-P-1} \rho(t) \left| \int_t^\infty (t-s)^k e^{z(t-s)} da_D^{(D-1)}(s) \right| dt \\ & \leq \int_0^\infty t^{Q-P-1} \rho(t) \int_t^\infty s^k |da_D^{(D-1)}(s)| dt \\ & \leq \int_0^\infty s^{k+Q-P} \rho(s) |da_D^{(D-1)}(s)| < \infty \quad (0 \leq k \leq P). \end{aligned}$$

From (4.7) and (4.14) we see that the integral term on the right side of (4.13) belongs to  $H(Q-P-1, 0; \rho)$ . Combining this with the fact that  $\alpha(t; D, P) \in S(Q; \rho)$ , we obtain  $R(t; a_D, P, z) \in S(Q-P-1; \rho)$  when  $a_D \in H(Q, D; \rho)$ ,  $1 \leq D \leq Q$ .

To complete the proof of Lemma 4.1 when  $P > 0$ , decompose  $A \in \mathcal{S}(Q; \rho)$  as in (3.3)–(3.5) with  $a = \sum_{D=0}^Q a_D$ . The formula

$$\hat{a}_0^{(k)}(z) = \int_0^\infty (-s)^k e^{-zs} a_0(s) ds, \quad \text{Re } z \geq 0, \quad 0 \leq k \leq P,$$

together with the binomial theorem, yields

$$R(t; a_0, P, z) = - \int_t^\infty (t-s)^P e^{z(t-s)} a_0(s) ds;$$

hence, the estimate in (4.14) with  $k = P$ , yields  $R(t; a_0, P, z) \in H(Q-P-1, 0; \rho)$ . Similarly  $R(t; h_A, P, z)$  and  $R(t; s_A, P, z)$  both belong to  $H(Q-P-1, 0; \rho)$ . Thus, linearity of the Laplace-Stieltjes transform yields  $R(t; A, P, z) \in S(Q-P-1; \rho)$ , and Lemma 4.1 is established.

*Proof of Theorem 3.1.* The proof is by induction on  $M$ . If  $M = 0$ , then  $A \in \mathcal{S}(0; \rho) = V_+[\rho]$ ,  $f \in S(0; \rho) = L^1(R^+; \rho)$ , and  $\hat{A}(z)$  has no zeros on  $\text{Re } z = 0$ ; hence, the result follows from the scalar case of part (i) of Theorem 2.1 of [10].

Now fix  $M \geq 0$  and assume that Theorem 3.1 holds for all nonnegative integers less than or equal to  $M$ . We show that it holds for  $M+1$ . Recall that  $S(M+1; \rho) \subseteq$

$S(M; \rho)$  and  $\mathcal{S}(2M + 2; \rho) \subseteq \mathcal{S}(2M + 1; \rho)$ ; hence, we may as well assume that  $\tilde{A}(z)$  has at least one zero of order  $M + 1$  on  $\text{Re } z = 0$ . Let  $z_1, \dots, z_L$  be all of the zeros of  $\tilde{A}(z)$  on  $\text{Re } z = 0$  having order  $M + 1$ , and define

$$(4.15) \quad S_j(t) = (z_j + 1) \int_0^t e^{z_j s} ds \quad (1 \leq j \leq L, t \in R^+).$$

Also, let  $J(t)$  be the unit step function

$$J(0) = 0, \quad J(t) = 1, \quad (t > 0),$$

and put

$$(4.16) \quad B(t) = (J + S_1) \otimes \dots \otimes (J + S_L)(t), \quad t \in R^+,$$

$$(4.17) \quad C(t) = A \otimes B(t), \quad t \in R^+.$$

An elementary Laplace transform argument [9, p. 604] shows that

$$(4.18) \quad B(t) = J(t) + \sum_{j=1}^L \alpha_j S_j(t),$$

where

$$\alpha_j = \prod_{\substack{k=1 \\ k \neq j}}^L (z_j + 1)/(z_j - z_k).$$

This fact combined with (4.17), (4.15) and  $\tilde{A}(z_j) = 0, (1 \leq j \leq L)$ , yields

$$C(t) = A(t) - \sum_{j=1}^L z_j^{-1} \alpha_j (z_j + 1) (e^{z_j t} \int_t^\infty e^{-z_j s} dA(s) + A(t)),$$

and, consequently,

$$(4.19) \quad dC(t) = dA(t) - \sum_{j=1}^L \alpha_j (z_j + 1) \int_t^\infty e^{-z_j s} dA(s) e^{z_j t} dt.$$

Since  $A(t) \in \mathcal{S}(2M + 2; \rho)$  and  $\tilde{A}(z_j) = 0, (1 \leq j \leq L)$ ,  $R(t; A, 0, z_j)$  defined in (4.1) belongs to  $S(2M + 1; \rho)$  and satisfies

$$R(t; A, 0, z_j) = - \int_t^\infty e^{z_j(t-s)} dA(s), \quad (1 \leq j \leq L);$$

hence, by (4.19),  $C(t) \in \mathcal{S}(2M + 1; \rho)$ .

From (4.16) and (4.17) we have

$$(4.20) \quad \tilde{C}(z) = \tilde{A}(z)\tilde{B}(z) = \tilde{A}(z) \prod_{j=1}^L \frac{z + 1}{z - z_j}$$

for  $\text{Re } z \geq 0, z \neq z_j (1 \leq j \leq L)$ . Thus, in  $\text{Re } z > 0$   $\tilde{C}(z)$  has the same zeros (including order) as  $\tilde{A}(z)$ ; on  $\text{Re } z = 0$  the zeros of  $\tilde{A}(z)$  of order  $M + 1$  are zeros of  $\tilde{C}(z)$  of order  $M$ , and the remaining zeros of  $\tilde{A}(z)$  are zeros of  $\tilde{C}(z)$  of the same order. Moreover, it follows from (4.17), the expression for  $\tilde{S}_j(z)$ , (4.18), and Taylor's formula with remainder, that

$$(4.21) \quad \tilde{C}^{(M)}(z_j) = \frac{\alpha_j (z_j + 1) \tilde{A}^{(M+1)}(z_j)}{M + 1}, \quad 1 \leq j \leq L.$$

From (4.20), we see that (3.6) holds with  $\tilde{A}$  replaced by  $\tilde{C}$ . Moreover, if we

express  $C = h_C + s_C + g_C$  as in (3.3), it follows from (4.19) that  $h_C = h_A$  and  $s_C = s_A$ ; hence, (3.7) holds with  $A$  replaced by  $C$ .

Now define

$$x^1(t) = x(t) - g(t),$$

where

$$g(t) = \sum_{j=1}^L \beta_j \hat{f}(z_j) t^M e^{z_j t},$$

with

$$(4.22) \quad \beta_j = (M + 1) / \tilde{A}^{(M+1)}(z_j).$$

From (4.17) and (3.1) we have

$$(4.23) \quad x^1 \otimes C(t) = f_1(t), \quad t \in \mathbb{R}^+,$$

with

$$f_1(t) \equiv f \otimes B(t) - g \otimes C(t), \quad t \in \mathbb{R}^+.$$

In order to apply the inductive hypothesis to (4.23), we must show that  $f_1 \in S(M; \rho)$ . Combining (4.18) with the definition of  $f_1$  yields

$$(4.24) \quad f_1(t) = f(t) + \sum_{j=1}^L \alpha_j (z_j + 1) \int_0^t e^{z_j(t-s)} f(s) ds - g \otimes C(t),$$

where

$$(4.25) \quad g \otimes C(t) = \sum_{j=1}^L \beta_j \hat{f}(z_j) \int_0^t (t - s)^M e^{z_j(t-s)} dC(s).$$

Since  $\tilde{C}^{(k)}(z_j) = 0, 0 \leq k \leq M - 1, 1 \leq j \leq L$ , and  $C \in \mathcal{S}(2M + 1; \rho)$ , Lemma 4.1 yields that

$$\int_0^t (t - s)^M e^{z_j(t-s)} dC(s) = R(t; C, M, z_j) + e^{z_j t} \tilde{C}^{(M)}(z_j),$$

with  $R(t; C, M, z_j) \in S(M; \rho), 1 \leq j \leq L$ . Combining this with (4.21), (4.22) and (4.25), we obtain

$$(4.26) \quad g \otimes C(t) - \sum_{j=1}^L \alpha_j (z_j + 1) \hat{f}(z_j) e^{z_j t} \in S(M; \rho).$$

Also, since  $f \in S(M + 1; \rho)$ , we may set  $F(t) = \int_0^t f(s) ds$  and apply Lemma 4.1 to get that

$$(4.27) \quad R(t; F, 0, z_j) = \int_0^t e^{z_j(t-s)} f(s) ds - e^{z_j t} \hat{f}(z_j) \in S(M; \rho)$$

for  $1 \leq j \leq L$ . Combining (4.24), (4.26), (4.27), and  $f \in S(M + 1; \rho)$ , we see that  $f_1 \in S(M; \rho)$ .

Our inductive hypothesis applied to (4.23) yields

$$(4.28) \quad x^1(t) = \sum_{j=1}^N p_j^1(t) e^{z_j t} + x_1(t),$$

where, for each  $j, p_j^1(t)$  is a polynomial and  $x_1(t) \in L^1(\mathbb{R}^+; \rho)$ . Moreover, the degree

of  $p_j^1(t)$  is at most  $M - 1$  for  $1 \leq j \leq L$ , and at most  $m_j - 1$  for the remaining zeros  $z_j$  of  $\tilde{A}(z)$  in  $\text{Re } z \geq 0$ . (Of course, there is no sum in (4.28) when  $L = N$  and  $M = 1$ .) It follows from our definitions of  $x^1$  and  $g$  that  $x$  has the form (3.8), with  $p_j(t) = p_j^1(t) + \beta_j \hat{f}(z_j)t^M$  for  $1 \leq j \leq L$ , and  $p_j(t) = p_j^1(t)$  for the remaining zeros  $z_j$  of  $\tilde{A}(z)$ . This completes the proof of Theorem 3.1.

It would be of interest to generalize Theorem 3.1 to a result that includes linear Volterra-Stieltjes systems. We remark that the technique employed in [10] to deduce the systems case of [10, Thm. 2.1] from the corresponding scalar case of the same theorem is not directly applicable in the setting of the present paper. The difficulty is caused by the fact that the classes  $\mathcal{S}(M; \rho)$  are not closed under convolution multiplication.

The following elementary lemma is needed for our proof of Theorem 3.2.

LEMMA 4.2. *Let  $Q$  be a nonnegative integer and set  $G(t) = e^{-t}$ ,  $t \in R^+$ . If  $A \in \mathcal{S}(Q; \rho)$ , then  $G \otimes A(t) \in \mathcal{S}(Q; \rho)$ .*

*Proof.* We begin by showing that whenever  $a_D \in H(Q, D; \rho)$ ,  $1 \leq D \leq Q$ , then

$$(4.29) \quad b_D(t) \equiv G * a_D(t) \in H(Q, D; \rho).$$

Here  $G * a_D(t) \equiv \int_0^t G(t-s) a_D(s) ds$ . To prove (4.29), differentiate  $b_D$  ( $D - 1$ )-times, and obtain, using  $a_D(t) \equiv 0$  on  $0 \leq t \leq 1$ ,  $b_D^{(D-1)}(t) = G * a_D^{(D-1)}(t)$ . Integrating the last expression by parts yields

$$b_D^{(D-1)}(t) = a_D^{(D-1)}(t) - \int_0^t G(t-s) da_D^{(D-1)}(s);$$

hence,

$$db_D^{(D-1)}(t) = \int_0^t G(t-s) da_D^{(D-1)}(s) dt.$$

Since  $a_D(t) = 0$  on  $0 \leq t \leq 1$ , we get, after an interchange of the order of integration and a change of variables, that

$$(4.30) \quad \int_0^\infty t^Q \rho(t) |db^{(D-1)}(t)| \leq \int_0^\infty \int_0^\infty (t+s)^Q \rho(t+s) G(t) dt |da_D^{(D-1)}(s)|.$$

Using  $A_D \in H(Q, D; \rho)$ , (2.1), and (2.2), we see that the expression on the right side of (4.30) is finite, and (4.29) is established.

Now decompose  $A \in \mathcal{S}(Q; \rho)$  as in (3.3)–(3.5). It follows from the calculation in (4.30) and (2.1), (2.2), (3.4) and  $a_0 \in H(Q, 0; \rho)$ , that

$$b_0(t) \equiv G \otimes [h_A + s_A](t) + G * a_0(t)$$

belongs to  $H(Q, 0; \rho)$ . Combining this with (4.29) completes the proof of Lemma 4.2.

*Proof of Theorem 3.2.* The proof uses a scalar version of the technique employed to deduce Theorem 2.2 of [10], together with Lemmas 4.1 and 4.2. Namely, let  $G(t) = e^{-t}$  and convolve both sides of (3.2) with  $G(t)$  to obtain

$$(4.31) \quad x' * G(t) + x * G \otimes A(t) = G * f(t), \quad x(0) = x_0, \quad t \in R^+.$$

Integrate the first term on the left side of (4.31) by parts, and rearrange the resulting equation to rewrite (4.31) as

$$(4.32) \quad x \otimes B(t) = k(t),$$

where

$$k(t) \equiv G * f(t) + x(0) G(t),$$

and

$$B(t) = J(t) + \int_0^t b(s) ds,$$

with  $J$  the unit step function and

$$b(t) = -G(t) + G \otimes A(t).$$

Since  $G \in H(Q, 0; \rho)$  for any  $Q$ , Lemma 4.2 with  $Q = 2M$  yields  $b \in S(2M; \rho)$ ; hence  $B \in \mathcal{S}(2M; \rho)$ . Similarly, we may set  $F(t) = \int_0^t f(s) ds$ , and apply Lemma 4.2 with  $Q = M$  to  $F$  to get  $G * f \in S(M; \rho)$ ; hence  $k \in S(M; \rho)$ .

From the definition of  $B$ , we see that  $h_B(t) = J(t)$  and  $s_B(t) = 0$  in the decomposition (3.3) of  $B$ . Also, the definitions of  $G, B$  and  $b$  show that  $\tilde{B}(z) = (1 + z)^{-1}(z + \tilde{A}(z))$  for  $\text{Re } z \geq 0$ ; hence,  $\tilde{B}(z)$  and  $z + \tilde{A}(z)$  have the same zeros (including order) in  $\text{Re } z \geq 0$ . Thus, Theorem 3.1 may be applied to equation (4.32) to yield that  $x(t)$  has the form (3.8) with  $x_1(t) \in L^1(R^+; \rho)$ .

It remains to show that  $x_1'(t) \in L^1(R^+; \rho)$ . Let  $w(t)$  denote the sum on the right side of (3.8), and use (3.8) and (3.2) to write

$$x_1'(t) = f(t) - x_1 \otimes A(t) - w(t) - w \otimes A(t).$$

Since  $f$  and  $x_1 \otimes A$  both belong to  $L^1(R^+; \rho)$ ,  $x_1'(t) \in L^1(R^+; \rho)$  follows from

$$(4.33) \quad w'(t) + w \otimes A(t) \in L^1(R^+; \rho).$$

To verify (4.33), first let  $z_j$  be a zero of  $z + \tilde{A}(z)$  of order  $m_j$  satisfying  $\text{Re } z_j = 0$ . By Lemma 4.1,

$$R(t; A, k, z_j) = \int_0^t (t - s)^k e^{z_j(t-s)} dA(s) - e^{z_j t} \sum_{p=0}^k \binom{k}{p} t^p \tilde{A}^{(k-p)}(z_j)$$

belongs to  $S(2M - k - 1; \rho) \subseteq L^1(R^+; \rho)$  for  $0 \leq k \leq m_j - 1$ . Combining this with the fact that  $z_j$  is a zero of  $z + \tilde{A}(z)$  of order  $m_j$ , we deduce that

$$(4.34) \quad (t^k e^{z_j t})' + \int_0^t (t - s)^k e^{z_j(t-s)} dA(s)$$

belongs to  $L^1(R^+; \rho)$  for  $0 \leq k \leq m_j - 1$ ; hence the terms in  $w' + w \otimes A$  contributed by zeros of  $z + \tilde{A}(z)$  on  $\text{Re } z = 0$  belong to  $L^1(R^+; \rho)$ .

The proof that the terms in  $w' + w \otimes A$  contributed by zeros of  $z + \tilde{A}(z)$  in  $\text{Re } z > 0$  also belong to  $L^1(R^+; \rho)$  is even easier. For if  $z_j, \text{Re } z_j > 0$ , is a zero of  $z + \tilde{A}(z)$  having order  $m_j$ , then the expression (4.34) may be written as

$$(4.35) \quad - \int_t^\infty (t - s)^k e^{z_j(t-s)} dA(s), \quad 0 \leq k \leq m_j - 1.$$

Here we have used the binomial theorem, the formula for  $\tilde{A}^{(i)}(z)$  in  $\text{Re } z > 0$ , the definition of a zero of  $z + \tilde{A}(z)$  of order  $m_j$ , and the fact that the integral (4.35) converges since  $\text{Re } z_j > 0$ . Using (2.2) and  $\text{Re } z_j > 0$ , one easily checks that the integral



(4.35) belongs to  $L^1(R^+; \rho)$ . This completes the proof of (4.33), and Theorem 3.2 is established.

**5. Proofs of Theorems 2.1 and 2.2.** In this section we deduce the theorems in Section 2 from those in § 3.

*Proof of Theorem 2.1.* When  $c = 0$ , Theorem 2.1 follows immediately from Theorem 3.1 by setting  $A(t) = J(t) + \int_0^t a(s) ds$ , where  $J$  is the unit step function, and rewriting (1.1) as

$$(5.1) \quad r_1 \circledast A(t) = a(t), \quad t \in R^+.$$

To establish Theorem 2.1 when  $c \neq 0$ , convolve (1.1) with  $G(t) = e^{-t}$ , subtract the resulting equation from (1.1), and use  $a(t) = c + b(t)$  to obtain

$$(5.2) \quad r_1(t) + r_1 * a_1(t) = f(t), \quad t \in R^+,$$

where

$$\begin{aligned} a_1(t) &= b(t) - b * G(t) + (c - 1)G(t), \\ f(t) &= a_1(t) + G(t). \end{aligned}$$

Lemma 4.2 with  $Q = 2M$  applied to  $B(t) \equiv \int_0^t b(s) ds$  yields  $b * G \in S(2M; \rho)$ ; hence  $a_1$  and  $f$  both belong to  $S(2M; \rho)$ . Define  $A_1(t) \equiv J(t) + \int_0^t a_1(s) ds$ , and rewrite (5.2) as

$$(5.3) \quad r_1 \circledast A_1(t) = f(t), \quad t \in R^+.$$

From the definitions of  $A_1$ ,  $a_1$  and  $G$ , we see that

$$\begin{aligned} \tilde{A}_1(z) &= (1 - \hat{G}(z))(1 + \hat{b}(z)) + c\hat{G}(z) \\ &= z(1 + z)^{-1} (1 + \hat{a}(z)), \quad \text{Re } z \geq 0. \end{aligned}$$

In particular,  $\tilde{A}_1(0) = c \neq 0$ ; hence,

$$\{z | \tilde{A}_1(z) = 0, \text{Re } z \geq 0\} = Z_1,$$

including order of the zeros in either set. Thus, Theorem 3.1 applied to (5.3) yields that  $r_1$  has the form (1.10) with  $u_1(t) \in L^1(R^+; \rho)$ , and Theorem 2.1 is established when  $c \neq 0$ .

*Proof of Theorem 2.2.* When  $c = 0$ , set  $A(t) = \int_0^t a(s) ds$ , rewrite (1.2) as

$$(5.4) \quad r_2'(t) + r_2 \circledast A(t) = 0, \quad r_2(0) = 1, \quad t \in R^+,$$

and apply Theorem 3.2 to (5.4) to see that  $r_2$  satisfies the conclusion of Theorem 2.2.

If  $c \neq 0$ , convolve (1.2) with  $G(t) = e^{-t}$ , and integrate the term  $r_2' * G(t)$  by parts to obtain

$$(5.5) \quad r_2(t) - r_2 * G(t) + r_2 * a * G(t) = G(t), \quad t \in R^+.$$

Next, subtract (5.5) from (1.2) and use  $a(t) = c + b(t)$ , to get

$$(5.6) \quad r_2'(t) - r_2(t) + r_2 * a_2(t) = -G(t), \quad r_2(0) = 1, \quad t \in R^+,$$

where

$$a_2(t) = b(t) - b * G(t) + (1 + c) G(t).$$

As in the proof of Theorem 2.1,  $a_2 \in S(2M; \rho)$ ; hence,  $A_2(t) \equiv -J(t) + \int_0^t a_2(s) ds$

belongs to  $\mathcal{S}(2M; \rho)$ , and (5.6) may be rewritten as

$$(5.7) \quad r_2'(t) + r_2 \circledast A_2(t) = -G(t), \quad r_2(0) = 1, \quad t \in R^+.$$

One easily verifies that  $\tilde{A}_2(z) = z(1+z)^{-1}(z + \hat{a}(z))$ ; hence  $\tilde{A}_2(0) = c \neq 0$ , and

$$\{z \mid z + \tilde{A}_2(z) = 0, \operatorname{Re} z \geq 0\} = Z_2$$

(including order). Now apply Theorem 3.2 to equation (5.7) to obtain the desired result.

**Acknowledgment.** The authors wish to thank Professor K. B. Hannsgen of Virginia Polytechnic Institute and State University for helpful discussions concerning the subject of this paper.

#### REFERENCES

- [1] I. M. GELFAND, D. A. RAIKOV, AND G. E. SHILOV, *Commutative Normed Rings*, Chelsea, New York, 1964.
- [2] G. GRIPENBERG, *On rapidly decaying resolvents of Volterra equations*, J. Integral Equations 1 (1979), pp. 241–247.
- [3] ———, *On the asymptotic behavior of resolvents of Volterra equations*, this Journal, 11 (1980), pp. 654–662.
- [4] S. I. GROSSMAN AND R. K. MILLER, *Perturbation theory for Volterra integrodifferential systems*, J. Differential Equations 8 (1970), pp. 457–474.
- [5] ———, *Nonlinear Volterra integrodifferential systems with  $L^1$ -kernels*, J. Differential Equations 13 (1973), pp. 551–566.
- [6] K. B. HANNSGEN, *A Volterra equation with completely monotonic convolution kernel*, J. Math. Anal. Appl. 31 (1970), pp. 459–471.
- [7] ———, *An  $L^1$  remainder theorem for an integrodifferential equation with asymptotically periodic solution*, Proc. Amer. Math. Soc. 73 (1979), pp. 331–337.
- [8] ———, *A Wiener-Lévy Theorem for quotients, with applications to Volterra equations*, Indiana Univ. Math. J. 29(1980), pp. 103–120.
- [9] G. S. JORDAN AND R. L. WHEELER, *Asymptotic behavior of unbounded solutions of linear Volterra integral equations*, J. Math. Anal. Appl. 55 (1976), pp. 596–615.
- [10] ———, *Structure of resolvents of Volterra integral and integrodifferential systems*, this Journal, 11 (1980), pp. 119–132.
- [11] ———, *Rates of decay of resolvents of Volterra equations with certain nonintegrable kernels*, J. Integral Equations, to appear.
- [12] R. K. MILLER, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
- [13] ———, *Structure of solutions of unstable linear Volterra integrodifferential equations*, J. Differential Equations, 15 (1974), pp. 129–157.
- [14] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 19, American Mathematical Society Providence, RI, 1934.
- [15] D. F. SHEA AND S. WAINGER, *Variants of the Wiener-Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.
- [16] G. S. W. WONG AND R. WONG, *Asymptotic solutions of linear Volterra integral equations with singular kernels*, Trans. Amer. Math. Soc., 189 (1974), pp. 185–200.

## PERIODIC SOLUTIONS OF A NONLINEAR AGE-DEPENDENT MODEL OF SINGLE SPECIES POPULATION DYNAMICS\*

K. E. SWICK†

**Abstract.** A system of nonlinear differential and integral equations, derived in an earlier paper as a model of single species population dynamics, is studied under the general assumption that the birth rate decreases or the death rate increases as the total population increases. Lower and upper uniform asymptotic bounds are determined for solutions of the model, and sufficient conditions are determined for the existence of periodic solutions.

**1. Introduction.** The renewal equation, as developed by Lotka [8], has been used extensively by demographers as a mathematical framework for determining the future female birth trajectory in a closed population when the initial population distribution and the birth and death moduli are known. In any particular application, the initial population distribution can easily be determined from census data. Present and past values for the birth and death moduli can also easily be determined from life tables, which are derived from existing census data, see Keyfitz [7, Chapt. 2]. To accurately determine the future birth trajectory, it is required that one know the birth and death moduli for future time; information which is, of course, not available. To obtain future birth trajectories, it is normally assumed that these moduli remain unchanged in the future, but since these rates can often change significantly, an accurate prediction of the future birth trajectory in these cases must accommodate these fluctuations.

Since it is impossible to know precisely the future mortality and maternity rates for any population, one must develop birth and death moduli which determine future rates in response to given exterior conditions; that is, the birth and death moduli are nonlinear functions of these variables.

Frauenthal [2] and Rorres [10] have looked at nonlinear generalizations of Lotka's model in which the birth moduli react to the cohort size. These models reflect an observation of Easterlin [1, Chapt. 5] that women born in large cohorts tend to produce fewer children than women born in small cohorts.

A population subject to limited resources or space often exhibits a decline in birth rate or increased mortality as a result of an increase in total population. In [3] Griffel derives an age-dependent model in which the mortality function exhibits this property, and determines certain asymptotic properties of solutions of this model.

Gurtin and MacCamy [4] derive a model, based on a partial differential equation proposed by Von Foerster [14], in which both birth and death moduli are functions of age as well as total population. In [13] this author derived a model, based on the Gurtin and MacCamy model, but satisfying the additional conditions: (i) there is a maximum life span  $L$  in the population, (ii) there is a time lag  $\tau \geq 0$  between conception and birth, and (iii) the birth and death moduli are also time-dependent. Properties of solutions were determined under the basic assumption that the birth rate increases with increasing population, a situation which occurs, for example, in the spread of certain contagious diseases. Threshold levels were established determining

---

\* Received by the editors November 9, 1979, and in final revised form February 25, 1980. This work was supported in part by the National Institutes of Health under Grant GM-24326.

† Department of Mathematics, Queens College, City University of New York, Flushing, New York 11367.

situations under which the population disappears and under which it increases without bound as  $t \rightarrow \infty$ .

We continue here the study of the model derived in [13], with the exception that it will be assumed here that  $L = +\infty$ .

In § 2 we sketch the derivation of the model to be studied. In § 3, upper and lower asymptotic bounds are determined for solutions of the system studied here under the general assumption that the birth rate decreases and/or the mortality rate increases when the total population increases. These results enable the use of an asymptotic fixed point theorem of Horn to establish the existence of periodic solutions of the system.

The existence of periodic solutions in non-age-dependent population dynamics is well known, and has been studied extensively. For models in which the mortality is age-specific and the maternity dependent on the size of the population, the existence of periodic solutions was established in [11] and [12]. Little is known, however, about the existence of periodic solutions in age-dependent dynamics, although such information is quite useful for a full understanding of the nature of the dynamics of a population.

In [5] Gurtin and MacCamy examined the autonomous system derived in [4] under the assumption that mortality is only dependent on the population size. They found there for two choices of the maternity, which were functions of age only, that the system had no nontrivial periodic solutions. They conjecture that the system has no closed orbits, under reasonable assumptions, even when maternity is population-dependent. In § 5 we give sufficient conditions, via a Hopf bifurcation, for the existence of periodic solutions for one of the cases studied in [5]. This condition requires that maternity be also population-dependent.

**2. The model.** If we assume that  $\rho(a, t)$  is the population at time  $t$  in the age interval  $(a, a + da)$ , and  $P(t)$  is the total population, then

$$(2.1) \quad P(t) = \int_0^\infty \rho(a, t) da.$$

If we set  $D\rho(a, t) = \lim (1/h)(\rho(a + h, t + h) - \rho(a, t)), h \rightarrow 0$ , then  $D\rho$  is the rate at which the population of age  $a$  at time  $t$  is changing with respect to time. If  $d(a, t)$  is the number of individuals of age  $a$  who die at time  $t$ , per unit age and time, and if we assume that  $d(a, t) = \lambda(a, P(t), t)\rho(a, t)$ , then we have Von Foerster's equation,

$$(2.2) \quad D\rho(a, t) + \lambda(a, P(t), t)\rho(a, t) = 0.$$

As a result of (ii) and (iii), the birth modulus  $\beta$  depends on the size of the population at time  $t - \tau$ , and assuming that the birth process is described by the "renewal equation," we have

$$(2.3) \quad B(t) = \rho(0, t) = \int_0^\infty \beta(a, P(t - \tau), t - \tau)\rho(a, t - \tau) da.$$

Here  $\beta(a, P, t)$ , the birth modulus, is the average number of offspring per unit population produced, at time  $t$ , by an individual of age  $a$ ,  $\rho(0, t) = B(t)$  is the birth rate, and  $\lambda$ , the death modulus, is the death rate at time  $t$ , per unit population of age  $a$ .

The model studied here is the system of equations (2.1)–(2.3) with the initial condition:

$$(2.4) \quad \rho(a, t) = \varphi(a, t) \quad a \geq 0, \quad -\tau \leq t \leq 0.$$

It was shown in [13] that if we set  $P(t) = \int_0^\infty \varphi(a, t) da$  and  $B(t) = \varphi(0, t) - \tau \cong t \cong 0$ , then

$$(2.5) \quad \rho(a, t - x) = \begin{cases} \varphi(a - t, -x) \exp \left[ - \int_0^t \lambda(a - t + s, P(s - x), s - x) ds \right] & \text{for } a \geq t, \\ B(t - x - a) \exp \left[ - \int_0^a \lambda(s, P(t - x - a + s), t - x - a + s) ds \right] & \text{for } a < t, \end{cases}$$

where  $x = 0$  or  $\tau$ . Substituting (2.5) into (2.1)–(2.3), we find that (2.1)–(2.3) is equivalent to

$$(2.6a) \quad P(t) = \int_0^t B(a) \exp \left[ - \int_a^t \lambda(s - a, P(s), s) ds \right] da + \int_0^\infty \varphi(a, 0) \exp \left[ - \int_0^t \lambda(s + a, P(s), s) ds \right] da$$

$$(2.6b) \quad B(t) = \int_0^t \beta(t - a, P(t - \tau), t - \tau) B(a - \tau) \cdot \exp \left[ - \int_a^t \lambda(s - a, P(s - \tau), s - \tau) ds \right] da + \int_0^\infty \beta(t + a, P(t - \tau), t - \tau) \varphi(a, -\tau) \cdot \exp \left[ - \int_0^t \lambda(s + a, P(s - \tau), s - \tau) ds \right] da.$$

It can be shown that (2.1)–(2.4) has a unique solution, existing for  $t \geq 0$ , if the following assumption is satisfied. (See [4] for the case  $\tau = 0$  and [13] for  $\tau > 0$ .) Let  $I_0 = [-\tau, 0]$  and  $I_1 = [-\tau, \infty)$ .

$H_0$ :  $\varphi \in L_1(R^+ \times I_0)$  is piecewise continuous;  $\lambda, \beta \in C(R^+ \times R^+ \times I_1)$ ,  $\lambda_P(a, P, t)$  and  $\beta_P(a, P, t)$  exist for  $a \geq 0, P \geq 0, t \geq -\tau$  and  $\lambda(\cdot, P, t), \beta(\cdot, P, t), \lambda_P(\cdot, P, t), \beta_P(\cdot, P, t)$ , as functions of  $P$  and  $t$ , belong to  $C(R^+ \times I_1 : L_\infty(R^+))$  for  $\varphi \geq 0, \lambda \geq 0$ , and  $\beta \geq 0$ ; and  $\sup \beta < \infty$ .

We will assume, henceforth, that  $H_0$  is satisfied.

By  $\rho(a, t; \varphi)$  we will mean the solution of (2.1)–(2.4), and by  $(P(t; \varphi), B(t; \varphi))$  the associated solution of (2.6), generated by the initial function  $\varphi$ .

**3. Contraction of solutions.** We consider first the case  $\tau = 0$ . Let  $(P, B)$  be a pair of functions satisfying (2.6), with  $\tau = 0$ , for  $t \geq 0$ . Differentiating (2.6a) and substituting the right hand side of (2.6b) for  $B(t)$  in the result, we get, noting that if  $\tau = 0$  then  $\varphi(a, t) = \varphi(a)$ ,

$$(3.1) \quad P'(t) = \int_0^t [\beta(t - a, P(t), t) - \lambda(t - a, P(t), t)] B(a) \cdot \exp \left[ - \int_a^t \lambda(s - a, P(s), s) ds \right] da + \int_0^\infty [\beta(t + a, P(t), t) - \lambda(t + a, P(t), t)] \varphi(a) \cdot \exp \left[ - \int_0^t \lambda(s + a, P(s), s) ds \right] da, \quad \text{for } t > 0.$$

Equation (3.1) provides a direct means of calculating bounds for solutions of (2.6). In particular, if  $\beta(a, P, t) - \lambda(a, P, t) \leq -\varepsilon < 0$  for  $a, t \geq 0$  when  $P \geq P_2 \geq 0$ , then it follows from (3.1) and (2.6a) that  $P'(t) \leq -\varepsilon P(t)$  when  $P(t) \geq P_2$ . Since when  $\tau = 0, \|\varphi\| = \int_0^\infty \varphi(a) da = P(0; \varphi)$ , it follows that for any  $M > 0$  there is  $T_0 \geq 0$  such that  $P(t; \varphi) \leq P_2$  for  $t \geq T_0$  if  $\|\varphi\| \leq M$ . A similar situation holds for the lower bound.

The following theorem gives some rather general conditions under which the solutions of (2.6) contract uniformly to a given set when  $\tau = 0$ .

**THEOREM 1.** *Suppose  $\tau = 0, H_0$  is satisfied and there are positive constants  $\varepsilon$  and  $P_1 < P_2$  such that  $\beta(a, P, t) - \lambda(a, P, t) \geq \varepsilon$  for  $a, t \geq 0$  if  $0 \leq P \leq P_1$  and  $\beta(a, P, t) - \lambda(a, P, t) \leq -\varepsilon$  for  $a, t \geq 0$  if  $P \geq P_2$ . Then there is  $T_0 \geq 0$ , depending only on  $\|\varphi\|, \varepsilon, P_1,$  and  $P_2$  such that  $P_1 \leq P(t; \varphi) \leq P_2$  for  $t \geq T_0$ .*

*Furthermore, if  $\beta_1\alpha_1(P) \leq \beta(a, P, t) \leq \beta_2\alpha_2(P)$  for  $a, P, t \geq 0$ , then*

$$(3.2) \quad \beta_1\alpha_1(P(t; \varphi))P(t; \varphi) \leq B(t; \varphi) \leq \beta_2\alpha_2(P(t; \varphi))P(t; \varphi) \text{ for } t \geq 0.$$

Equation (3.1) can also be used to show that the solutions of (2.6) behave as we would expect when the birth rate always exceeds the death rate or vice versa.

**THEOREM 2.** *If  $\tau = 0, H_0$  holds and  $\beta(a, P, t) - \lambda(a, P, t) \geq \varepsilon > 0$  for  $a, P, t \geq 0$ , then for any  $\varphi, P(t; \varphi) \rightarrow \infty$  as  $t \rightarrow \infty$ . If  $\beta(a, P, t) - \lambda(a, P, t) \leq -\varepsilon < 0$  for  $a, P, t \geq 0$ , then for any  $\varphi P(t; \varphi) \rightarrow 0$  as  $t \rightarrow \infty$ .*

When  $\tau > 0$ , the equations which generate the upper and lower bounds for solutions of (2.6) are delay differential equations. As would be expected, additional structure is required for  $\beta$  and  $\lambda$  to get the contraction of solutions described for  $\tau = 0$  in this case.

To this point the biological motivations for  $\beta$  and  $\lambda$  have been ignored. When space and/or natural resources are limited, a common reaction is that an increase in total population produces a decrease in the birth rate and/or an increase in the death rate. Theorem 1 can easily be applied in such a situation, and can just as easily be applied when this situation holds only for very small and very large values of  $P$ . With this condition in mind, we add the following assumptions about  $\beta$  and  $\lambda$ .

$H_1$ : There are functions  $\alpha, \gamma \in C(R^+, R^+)$ , where  $\alpha(P)$  is a nonincreasing and  $\alpha(P)P$  and  $\gamma(P)$  are nondecreasing functions of  $P$ , such that  $0 < \underline{\beta} \leq \beta_1\alpha(P) \leq \beta(a, P, t) \leq \beta_2\alpha(P) \leq \bar{\beta}$  and  $0 < \underline{\lambda} \leq \lambda_1\gamma(P) \leq \lambda(a, P, t) \leq \lambda_2\gamma(P) \leq \bar{\lambda}$  for  $a, P \geq 0$  and  $t \geq -\tau$ .

Since  $\varphi \in L_1(R^+ \times I_0)$  when  $\tau > 0$ , we take

$$\|\varphi\| = \sup_{t \in I_0} \int_0^\infty \varphi(a, t) da = \sup_{t \in I_0} P(t; \varphi).$$

We also define

$$\|\varphi\|_* = \inf_{t \in I_0} \int_0^\infty \varphi(a, t) da = \inf_{t \in I_0} P(t; \varphi).$$

The following lemma provides the information needed to establish contractive bounds for solutions of (2.6) when  $\tau > 0$ .

**LEMMA 1.** *If  $H_0 - H_1$  are satisfied and if  $\|\varphi\| \leq M$  and  $\varphi \leq M$ , then*

$$P'(t; \varphi) \leq \beta_2\alpha(P(t - \tau; \varphi))P(t - \tau; \varphi) - \lambda_1\gamma(P(t; \varphi))P(t; \varphi) + 2\bar{\beta}M \exp(-\underline{\lambda}t),$$

and

$$P'(t; \varphi) \cong \beta_1\alpha(P(t - \tau; \varphi))P(t - \tau; \varphi) - \lambda_2\gamma(P(t; \varphi))P(t; \varphi) - \underline{\beta}M \exp(-\underline{\lambda}t)$$

for  $t > 0$ .

**THEOREM 3.** *Suppose  $\tau \geq 0$ ,  $H_0 - H_1$  are satisfied,  $0 < M_1 \leq \|\varphi\|_* \leq \|\varphi\| \leq M_2$ ,  $\varphi \leq M_2$  and there are positive constants  $\varepsilon$ ,  $P_1 < P_2$  such that*

$$(3.3) \quad \beta_1\alpha(P_1) - \lambda_2\gamma(P_1) \geq \varepsilon \quad \text{and} \quad \beta_2\alpha(P_2) - \lambda_1\gamma(P_2) \leq -\varepsilon.$$

*Then there is  $T_0 > 0$ , depending only on  $M_1, M_2, \varepsilon, P_1$  and  $P_2$  such that*

$$P_1 \leq P(t; \varphi) \leq P_2 \quad \text{for } t \geq T_0,$$

and

$$\beta_1\alpha(P(t - \tau; \varphi))P(t - \tau; \varphi) \leq B(t; \varphi) \leq \beta_2\alpha(P(t - \tau; \varphi))P(t - \tau; \varphi), \quad t > 0.$$

A proof similar to that used to establish Theorem 3 can be used to extend Theorem 2 to the case  $\tau > 0$ .

**THEOREM 4.** *If  $H_1$  is satisfied then Theorem 2 holds for  $\tau \geq 0$ .*

One of the most concise, but yet realistic, realizations for  $\beta$  and  $\lambda$  of the assumption of a finite carrying capacity for the environment is the  $s$  shaped curve given by  $y = (bx + c)/(x + d)$ , where  $b, c, d > 0$ . We assume next that  $\beta$  and  $\lambda$  are bounded by such functions of  $P$  as follows:

$H_2$ : There are positive constants  $\beta_i, \lambda_i, i = 1, 2$ , and  $b_i, c_i, i = 1, 2, 3$ , such that  $\beta_1\alpha(P) \leq \beta(a, P, t) \leq \beta_2\alpha(P)$  and  $\lambda_1\gamma(P) \leq \lambda(a, P, t) \leq \lambda_2\gamma(P)$  for  $a, P \geq 0$ , and  $t \geq -\tau$ , where  $\alpha(P) = (b_0P + b_1)/(P + b_2)$ ,  $\gamma(P) = (c_0P + c_1)/(P + c_2)$ ,  $b_0b_2 \leq b_1$ , and  $c_1 \leq c_0c_2$ .

We will say that solutions of (2.6) uniformly approach the set  $[P_1, P_2] \times [B_1, B_2]$  if for any  $\varepsilon > 0$  and  $M > 1$  there is  $T_0 > 0$  such that if  $1/M \leq \|\varphi\|_* \leq \|\varphi\| \leq M$  and  $\varphi \leq M$  then  $(P(t; \varphi), B(t; \varphi)) \in [P_1 - \varepsilon, P_2 + \varepsilon] \times [B_1 - \varepsilon, B_2 + \varepsilon]$  for  $t \geq T_0$ .

The following is a direct consequence of Theorems 3 and 4 since  $H_2$  implies  $H_1$ .

**THEOREM 5.** *Assume  $\tau \geq 0$  and  $H_0, H_2$  are satisfied.*

- (i) *If  $b_1\beta_2/b_2 < c_1\lambda_1/c_2$ , then  $(P(t; \varphi), B(t; \varphi)) \rightarrow 0$  as  $t \rightarrow \infty$ .*
- (ii) *If  $b_0\beta_1 > c_0\lambda_2$ , then  $P(t; \varphi), B(t; \varphi) \rightarrow \infty$  as  $t \rightarrow \infty$ .*
- (iii) *If  $b_0\beta_2 < c_0\lambda_1$  and  $b_1\beta_1/b_2 > c_1\lambda_2/c_2$ , then the solutions of (2.6) uniformly approach the set  $[P_1, P_2] \times [\beta_1P_1\alpha(P_1), \beta_2P_2\alpha(P_2)]$ , where  $P_1, P_2$  are the unique positive solutions of  $\beta_1\alpha(P) = \lambda_2\gamma(P)$  and  $\beta_2\alpha(P) = \lambda_1\gamma(P)$  respectively.*

**4. Periodic solutions.** Since a periodic solution in a population model represents a state of equilibrium for the system described by the model, it has both theoretical and practical importance. We can use the preceding results along with an asymptotic fixed point theorem of Horn [6] to establish the existence of a periodic solution of (2.1)–(2.4).

We will need the following:

$H_3$ : The functions  $\beta$  and  $\lambda$  satisfy a Lipschitz condition in each variable for  $a, P \geq 0$ , and  $t \geq -\tau$ .

$H_4$ : There is  $\omega > 0$  such that  $\beta(a, P, t + \omega) = \beta(a, P, t)$ , and  $\lambda(a, P, t + \omega) = \lambda(a, P, t)$ , for  $a, P \geq 0$  and  $t \geq -\tau$ .

The key condition in the application of Horn's theorem is that solutions of (2.1)–(2.4) contract uniformly from a given set in  $L_1(R^+ \times [-\tau, 0])$  to a smaller set

contained within the original set. Necessary conditions for these contractions are contained in each of Theorems 1, 3 and 5. Condition  $H_3$  is a smoothness condition required for the application of Horn's theorem; it says that  $\beta$  and  $\lambda$  do not vary "too fast," a situation which is usually the case in nature.

**THEOREM 6.** *Suppose  $\tau \geq 0$ ,  $H_0, H_1, H_3$  and  $H_4$  are satisfied,  $\omega > \tau$ , and there are positive constants  $P_1 < P_2$  such that  $\beta_1\alpha(P_1) - \lambda_2\gamma(P_1) > 0$  and  $\beta_2\alpha(P_2) - \lambda_1\gamma(P_2) < 0$ . Then there is a positive periodic solution  $\rho(a, t; \bar{\varphi})$  of (2.1)–(2.4) of period  $\omega$ .*

Also from the proof of Theorem 6 we have:

**THEOREM 7.** *If the hypotheses of Theorem 6 are satisfied ( $H_4$  is not needed), then the solution  $\rho(a, t; \varphi)$  of (2.1)–(2.4) is a continuous functional of the function  $\varphi$ .*

Theorem 3 says that we need look only at  $(P, B) \in [P_1, P_2] \times [\beta_1 P_1 \alpha(P_1), \beta_2 P_2 \alpha(P_2)]$  to find the salient features of the model. Theorem 6 describes a state of equilibrium within this set. Two obvious questions are left unanswered; namely, is the periodic solution unique, and if it is, do nearby solutions converge to it as  $t \rightarrow \infty$ ?

**5. Periodic solutions when (2.1)–(2.3) is autonomous.** When  $\tau = 0$  and  $\beta$  and  $\lambda$  are independent of  $t$ , (2.1)–(2.4) reduces to the system derived by Gurtin and MacCamy [4]. When it is assumed further that  $\lambda(a, P) = \lambda(P)$ , these authors showed in [5] that the system reduces to systems of ordinary differential equations. In particular, they showed that if  $g$  satisfies  $g(a)\rho(a, t) \rightarrow 0$  as  $a \rightarrow \infty$ , and

$$G(t) = \int_0^\infty g(a)\rho(a, t) da, \quad H(t) = \int_0^\infty g'(a)\rho(a, t) da,$$

then

$$\dot{G} + \lambda(P)G - g(0)B = H.$$

If it is assumed that  $\lambda(a, P) = \lambda(P)$  and  $\beta(a, P) = \beta(P)e^{-\alpha a}$ ,  $\alpha > 0$ , then (2.1)–(2.3) reduces to the following pair of differential equations:

$$(5.1) \quad \begin{aligned} \dot{P} &= -\lambda(P)P + \beta(P)G, \\ \dot{G} &= [-\lambda(P) + \beta(P) - \alpha]G, \end{aligned}$$

where  $B = \beta(P)G$ .

When  $\beta(a, P) = \beta(a)ae^{-\alpha a}$ , (2.1)–(2.3) reduces to the system

$$(5.2) \quad \begin{aligned} \dot{P} &= -\lambda(P)P + \beta(P)A, \\ \dot{G} &= -[\lambda(P) + \alpha]G + \beta(P)A, \\ \dot{A} &= -[\lambda(P) + \alpha]A + G, \end{aligned}$$

where  $B = \beta(P)A$  and  $G = \int_0^\infty e^{-\alpha a}\rho(a, t) da$ .

It is shown in [5] that if  $\beta(P) = \beta_0$ , then neither (5.1) nor (5.2) can have closed orbits; i.e., (2.1)–(2.4) cannot have a periodic solution. The authors conjecture that the general system (5.2) has no closed orbits.

We show next that if  $\beta(P)$  is not constant, then closed orbits can occur for (5.1). It will be assumed that  $\lambda, \beta \in C^3(R_+)$ , and that  $\lambda(P) = \lambda_0 - q(P - 1) + \lambda_1(P - 1)$  and  $\beta(P) = \beta_0 - r(P - 1) + \beta_1(P - 1)$ ; here  $\lambda_0, \beta_0$  are positive,  $\lambda_1(0) = 0 = \lambda'_0(0)$ , and  $\beta_1(0) = 0 = \beta'_1(0)$ , i.e.,  $\lambda_1$  and  $\beta_1$  contain only higher order terms in  $(P - 1)$ . System (5.1) has the positive equilibrium  $(P, B) = (1, \lambda_0/\beta_0)$  if and only if  $\beta_0 = \alpha +$



$\lambda_0$ . Setting  $x = P - 1$  and  $y = G - \lambda_0/\beta_0$ , reduces (5.1) to

$$(5.3) \quad \begin{aligned} \dot{x} &= (q - \lambda_0 - \lambda_0 r/\beta_0)x + \beta_0 y + g_1(x, y), \\ \dot{y} &= (\lambda_0/\beta_0)(q - r)x + g_2(x, y), \end{aligned}$$

where  $g_1$  and  $g_2$  contain only higher order terms in  $x$  and  $y$ .

The characteristic equation of the linear part of (5.3) is  $z^2 + (\lambda_0 - q + \lambda_0 r/\beta_0)z + \lambda_0(r - q) = 0$ , which has roots

$$z = -\frac{1}{2} \left( \lambda_0 - q + \frac{\lambda_0 r}{\beta_0} \right) \pm \frac{1}{2} \sqrt{\left( \lambda_0 - q + \frac{\lambda_0 r}{\beta_0} \right)^2 - 4 \lambda_0(r - q)}.$$

If  $r > \beta_0 \lambda_0/\alpha$ , then  $r > (\lambda_0/\beta_0)(\beta_0 + r) \equiv q_*$ ;  $\text{Re}(z)$  is negative for  $q < q_*$  and positive for  $q > q_*$ , and at  $q = q_*$   $z = \pm i\sqrt{\lambda_0(r - q_*)}$ . Since  $d/dq \text{Re}(z) = \frac{1}{2}$ , a Hopf bifurcation occurs at  $q = q_*$  to a periodic solution of (5.1) of period  $\cong 2\pi/\sqrt{\lambda_0(r - q_*)}$ . See e.g. [9]. It is also easily seen that  $\text{Re}(z) < 0$  for  $q < q_*$  as long as  $z$  is complex, and both roots are negative if  $q < q_*$  when the roots are real.

**THEOREM 8.** *Suppose  $\lambda, \beta \in C^3(R_+)$ , and admit the preceding expansion where  $\lambda_0 > 0$ ,  $\beta_0 = \alpha + \lambda_0$  and  $r > \beta_0 \lambda_0/\alpha$ . Then the equilibrium  $(1, \lambda_0/\beta_0)$  is asymptotically stable for  $q < q_*$  and bifurcates at  $q = q_*$  to a periodic solution of (5.1) of period near  $2\pi/\sqrt{\lambda_0(r - q_*)}$ .*

Gurtin and MacCamy [5] assumed  $\beta(P) = \beta_0$ ; in that case  $r = 0$ , and it is not possible to get closed orbits, as they showed. As was noted earlier, one would usually expect  $\lambda$  to be an increasing function of  $P$ , while the periodic solutions occur here only when  $\lambda$  is a decreasing function of  $P$  near  $P = 1$ . There are, however, reasonable applications of this result, for example the case examined by Gurtin and MacCamy [5, Fig. 4], where it is assumed that  $\lambda$  is large for small and large  $P$  and relatively small between these extreme values of  $P$ .

It is not known whether (2.6) can have periodic solutions when (2.6) is autonomous,  $\lambda$  is an increasing and  $\beta$  a decreasing function of  $P$ .

**Appendix.**

*Proof of Lemma 1.* Equation (2.6b) can be rewritten in the form

$$\begin{aligned} B(t + \tau) &= \int_0^t \beta(t - a, P(t), t)B(a) \exp \left[ - \int_a^t \lambda(s - a, P(s), s) ds \right] da \\ &+ \int_0^\infty \beta(t + a, P(t), t)\varphi(a, 0) \exp \left[ - \int_0^t \lambda(s + a, P(s), s) ds \right] da \\ &+ \int_{-\tau}^0 \beta(t - a, P(t), t)\varphi(0, a) \exp \left[ - \int_a^t \lambda(s - a, P(s), s) ds \right] da \\ &+ \int_0^\infty \beta(t + \tau + a, P(t), t)\varphi(a, -\tau) \\ &\quad \cdot \exp \left[ - \int_0^{t+\tau} \lambda(s + a, P(s - \tau), s - \tau) ds \right] da \\ &- \int_0^\infty \beta(t + a, P(t), t)\varphi(a, 0) \exp \left[ - \int_0^t \lambda(s + a, P(s), s) ds \right] da. \end{aligned}$$

Since  $\lambda \cong \underline{\lambda} > 0$ ,  $0 < \underline{\beta} \leq \beta \leq \bar{\beta}$ ,  $0 \leq \varphi \leq M$ , and  $\|\varphi\| \leq M$ , we get the inequalities given in Lemma 1.

*Proof of Theorem 4.* For a fixed  $\varphi$ , consider  $P(t) = P(t; \varphi)$  for  $t \geq 0$ . It follows

from (3.3), Lemma 1, and the fact that  $\alpha(P)$  is a nonincreasing and  $\alpha(P)P$  and  $\gamma(P)$  are nondecreasing functions of  $P$ , that  $P'(t) \leq -\varepsilon P_2$  if  $P(t) \geq P_2$  and  $P(t - \tau) \leq P(t)$ . Now  $f(x, y) = \beta_2\alpha(x)x - \lambda_1\gamma(y)y$  is a continuous function of  $x$  and  $y$ , and since  $f(x, y) \leq -\varepsilon P_2$  for  $0 \leq x \leq y$  and  $P_2 \leq y \leq M$ , it follows that there is  $\delta > 0$  such that  $f(x, y) \leq -P_2\varepsilon/2$  if  $x \leq y + \delta$  and  $P_2 \leq y \leq M$ .

Now consider  $P(t)$  for  $t \geq t_0 > \tau$ , and assume that  $\sup_{s \in I_0} P(t_0 + s) > P_2$ . At each point  $t > t_0$ , one of the following must hold: 1)  $P(t) < P_2$ , 2)  $P(t) < P(t - \tau) - \delta$  or 3)  $P'(t) \leq -P_2\varepsilon/2$ . It follows that if  $\eta = \min(\delta, P_2\varepsilon/2)$  and  $P_* = \max(P_2, \sup_{s \in I_0} P(t_0 + s) - \eta)$ , then  $P(t) < \sup_{s \in I_0} P(t_0 + s)$  for  $t > t_0$ , and  $P(t) \leq P_*$  for  $t \geq t_0 + \tau$ . This argument can be repeated sufficiently often, (the next iteration would start at  $t = t_0 + 2\tau$ ), to find  $T_1 > T_0$ , depending only on  $\varepsilon, P_2$  and  $M_2$ , such that  $P(t) < P_2$  for  $T_1 - \tau \leq t \leq T_1$ . Suppose now that  $P(t) = P_2$  for some  $t_1 > T_1$ , and we can assume that  $P(t) < P_2$  for  $T_1 - \tau \leq t < t_1$ . Then, since  $P(t_1 - \tau) < P_2 = P(t_1)$ , it follows that  $P'(t_1) < -P_2\varepsilon/2$ , which is clearly impossible in view of the choice of  $t_1$ . It follows that  $P(t) < P_2$  for  $t > T_1$ .

The proof establishing the lower bound is similar.

*Proof of Theorem 3.* Since  $H_2$  implies  $H_1$ , it is easily seen that i) and ii) follow from Theorem 2, and that Theorem 3 and the inequalities given in iii) imply that the solutions of (2.6) approach a set of the form given. To see that  $P_1$  and  $P_2$  give such a set, note that after simplifying, the equations  $\beta_1\alpha(P) = \lambda_2\gamma(P)$  and  $\beta_2\alpha(P) = \lambda_1\gamma(P)$  can be put in the form  $A_1P^2 + A_2P + A_3 = 0$  and  $A_4P^2 + A_5P + A_6 = 0$ , respectively. Using  $H_0$  and the inequalities given in iii), we can show that  $A_1 < A_4 < 0 < A_3 < A_6$ . It follows that each equation has one positive root, and that  $P_1 < P_2$ .

To establish Theorem 6, we need the following:

LEMMA 2. Horn [6]. Let  $S_0 \subset S_1 \subset S_2$  be convex subsets of a Banach space  $X$  with  $S_0, S_2$  compact and  $S_1$  open in  $S_2$ . Let  $T : S_2 \rightarrow X$  be a continuous mapping such that for some integer  $m > 0, T^j(S_1) \subset S_2, 0 \leq j \leq m - 1, T^j(S_1) \subset S_0, m \leq j \leq 2m - 1$ . Then  $T$  has a fixed point.

*Proof of Theorem 6.* Let  $X$  be the Banach space of continuous functions in  $L_1(R^+ \times I_0)$ . Since the hypotheses of Theorem 6 satisfy Theorem 3, if  $0 < M_1 < P_1 < P_2 < M_2$  and if  $\varphi \in X$  with  $M_1 \leq \|\varphi\|_* \leq \|\varphi\| \leq M_2$ , then  $M_1 \leq P(t; \varphi) \leq M_2$  and  $\beta_1M_1\alpha(M_1) \leq B(t; \varphi) \leq \beta_2M_2\alpha(M_2)$  for  $t \geq 0$ . It follows that  $M_1 \leq \|\rho_t(\varphi)\|_* \leq \|\rho_t(\varphi)\| \leq M_2$  for  $t \geq 0$ , and from (2.5) that  $0 \leq \rho(a, t; \varphi) \leq M_4$ ; here  $M_4 = \max(M_2, \beta_2M_2\alpha(M_2))$ , if  $0 \leq \varphi \leq M_4$ , and  $\rho_t(\varphi) \in X$  is defined by  $\rho_t(\varphi) = \rho(a, t + s; \varphi)$  for  $a \geq 0$  and  $s \in I_0$ .

Let  $S = \{\varphi \in X \mid \|\varphi\| \leq M_2 \text{ and } 0 \leq \varphi \leq M_4\}$ , and define  $T : S \rightarrow S$  by  $T\varphi = \rho_\omega(\varphi)$ . We also define the sets

$$S_2(K) = \{\varphi \in S \mid |\varphi(a_1, t_1) - \varphi(a_2, t_2)| \leq K(|a_1 - a_2| + |t_1 - t_2|) \text{ for } a_1, a_2 \geq 0 \text{ and } t_1, t_2 \in I_0\},$$

where  $K$  is a positive constant to be determined later in the proof,

$$S_1 = \{\varphi \in S_2(K) \mid M_1 < \|\varphi\|_* \leq \|\varphi\| < M_2\},$$

and

$$S_0 = \{\varphi \in S_2(K) \mid P_1 \leq \|\varphi\|_* \leq \|\varphi\| \leq P_2\}.$$

It follows from Theorem 3 that there is  $m > 0$  such that  $P_1 \leq \|T^j\varphi\|_* \leq \|T^j\varphi\| \leq P_2$  for  $j \geq m$ , if  $\varphi \in S$  and  $M_1 \leq \|\varphi\|_* \leq \|\varphi\| \leq M_2$ .

Since the proof that  $T$  is continuous and that  $\rho$  satisfies the appropriate Lipschitz condition involves rather lengthy calculations, we only sketch the main ideas.

Let  $(P(t; \varphi), B(t; \varphi))$  and  $(P(t; \psi), B(t; \psi))$  be solutions of (2.6), with  $\varphi, \psi \in S$ , and set  $x_1(t) = |P(t; \varphi) - P(t; \psi)|$  and  $y_1 = |B(t; \varphi) - B(t; \psi)|$  for  $t \geq -\tau$ . One can show that there is  $K_1 > 0$ , such that for  $0 \leq t \leq \omega$ ,

$$\begin{aligned} x_1(t) &\leq \int_0^t \left\{ y_1(a) + K_1 \int_a^t x_1(s) ds \right\} da \\ &\quad + \int_0^\infty \left\{ |\varphi(a, 0) - \psi(a, 0)| + \psi(a, 0) \int_0^t x_1(s) ds \right\} da, \\ y_1(t) &\leq K_1 \int_0^t \left\{ x_1(t - \tau) + y_1(a - \tau) + \int_a^t x_1(s - \tau) ds \right\} da \\ &\quad + K_1 \int_0^\infty \left\{ |\varphi(a, -\tau) - \psi(a, -\tau)| \right. \\ &\quad \left. + \varphi(a, -\tau)x_1(t - \tau) + \psi(a, -\tau) \int_0^t x_1(s - \tau) ds \right\} da \end{aligned}$$

or, there is  $K_2 > 0$  such that

$$\begin{aligned} (A.1) \quad x_1(t) &\leq \|\varphi - \psi\| + K_2 \int_0^t \left\{ x_1(a) + y_1(a) + \int_a^t x_1(s) ds \right\} da, \\ y_1(t) &\leq K_2 \|\varphi - \psi\| + K_2 \int_0^t \left\{ x_1(t - \tau) + y_1(a - \tau) + \int_a^t x_1(s - \tau) ds \right\} da. \end{aligned}$$

If  $(x, y)$  is the solution of (A.1) with “=” replacing “ $\leq$ ”, then

$$\begin{aligned} (A.2) \quad \dot{x}(t) &= K_2(x(t) + y(t) + tx(t)), \\ \dot{y}(t) &= K_2(K_2 + 1)(t + 1)(x(t - \tau) + y(t - \tau)), \\ x(t) &= \|\varphi - \psi\|, y(t) = K_2 \|\varphi - \psi\|, \quad -\tau \leq t \leq 0. \end{aligned}$$

Since the solutions of (A.2) exist for  $t \geq 0$  and depend continuously on the initial function, there is  $K_3 > 0$  such that

$$\begin{aligned} (A.3) \quad |P(t; \varphi) - P(t; \psi)| &\leq K_3 \|\varphi - \psi\|, \\ |B(t; \varphi) - B(t; \psi)| &\leq K_3 \|\varphi - \psi\|, \quad t \geq 0. \end{aligned}$$

Now, from (2.5),

$$\begin{aligned} \|T\varphi - T\psi\| &= \sup_{s \in I_0} \int_0^\infty |\rho(a, \omega + s; \varphi) - \rho(a, \omega + s; \psi)| da \\ &= \sup_{s \in I_0} \int_0^{\omega+s} \left| B(\omega + s - a; \varphi) \exp \left( - \int_0^a \lambda(\varphi) ds \right) \right. \\ &\quad \left. - B(\omega + s - a; \psi) \exp \left( - \int_0^a \lambda(\psi) ds \right) \right| da \\ &\quad + \sup_{s \in I_0} \int_{\omega+s}^\infty \left| \varphi(a - \omega - s, 0) \exp \left( - \int_0^{\omega+s} \lambda(\varphi) ds \right) \right. \\ &\quad \left. - \psi(a - \omega - s, 0) \exp \left( - \int_0^{\omega+s} \lambda(\psi) ds \right) \right| da. \end{aligned}$$

(A.3) can be used to find a positive constant  $K_4$  such that each of these integrals is bounded by  $K_4\|\varphi - \psi\|$ , which establishes the continuity of  $T$ .

Since  $\alpha$  and  $\gamma$  are continuous, it follows from  $H_1$  that there is  $K_5 > 0$  such that  $|\beta(a, P, t) - \lambda(a, P, t)| \leq K_5$  for  $P \in S$ , which, with (3.1) shows that there is  $K_6 > 0$  such that

$$(A.4) \quad |P'(t; \varphi)| \leq K_6 \quad \text{for } \varphi \in S.$$

Using (A.4), one can show that there is  $K_7 > 0$  such that for  $\varphi \in S$ ,

$$(A.5) \quad |B(t_1; \varphi) - B(t_2; \varphi)| \leq K_7|t_1 - t_2|.$$

We need to show not only that  $\rho(a, \omega + s; \varphi)$  is equicontinuous for  $s \in I_0$  and  $a \geq 0$  on  $S_2(K)$ , but that there is  $K > 0$  such that if  $\varphi \in S_2(K)$ , then  $|\rho(a_1, \omega + s_1; \varphi) - \rho(a_2, \omega + s_2; \varphi)| \leq K(|a_1 - a_2| + |s_1 - s_2|)$  for  $a_1, a_2 \geq 0$  and  $s_1, s_2 \in I_0$ .

Using (2.5), (A.4) and (A.5), we can show that there is  $K_8 > 0$ , which is independent of  $K$  since it is valid for all  $\varphi \in S$ , such that

$$\begin{aligned} & |\rho(a_1, \omega + s_1; \varphi) - \rho(a_2, \omega + s_2; \varphi)| \\ & \leq \exp(-\underline{\lambda}(\omega - \tau))(K + K_8)(|a_1 - a_2| + |t_1 - t_2|) \quad \text{for } a_1, a_2 \geq 0, \quad s_1, s_2 \in I_0. \end{aligned}$$

If we select  $K > K_8/[\exp(\underline{\lambda}(\omega - \tau)) - 1]$ , then we have  $T^j(S_1) \subset S_2(K) \quad j \geq 1$ , and  $T^j(S_1) \subset S_0$  for  $j \geq m$ . It follows now from Lemma 2 that  $T$  has a fixed point  $\tilde{\varphi}$ , and clearly  $\rho(a, t; \tilde{\varphi})$  is the periodic solution sought.

REFERENCES

[1] R. A. EASTERLIN, *The current fertility decline and projected fertility changes*, in Population, Labor Force and Long Swings in Economic Growth: The American Experience, Columbia University Press, New York, 1968.

[2] J. C. FRAUENTHAL, *A dynamic model for human population growth*, Theoret. Population Biology, 8 (1975), pp. 64-73.

[3] D. H. GRIFFEL, *Age-dependent population growth*, J. Inst. Math. Appl., 17 (1976), pp. 141-152.

[4] M. E. GURTIN AND R. C. MACCAMY, *Non-linear age-dependent population dynamics*, Arch. Rational Mech. Anal., 54 (1974), pp. 281-300.

[5] ———, *Some simple models for nonlinear age-dependent population dynamics*, Math Biosci. 43 (1979), pp. 199-211.

[6] W. A. HORN, *Some fixed point theorems for compact mappings and flows on a Banach space*, Trans. Amer. Math. Soc., 149 (1970), pp. 391-404.

[7] N. KEYFITZ, *Introduction to the Mathematics of Population*, Addison-Wesley, Reading, MA., 1968.

[8] A. J. LOTKA, *Théorie analytique des associations biologiques, Part II, Analyse démographique avec applications particulière à l'espèce humaine*, Actualités Scientifiques et Industrielles. No. 780 (1939), Hermann and Cie, Paris.

[9] J. E. MARSDEN AND M. MCCRAKEN, *The Hopf Bifurcation and its Applications*, Springer-Verlag, New York, 1976.

[10] C. RORRES, *Stability of an age specific population with density dependent fertility*, Theoret. Population Biology, 10 (1976), pp. 26-46.

[11] H. L. SMITH, *On periodic solutions of delay integral equations modelling epidemics and population growth*, Ph.D. Thesis, University of Iowa, Iowa City IO, 1976.

[12] K. E. SWICK, *A model of single species population growth*, this Journal, 7 (1976), pp. 565-576.

[13] ———, *A non-linear age-dependent model of single species population dynamics*, SIAM J. Appl. Math., 32 (1977), pp. 484-498.

[14] H. VON FOERSTER, *Some remarks on changing populations*, in The Kinetics of Cellular Proliferation, Grune and Stratton, New York, 1959, pp. 382-407.

**NOTE ON THE ASYMPTOTIC BEHAVIOR OF MULTIDIMENSIONAL  
 LAPLACE INTEGRALS\***

L. A. SKINNER†

**Abstract.** The leading terms of a uniformly valid asymptotic expansion for a class of multidimensional integrals of Laplace type are obtained by a procedure derived from singular perturbation theory. The expansion describes the smooth transition between distinctly different forms of asymptotic behavior which occurs when the critical point of the integrand crosses the boundary of the domain of integration.

**1. Introduction.** This paper is concerned with the asymptotic evaluation of integrals in  $m$ -dimensional Euclidean space  $E^m$  of the form

$$(1.1) \quad I(\mathbf{x}, \nu) = \left(\frac{\nu}{2\pi}\right)^{m/2} \int_D g(\mathbf{x}, \boldsymbol{\xi}) \exp[-\nu h(\mathbf{x}, \boldsymbol{\xi})] d\boldsymbol{\xi},$$

as  $\nu \rightarrow \infty$ . We shall assume

- (i)  $g(\mathbf{x}, \boldsymbol{\xi})$  and  $h(\mathbf{x}, \boldsymbol{\xi})$  are real-valued functions of class  $C^\infty$  on  $E^m \times E^m$ ;
- (ii)  $h(\mathbf{x}, \boldsymbol{\xi}) > h(\mathbf{x}, \mathbf{x}) = 0$  for  $\boldsymbol{\xi} \neq \mathbf{x}$ , and (1.1) converges absolutely for sufficiently large  $\nu$ ;
- (iii) the matrix  $A(\mathbf{x})$  of partial derivatives  $a_{ij}(\mathbf{x}) = h_{\xi_i \xi_j}(\mathbf{x}, \mathbf{x})$  is positive definite;
- (iv)  $\partial D$ , the  $(m - 1)$ -dimensional boundary of  $D$ , is of class  $C^\infty$ .

Thus for each  $\mathbf{x} \in E^m$ , the minimum of  $h(\mathbf{x}, \boldsymbol{\xi})$  occurs at  $\boldsymbol{\xi} = \mathbf{x}$ , and it is a simple minimum. Also,  $\partial D$  has a well-defined normal at every point. There is no loss of generality in assuming  $h(\mathbf{x}, \mathbf{x}) = 0$ .

Under the above conditions, as has essentially been known for some time (cf. [3]),

$$(1.2) \quad I(\mathbf{x}, \nu) = \gamma(\mathbf{x}) + O(\nu^{-1})$$

if  $\mathbf{x}$  is an interior point of  $D$ , and

$$(1.3) \quad I(\mathbf{x}, \nu) = \frac{1}{2} \gamma(\mathbf{x}) + O(\nu^{-1/2})$$

if  $\mathbf{x} \in \partial D$ , where

$$(1.4) \quad \gamma(\mathbf{x}) = |A(\mathbf{x})|^{-1/2} g(\mathbf{x}, \mathbf{x}).$$

Also, if  $\mathbf{x}$  is an exterior point of  $D$  then, clearly,

$$(1.5) \quad I(\mathbf{x}, \nu) = o(\nu^{-\infty}),$$

i.e.,  $I(\mathbf{x}, \nu) = o(\nu^{-N})$  as  $\nu \rightarrow \infty$  for any  $N$ . Bleistein and Handelsman [2] have recently developed a method of integration by parts which establishes the general form of the asymptotic expansions corresponding to (1.2) and (1.3). However, due to the complexity of the calculations required for their results, the next term of (1.2) is the only additional one they determined explicitly.

As in the one-dimensional analogue of (1.1), which has been studied by Bleistein [1] and Wong [5], the transitions from (1.2) to (1.3) to (1.5) as  $\mathbf{x}$  approaches and then crosses  $\partial D$  do not happen abruptly. There are boundary layers on either side of  $\partial D$ . Together they form a shock layer. Our objective in this paper is to give a precise

\* Received by the editors June 5, 1979, and in final revised form February 25, 1980.

† Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

description of this boundary layer behavior. In other words, we seek an asymptotic expansion for  $I(\mathbf{x}, \nu)$  which is uniformly valid in  $\mathbf{x}$ . In the process we shall establish the next term of (1.3), and it will be clear that our procedure also could be employed to obtain more terms for both (1.2) and (1.3).

The basis for our analysis is the following theorem, which is a straightforward generalization of the theorem given in [4] to study Laplace integrals in  $E^1$  with coalescing saddle points.

**THEOREM 1.** *Let  $\phi(\zeta, \mathbf{Z}, \mathbf{x})$  be a real-valued function of class  $C^\infty$  on  $B \times E^m \times \Delta$ , where  $B$  and  $\Delta$  are closed subsets of  $E^m$  with  $\mathbf{0} \in B$ , and let  $\Phi_k(\mathbf{Z}, \mathbf{x})$  denote the coefficient of  $\nu^{-k/2}$  in the expansion of  $\phi(\nu^{-1/2}\mathbf{Z}, \mathbf{Z}, \mathbf{x})$  as  $\nu \rightarrow \infty$ . If  $\phi(\zeta, \mathbf{Z}, \mathbf{x})$  is uniformly  $o(|\mathbf{Z}|^{-\infty})$  as  $|\mathbf{Z}| \rightarrow \infty$ , then*

$$(1.6) \quad \phi(\zeta, \nu^{1/2}\zeta, \mathbf{x}) = \sum_{k=0}^{N-1} \nu^{-k/2} \Phi_k(\nu^{1/2}\zeta, \mathbf{x}) + O(\nu^{-N/2})$$

uniformly, for all  $(\zeta, \mathbf{x}) \in B \times \Delta$ , as  $\nu \rightarrow \infty$ .

This theorem has an immediate interpretation in singular perturbation theory. The assumption  $\phi(\zeta, \mathbf{Z}, \mathbf{x}) = o(|\mathbf{Z}|^{-\infty})$  is equivalent to assuming that the (outer) expansion of  $w(\zeta, \mathbf{x}, \varepsilon) = \phi(\zeta, \varepsilon^{-1}\zeta, \mathbf{x})$  with respect to the asymptotic sequence  $\{1, \varepsilon, \varepsilon^2, \dots\}$  is identically zero. Thus (1.6) simply says that, this being the case, the  $N$ -term inner expansion of  $w(\zeta, \mathbf{x}, \varepsilon)$ , i.e., the expansion to order  $\varepsilon^{N-1}$  of  $w(\varepsilon\mathbf{Z}, \mathbf{x}, \varepsilon)$  followed by the substitution  $\mathbf{Z} = \varepsilon^{-1}\zeta$ , is uniformly valid.

**2. Preliminary transformations.** Since  $A(\mathbf{x})$  is a symmetric positive definite matrix of class  $C^\infty$  on  $E^m$ , all its eigenvalues  $\lambda_1(\mathbf{x}), \dots, \lambda_m(\mathbf{x})$  are positive, and there exists an orthogonal matrix  $Q(\mathbf{x})$  of class  $C^\infty$  such that

$$(2.1) \quad Q^T(\mathbf{x})A(\mathbf{x})Q(\mathbf{x}) = \text{diag} [\lambda_1(\mathbf{x}), \dots, \lambda_m(\mathbf{x})].$$

As in [2], let

$$(2.2) \quad R(\mathbf{x}) = \text{diag} [\lambda_1^{-1/2}(\mathbf{x}), \dots, \lambda_m^{-1/2}(\mathbf{x})].$$

Then  $P(\mathbf{x}) = Q(\mathbf{x})R(\mathbf{x})S(\mathbf{x})$ , where  $S(\mathbf{x})$  is another  $C^\infty$  orthogonal matrix, satisfies

$$(2.3) \quad P^T(\mathbf{x})A(\mathbf{x})P(\mathbf{x}) = I,$$

and

$$(2.4) \quad C(\mathbf{x}) = P(\mathbf{x})P^T(\mathbf{x}) = A^{-1}(\mathbf{x}).$$

The first step in our analysis of (1.1) is to change the variable of integration from  $\xi$  to  $\zeta$  by setting

$$(2.5) \quad \xi = \mathbf{x} + P(\mathbf{x})\zeta.$$

The Jacobian determinant of this transformation is  $|A(\mathbf{x})|^{-1/2}$  and, in view of (2.3),

$$(2.6) \quad h(\mathbf{x}, \mathbf{x} + P(\mathbf{x})\zeta) = \frac{1}{2} \zeta_i \zeta_j \delta_{ij} + \frac{1}{6} \zeta_i \zeta_j \zeta_k \eta_{ijk}(\mathbf{x}) + O(|\zeta|^4),$$

where

$$(2.7) \quad \eta_{ijk}(\mathbf{x}) = p_{ri}(\mathbf{x})p_{sj}(\mathbf{x})p_{tk}(\mathbf{x})h_{\xi_r, \xi_s, \xi_t}(\mathbf{x}, \mathbf{x}).$$

Here we are employing the convention that repeated italic indices denote summation from 1 to  $m$ . Similarly, we shall use Greek letters for indices which range from 2 to  $m$ . Thus, for example,

$$(2.8) \quad \zeta_i \zeta_j \delta_{ij} = \zeta_1^2 + \zeta_\alpha \zeta_\alpha = |\zeta|^2.$$

We also have

$$(2.9) \quad G(\mathbf{x}, \boldsymbol{\zeta}) = \gamma(\mathbf{x}) + \zeta_i \gamma_i(\mathbf{x}) + O(|\boldsymbol{\zeta}|^2),$$

where

$$(2.10) \quad G(\mathbf{x}, \boldsymbol{\zeta}) = |A(\mathbf{x})|^{-1/2} g(\mathbf{x}, \mathbf{x} + P(\mathbf{x}) \boldsymbol{\zeta})$$

and

$$(2.11) \quad \gamma_i(\mathbf{x}) = |A(\mathbf{x})|^{-1/2} p_{ri}(\mathbf{x}) g_{\xi}(\mathbf{x}, \mathbf{x}).$$

Furthermore, from (2.6), since

$$(2.12) \quad O(|\boldsymbol{\zeta}|^4) = \zeta_i \zeta_j \delta_{ij} O(|\boldsymbol{\zeta}|^2),$$

there exists a  $C^\infty$  matrix  $U(\mathbf{x}, \boldsymbol{\zeta})$  such that

$$(2.13) \quad h(\mathbf{x}, \mathbf{x} + P(\mathbf{x}) \boldsymbol{\zeta}) = \zeta_i \zeta_j u_{ij}(\mathbf{x}, \boldsymbol{\zeta}),$$

and the elements of  $U(\mathbf{x}, \boldsymbol{\zeta})$  satisfy

$$(2.14) \quad u_{ij}(\mathbf{x}, \boldsymbol{\zeta}) = \frac{1}{2} \delta_{ij} + \frac{1}{6} \zeta_k \eta_{ijk}(\mathbf{x}) + O(|\boldsymbol{\zeta}|^2).$$

Now let  $\Delta = \{\mathbf{x}: |\mathbf{x} - \mathbf{x}_0| \leq \delta\}$ , where  $\mathbf{x}_0$  is a point on the boundary of  $D$ , and let  $B = \{\boldsymbol{\zeta}: |\zeta_i| \leq b\}$ . The reason for including  $S(\mathbf{x})$  in the definition of  $P(\mathbf{x})$  is that for all  $\mathbf{x}$  in a sufficiently small neighborhood of  $\mathbf{x}_0$  we can define  $S(\mathbf{x})$  so that the  $\zeta_1$  axis is always perpendicular to  $\partial D(\mathbf{x})$ , where  $D(\mathbf{x})$  is the image of  $D$  under (2.5). This allows us to express  $\partial D(\mathbf{x})$  near the image of  $\mathbf{x}_0$  as  $\zeta_1 = f(\mathbf{x}, \boldsymbol{\zeta}')$ , where  $\boldsymbol{\zeta}' = (\zeta_2, \zeta_3, \dots, \zeta_m)$ . Let us agree to orient the  $\zeta_1$ -axis so that  $\sigma(\mathbf{x}) = f(\mathbf{x}, \mathbf{0}') < 0$  when  $\mathbf{x}$  is an interior point of  $D$ , and  $\sigma(\mathbf{x}) > 0$  when  $\mathbf{x}$  is an exterior point. Then, provided  $\delta > 0$  is sufficiently small, there exists  $b > \delta$  such that  $B \cap D(\mathbf{x}) = \{\boldsymbol{\zeta} : f(\mathbf{x}, \boldsymbol{\zeta}') \leq \zeta_1 \leq b, |\zeta_\alpha| \leq b\}$ ; hence

$$(2.15) \quad I(\mathbf{x}, \nu) = \left(\frac{\nu}{2\pi}\right)^{m/2} \int_{B \cap D(\mathbf{x})} \phi(\boldsymbol{\zeta}, \nu^{1/2} \boldsymbol{\zeta}, \mathbf{x}) d\boldsymbol{\zeta} + o(\nu^{-\infty}),$$

for all  $\mathbf{x} \in \Delta$ , where

$$(2.16) \quad \phi(\boldsymbol{\zeta}, \mathbf{Z}, \mathbf{x}) = G(\mathbf{x}, \boldsymbol{\zeta}) \exp[-Z_i Z_j u_{ij}(\mathbf{x}, \boldsymbol{\zeta})].$$

The desired uniformly valid expansion of  $I(\mathbf{x}, \nu)$  can now be obtained by expanding  $\phi(\boldsymbol{\zeta}, \nu^{1/2} \boldsymbol{\zeta}, \mathbf{x})$  according to Theorem 1 and integrating term by term.

Referring to (2.9) and (2.14) we see for (2.16) that

$$(2.17) \quad \Phi_0(\mathbf{Z}, \mathbf{x}) = \gamma(\mathbf{x}) \exp\left(-\frac{1}{2} |\mathbf{Z}|^2\right)$$

and

$$(2.18) \quad \Phi_1(\mathbf{Z}, \mathbf{x}) = [Z_i \gamma_i(\mathbf{x}) - \frac{1}{6} Z_i Z_j Z_k \eta_{ijk}(\mathbf{x}) \gamma(\mathbf{x})] \exp\left(-\frac{1}{2} |\mathbf{Z}|^2\right).$$

Let

$$(2.19) \quad F(\mathbf{x}, \nu) = \left(\frac{\nu}{2}\right)^{m/2} \int_{B \cap D(\mathbf{x})} \exp\left(-\frac{1}{2} \nu |\boldsymbol{\zeta}|^2\right) d\boldsymbol{\zeta},$$

and, in analogy with (2.13) and (2.14), note that

$$(2.20) \quad f(\mathbf{x}, \boldsymbol{\zeta}') = \sigma(\mathbf{x}) + \zeta_\alpha v_\alpha(\mathbf{x}, \boldsymbol{\zeta}'),$$

where

$$(2.21) \quad v_\alpha(\mathbf{x}, \boldsymbol{\zeta}') = \frac{1}{2} \zeta_\beta \sigma_{\alpha\beta}(\mathbf{x}) + O(|\boldsymbol{\zeta}'|^2)$$

and

$$(2.22) \quad \sigma_{\alpha\beta}(\mathbf{x}) = f_{\zeta_\alpha \zeta_\beta}(\mathbf{x}, \mathbf{0}').$$

In the next section we will show that

$$(2.23) \quad F(\mathbf{x}, \nu) = \frac{1}{2} \operatorname{erfc} [(\nu/2)^{1/2} \sigma(\mathbf{x})] - \frac{1}{2} (2\pi\nu)^{-1/2} \sigma_{\alpha\alpha}(\mathbf{x}) \exp \left[ -\frac{1}{2} \nu\sigma^2(\mathbf{x}) \right] + O(\nu^{-1}).$$

Similarly,

$$(2.24) \quad F_i(\mathbf{x}, \nu) = \left(\frac{\nu}{2\pi}\right)^{m/2} \int_{B \cap D(\mathbf{x})} \nu^{1/2} \zeta_i \exp \left(-\frac{1}{2} \nu|\zeta|^2\right) d\zeta = (2\pi)^{-1/2} \delta_{i1} \exp \left[-\frac{1}{2} \nu\sigma^2(\mathbf{x})\right] + O(\nu^{-1/2}),$$

and

$$(2.25) \quad F_{ijk}(\mathbf{x}, \nu) = \left(\frac{\nu}{2\pi}\right)^{m/2} \int_{B \cap D(\mathbf{x})} \nu^{3/2} \zeta_i \zeta_j \zeta_k \exp \left(-\frac{1}{2} \nu|\zeta|^2\right) d\zeta = (2\pi)^{-1/2} \{\lambda_{ijk} - \mu_{ijk}[\nu\sigma^2(\mathbf{x}) - 1]\} \exp \left[-\frac{1}{2} \nu\sigma^2(\mathbf{x})\right] + O(\nu^{-1/2})$$

where

$$(2.26) \quad \lambda_{ijk} = \delta_{ij} \delta_{k1} + \delta_{jk} \delta_{i1} + \delta_{ik} \delta_{j1}, \quad \mu_{ijk} = \delta_{i1} \delta_{j1} \delta_{k1}.$$

Also, for any  $k$ , because of the exponential factor  $\exp(-\frac{1}{2} \nu|\zeta|^2)$ ,

$$(2.27) \quad \left(\frac{\nu}{2\pi}\right)^{m/2} \int_{B \cap D(\mathbf{x})} \Phi_k(\nu^{1/2} \zeta, \mathbf{x}) d\zeta = O(1).$$

Therefore, collecting these results together, we obtain

$$(2.28) \quad I(\mathbf{x}, \nu) = \frac{1}{2} \gamma(\mathbf{x}) \operatorname{erfc} \left[\left(\frac{\nu}{2}\right)^{1/2} \sigma(\mathbf{x})\right] + \nu^{-1/2} \theta(\mathbf{x}, \nu^{1/2} \sigma(\mathbf{x})) + O(\nu^{-1}),$$

where

$$(2.29) \quad \theta(\mathbf{x}, \Sigma) = (2\pi)^{1/2} \{\gamma_1(\mathbf{x}) - \frac{1}{2} \gamma(\mathbf{x})[\sigma_{\alpha\alpha}(\mathbf{x}) + \eta_{1kk}(\mathbf{x}) + \frac{1}{3} (\Sigma^2 - 1)\eta_{111}(\mathbf{x})]\} \exp \left(-\frac{1}{2} \Sigma^2\right).$$

**3. Evaluation of leading terms.** Note that

$$(3.1) \quad \int_{f(\mathbf{x}, \zeta')}^b \exp \left(-\frac{1}{2} \nu \zeta_1^2\right) d\zeta_1 = \frac{1}{2} \left(\frac{2\pi}{\nu}\right)^{1/2} \operatorname{erfc} \left[\left(\frac{\nu}{2}\right)^{1/2} f(\mathbf{x}, \zeta')\right] + o(\nu^{-\infty}),$$

for all  $(\mathbf{x}, \zeta') \in \Delta \times B'$  where  $B' = \{\zeta' : |\zeta_\alpha| \leq b\}$ . Thus, if we set

$$(3.2) \quad \psi(\zeta', \mathbf{Z}', \mathbf{x}, \Sigma) = \operatorname{erfc} \{2^{-1/2} [\Sigma + Z_\alpha v_\alpha(\mathbf{x}, \zeta')]\} \exp \left(-\frac{1}{2} |\mathbf{Z}'|^2\right)$$

then, from (2.19),



$$(3.3) \quad F(\mathbf{x}, \nu) = \frac{1}{2} \left( \frac{\nu}{2\pi} \right)^{m'/2} \int_{B'} \psi(\boldsymbol{\zeta}', \nu^{1/2} \boldsymbol{\zeta}', \mathbf{x}, \nu^{1/2} \sigma(\mathbf{x})) d\boldsymbol{\zeta}' + o(\nu^{-\infty})$$

where  $m' = m - 1$ . Furthermore,  $\psi(\boldsymbol{\zeta}', \mathbf{Z}', \mathbf{x}, \Sigma)$  is of class  $C^\infty$  on  $B' \times E^{m-1} \times \Delta \times [-\infty, \infty]$ . Hence, by a slight variation of Theorem 1,

$$(3.4) \quad F(\mathbf{x}, \nu) = \frac{1}{2} (2\pi)^{-m'/2} \int_{|\mathbf{Z}'| \geq 0} [\Psi_0(\mathbf{Z}', \mathbf{x}, \nu^{1/2} \sigma(\mathbf{x})) + \nu^{-1/2} \Psi_1(\mathbf{Z}', \mathbf{x}, \nu^{1/2} \sigma(\mathbf{x}))] d\mathbf{Z}' + O(\nu^{-1})$$

where

$$(3.5) \quad \Psi_0(\mathbf{Z}', \mathbf{x}, \Sigma) = \operatorname{erfc}(2^{-1/2} \Sigma) \exp\left(-\frac{1}{2} |\mathbf{Z}'|^2\right),$$

and

$$(3.6) \quad \Psi_1(\mathbf{Z}', \mathbf{x}, \Sigma) = -(2\pi)^{-1/2} Z_\alpha Z_\beta \sigma_{\alpha\beta}(\mathbf{x}) \exp\left[-\frac{1}{2} (\Sigma^2 + |\mathbf{Z}'|^2)\right].$$

Since

$$(3.7) \quad \int_{|\mathbf{Z}'| \geq 0} \exp\left(-\frac{1}{2} |\mathbf{Z}'|^2\right) d\mathbf{Z}' = \left(\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} Z^2\right) dZ\right)^{m'} = (2\pi)^{m'/2},$$

and

$$(3.8) \quad \int_{|\mathbf{Z}'| \geq 0} Z_\alpha Z_\beta \exp\left(-\frac{1}{2} |\mathbf{Z}'|^2\right) d\mathbf{Z}' = (2\pi)^{m'/2} \delta_{\alpha\beta},$$

we therefore obtain (2.23). The same procedure yields

$$(3.9) \quad F_i(\mathbf{x}, \nu) = (2\pi)^{-m/2} \int_{K(\mathbf{x}, \nu)} Z_i \exp\left(-\frac{1}{2} |\mathbf{Z}|^2\right) d\mathbf{Z} + O(\nu^{-1/2}),$$

and

$$(3.10) \quad F_{ijk}(\mathbf{x}, \nu) = (2\pi)^{-m/2} \int_{K(\mathbf{x}, \nu)} Z_i Z_j Z_k \exp\left(-\frac{1}{2} |\mathbf{Z}|^2\right) d\mathbf{Z} + O(\nu^{-1/2})$$

where  $K(\mathbf{x}, \nu) = \{\mathbf{Z} : Z_1 \leq \nu^{1/2} \sigma(\mathbf{x}), |\mathbf{Z}'| \geq 0\}$ .

To complete the evaluation of  $I(\mathbf{x}, \nu)$  we need to express  $\sigma(\mathbf{x})$  and the quantities occurring in (2.29) independently of the transformation matrix  $P(\mathbf{x})$ . Let  $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}(\mathbf{x})$  be the point on  $\partial D$  closest to  $\mathbf{x} \in \Delta$ , and let  $w(\boldsymbol{\xi}) = 0$  be the local representation for  $\partial D$ . Assume  $w(\boldsymbol{\xi}) < 0$  inside  $D$  and  $w(\boldsymbol{\xi}) > 0$  outside  $D$ . Then

$$(3.11) \quad n_i(\mathbf{x}) = [w_{\xi_k}(\hat{\boldsymbol{\xi}}) w_{\xi_k}(\hat{\boldsymbol{\xi}})]^{-1/2} w_{\xi_i}(\hat{\boldsymbol{\xi}})$$

is the  $i$ th component of the unit outer normal to  $\partial D$  at  $\hat{\boldsymbol{\xi}}$ . Similarly, the outer normal to  $\partial D(\mathbf{x})$  is given by the gradient (with respect to  $\boldsymbol{\zeta}$ ) of  $w(\mathbf{x} + P(\mathbf{x})\boldsymbol{\zeta})$ . Thus

$$(3.12) \quad \mu(\mathbf{x}) p_{jk}(\mathbf{x}) n_j(\mathbf{x}) = -\delta_{k1},$$

since the  $\zeta_1$  axis is perpendicular to  $\partial D(\mathbf{x})$ , and therefore

$$(3.13) \quad p_{i1}(\mathbf{x}) = p_{ik}(\mathbf{x}) \delta_{k1} = -\mu(\mathbf{x}) c_{ij}(\mathbf{x}) n_j(\mathbf{x}),$$

where  $C(\mathbf{x}) = A^{-1}(\mathbf{x})$  as noted in (2.4), and

$$(3.14) \quad \mu(\mathbf{x}) = [c_{ij}(\mathbf{x}) n_i(\mathbf{x}) n_j(\mathbf{x})]^{-1/2}.$$

Hence

$$(3.15) \quad \gamma_1(\mathbf{x}) = -|A(\mathbf{x})|^{-1/2} \mu(\mathbf{x})c_{ri}(\mathbf{x})n_i(\mathbf{x})g_{\xi_r}(\mathbf{x}, \mathbf{x}),$$

$$(3.16) \quad \eta_{1kk}(\mathbf{x}) = -\mu(\mathbf{x})c_{ri}(\mathbf{x})c_{st}(\mathbf{x})n_i(\mathbf{x})h_{\xi_r, \xi_s, \xi_t}(\mathbf{x}, \mathbf{x}),$$

and

$$(3.17) \quad \eta_{111}(\mathbf{x}) = -\mu^3(\mathbf{x})c_{ri}(\mathbf{x})c_{sj}(\mathbf{x})c_{tk}(\mathbf{x})n_i(\mathbf{x})n_j(\mathbf{x})n_k(\mathbf{x})h_{\xi_r, \xi_s, \xi_t}(\mathbf{x}, \mathbf{x}).$$

Also, since

$$(3.18) \quad \hat{\xi}_i = x_i + p_{ij}(\mathbf{x})\hat{\xi}_j,$$

where  $\hat{\xi}_j = \sigma(\mathbf{x})\delta_{j1}$ , it follows that

$$(3.19) \quad \sigma(\mathbf{x}) = \frac{1}{\mu(\mathbf{x})} [c_{ij}(\mathbf{x})c_{ik}(\mathbf{x})n_j(\mathbf{x})n_k(\mathbf{x})]^{-1/2} d(\mathbf{x}),$$

where  $d(\mathbf{x})$  is the (directed) distance from  $\mathbf{x}$  to  $\partial D$ . Finally, since  $\zeta_1 = f(\mathbf{x}, \zeta')$  is a solution of  $w(\mathbf{x} + P(\mathbf{x})\zeta) = 0$ , and  $f_{\xi_r}(\mathbf{x}, \mathbf{0}') = 0$ , it follows that

$$(3.20) \quad p_{j1}(\mathbf{x})w_{\xi_j}(\hat{\xi})\sigma_{\alpha\alpha}(\mathbf{x}) + p_{i\alpha}(\mathbf{x})p_{j\alpha}(\mathbf{x})w_{\xi_i, \xi_j}(\hat{\xi}) = 0.$$

Therefore

$$(3.21) \quad \sigma_{\alpha\alpha}(\mathbf{x}) = \mu(\mathbf{x})[c_{ij}(\mathbf{x}) - \mu^2(\mathbf{x})c_{ri}(\mathbf{x})c_{sj}(\mathbf{x})n_r(\mathbf{x})n_s(\mathbf{x})]\kappa_{ij}(\mathbf{x}),$$

where

$$(3.22) \quad \kappa_{ij}(\mathbf{x}) = [w_{\xi_k}(\hat{\xi})w_{\xi_k}(\hat{\xi})]^{-1/2} w_{\xi_i, \xi_j}(\hat{\xi}),$$

and we have used the fact that

$$(3.23) \quad c_{ij}(\mathbf{x}) = p_{i1}(\mathbf{x})p_{j1}(\mathbf{x}) + p_{i\alpha}(\mathbf{x})p_{j\alpha}(\mathbf{x}).$$

Now suppose  $\mathbf{x}$  is any point of  $E^m$  for which  $\hat{\xi}$  is uniquely determined. According to the analysis just completed, if  $\mathbf{x}$  is in a sufficiently small neighborhood of some point  $\mathbf{x}_0$  on  $\partial D$ , then (2.28) holds with  $\sigma(\mathbf{x})$  given by (3.19), and the terms in (2.29) given by (3.15)–(3.17) and (3.21). On the other hand, if  $\mathbf{x}$  is not in such a neighborhood, this same expression is asymptotically equivalent to either (1.2) or (1.4), depending on whether  $d(\mathbf{x}) < 0$  or  $d(\mathbf{x}) > 0$ . In other words, the expansion is valid in a larger region than we assumed for its derivation. Indeed we can summarize our results as follows.

**THEOREM 2.** *Let  $I(\mathbf{x}, \nu)$  be the function defined by (1.1) with conditions (i)–(iv), and let  $R$  be a closed subset of  $E^m$  such that*

(v) *for each  $x \in R$  there exists a unique point  $\hat{\xi} = \hat{\xi}(\mathbf{x}) \in R \cap \partial D$  which is closest to  $\mathbf{x}$ ;*

(vi)  *$R \cap \partial D = \{\xi \in R : w(\xi) = 0\}$ , where  $w(\xi)$  is of class  $C^\infty$  on a domain containing  $R \cap \partial D$ , and  $\text{grad } w(\xi)$  is an outer normal.*

*Then the asymptotic expansion (2.28), with  $\sigma(\mathbf{x})$  given by (3.19) and the terms in (2.29) given by (1.4), (3.15)–(3.17) and (3.21), holds uniformly as  $\nu \rightarrow \infty$  for all  $\mathbf{x} \in R$ .*

As an example consider the solution,

$$(3.24) \quad u(\mathbf{x}, t) = (4\pi t)^{-3/2} \int_D \phi(\xi) \exp \left[ -\frac{1}{4t} |\mathbf{x} - \xi|^2 \right] d\xi,$$

of the diffusion equation

$$(3.25) \quad u_t = u_{x_1x_1} + u_{x_2x_2} + u_{x_3x_3}, \quad t > 0, \quad |\mathbf{x}| \geq 0,$$

with the initial condition

$$(3.26) \quad u(\mathbf{x}, 0) = \begin{cases} \phi(\mathbf{x}), & \mathbf{x} \in D, \\ 0, & \mathbf{x} \notin D. \end{cases}$$

If we put  $\nu = (2t)^{-1}$ , then  $u(\mathbf{x}, t) = I(\mathbf{x}, \nu)$ , with

$$(3.27a,b) \quad h(\mathbf{x}, \xi) = \frac{1}{2} |\mathbf{x} - \xi|^2, \quad g(\mathbf{x}, \xi) = \phi(\xi).$$

Thus in this example we have  $A(\mathbf{x}) = I$ . Therefore  $C(\mathbf{x}) = I$  and  $\mu(\mathbf{x}) = 1$ . It follows from (1.4) that  $\gamma(\mathbf{x}) = \phi(\mathbf{x})$  and from (3.15),  $\gamma_1(\mathbf{x}) = -\phi_n(\mathbf{x})$  where  $\phi_n(\mathbf{x})$  is the derivative of  $\phi(\mathbf{x})$  in the direction of the outer normal to  $D$  at  $\hat{\xi}$ . Similarly, (3.19) reduces to  $\sigma(\mathbf{x}) = d(\mathbf{x})$  and from (3.21),  $\sigma_{\alpha\alpha}(\mathbf{x}) = 2H(\mathbf{x})$  where

$$(3.28) \quad H(\mathbf{x}) = \frac{1}{2} (\delta_{ij} - n_i n_j) \kappa_{ij}(\mathbf{x})$$

is the mean curvature of  $\partial D$  at  $\hat{\xi}$ . We also have  $\eta_{1kk}(\mathbf{x}) = \eta_{111}(\mathbf{x}) = 0$ , since  $h(\mathbf{x}, \xi)$  is quadratic. Hence, substituting into (2.28),

$$(3.29) \quad u(\mathbf{x}, t) = \frac{1}{2} \phi(\mathbf{x}) \operatorname{erfc} [(4t)^{-1/2} d(\mathbf{x})] - \left(\frac{t}{\pi}\right)^{1/2} [\phi_n(\mathbf{x}) + \phi(\mathbf{x})H(\mathbf{x})] \cdot \exp [-(4t)^{-1} d^2(\mathbf{x})] + O(t).$$

If  $D$  is a sphere of radius  $a$  centered at the origin and if  $\phi(\mathbf{x}) = f(r)$  where  $r = (x_i x_i)^{1/2}$ , then (3.29) reduces to

$$(3.30) \quad u(\mathbf{x}, t) = \frac{1}{2} f(r) \operatorname{erfc} \left[ \frac{r-a}{(4t)^{1/2}} \right] - \left(\frac{t}{\pi}\right)^{1/2} [f'(r) + a^{-1}f(r)] \exp \left[ \frac{-(r-a)^2}{4t} \right] + O(t),$$

which can be verified directly.

REFERENCES

[1] N. BLEISTEIN, *Uniform asymptotic expansions of integrals with stationary point near algebraic singularity*, *Comm. Pure Appl. Math.*, 19 (1966), pp. 353-370.  
 [2] N. BLEISTEIN AND R. A. HANDELSMAN, *Asymptotic Expansions of Integrals*. Holt, Rinehart, and Winston, New York, 1975.  
 [3] L. C. HSU, *A theorem on the asymptotic behavior of a multiple integral*, *Duke Math. J.*, 15 (1948), pp. 623-632.  
 [4] L. A. SKINNER, *Uniformly valid expansions for Laplace integrals*, this Journal, to appear.  
 [5] R. WONG, *On uniform asymptotic expansion of definite integrals*, *J. Approx. Theo.*, 7 (1971), pp. 76-86.

## THE ALGEBRA OF LINEAR PARTIAL DIFFERENCE OPERATORS AND ITS APPLICATIONS\*

DORON ZEILBERGER†

*Dedicated to Richard J. Duffin*

**Abstract.** The algebra of linear partial difference operators is investigated, and an elimination procedure demonstrated. Applications to combinatorics are given. In particular, a new proof and a  $q$ -analogue of MacMahon's Master Theorem are given.

**Introduction.** In this paper the algebra of partial difference operators will be considered, and applications to combinatorics demonstrated. It is surprising that partial difference equations have received so little attention while partial differential equations flourished. The only serious study of partial difference equations was in numerical analysis, and then only as a tool for solving partial differential equations numerically. Specific partial difference equations arose in random walk and combinatorics, but no unified theory such as in Hörmander [6] was attempted.

We hope to show here that partial difference operators are much more comfortable to work with, since the shift operator  $Xf(m) = f(m + 1)$  has a simple "Leibnitz rule"  $X^\alpha(fg) = (X^\alpha f)(X^\alpha g)$ . This is so much simpler than the continuous Leibnitz rule:

$$D^n(fg) = \sum \binom{n}{k} (D^k f)(D^{n-k} g).$$

A general theory of linear partial difference equations will not be attempted here. Instead we shall study the algebra of linear partial difference operators, and describe how to extend the elimination procedure in the algebra of polynomials (Van der Waerden [8]) to the algebra of partial difference operators. This will be followed by various applications of the elimination procedure. Unfortunately, in most cases, the algorithm is too cumbersome to be done by hand. However, since the algorithm involves nothing more complicated than multiplication by polynomials, it would be possible to employ a "symbolic" computer (such as MIT's MACSYMA) to solve some open problems in combinatorics.

This paper is dedicated to Richard J. Duffin whose pioneering work in partial difference equations prompted the author's interest in them. The author is also indebted to Richard A. Askey who challenged him to prove Andrews' [1] conjecture (§ 5). The present paper is a result of attempts to prove this conjecture, which is a  $q$ -generalization of the already resolved Dyson's conjecture. We attempted to  $q$ -generalize Good's [12] elegant proof of Dyson's conjecture.

Although our algorithm is capable of doing it, in principle, for any given  $n$ , it turns out to be too involved to do by hand. However, our algorithm turns out to be useful in other situations, as we hope to show later.

Next, let us describe briefly the content of the paper. In § 1 the algebra of linear ordinary difference operators is introduced and we show how to take an inverse of an operator. This is generalized to linear partial difference operators in § 2. Following this is a description of elimination in the algebra of linear partial difference operators with

---

\* Received by the editors January 26, 1979, and in final revised form March 24, 1980.

† Weizmann Institute of Science, Rehovot, Israel. This work was performed while the author was at the School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

constant coefficients. This algorithm is then applied to give a new proof of MacMahon’s [7] Master Theorem. Good’s proof of Dyson’s conjecture is quoted in subsection 3.4, and § 3 ends with the consideration of other possible generalizations to Dyson’s [8] conjecture B.

Section 4 gives a generalization of MacMahon’s Master Theorem which answers, in particular, a question posed by Andrews [1, p. 213].

Andrews’ [1] conjecture about the  $q$ -generalization of Dyson’s conjecture is considered in § 5. Although we failed to prove it, we describe how, in principle, for a given  $n$ , the algorithm is capable of settling it. Section 6 presents the algorithm of elimination in the most general setting, in the algebra of linear partial difference operators with variable coefficients.

We are very grateful to George Gasper, whose valuable criticism and comments transformed a disastrous first draft into a hopefully reasonable final version.

**1. The algebra of ordinary difference operators.** Let  $Z$  be the set of integers. We shall here consider the vector space of functions  $\mathcal{F}_1 = \{f : Z \rightarrow \mathbb{C}\}$ . A linear (ordinary) difference operator is a mapping  $P : \mathcal{F}_1 \rightarrow \mathcal{F}_1$  of the form

$$(1.1) \quad Pf(m) = \sum_{|\alpha| < M} c_\alpha(m) f(m + \alpha), \quad f \in \mathcal{F}_1,$$

where  $M$  is a positive integer, and  $c_\alpha$  are elements of  $\mathcal{F}_1$ . In case all the  $c_\alpha$ ’s are constants (polynomials), we have a linear difference operator with constant (polynomial) coefficients. Introducing the shift operator  $Xf(m) = f(m + 1)$ , we can write (1.1) in the form

$$(1.2) \quad Pf = \sum_{|\alpha| < M} c_\alpha X^\alpha f.$$

The set of all such operators will be denoted by  $\mathcal{P}_1$ . Note that the operator  $X$  has a particularly simple “Leibnitz rule”,

$$(1.3) \quad X^\alpha (fg) = (X^\alpha f)(X^\alpha g),$$

which proves that  $\mathcal{P}_1$  is an algebra:

$$(1.4) \quad \left( \sum c_\alpha X^\alpha \right) \left( \sum \delta_\beta X^\beta \right) = \sum_\alpha \sum_\beta c_\alpha (X^\alpha \delta_\beta) X^{\alpha+\beta}.$$

Let  $+\mathcal{P}_1, -\mathcal{P}_1$  be the subalgebras

$$+\mathcal{P}_1 = \left\{ \sum_{0 \leq \alpha \leq M} c_\alpha X^\alpha, \text{ for some } M \right\}$$

$$-\mathcal{P}_1 = \left\{ \sum_{-M \leq \alpha \leq 0} c_\alpha X^\alpha, \text{ for some } M \right\}.$$

$+\mathcal{P}_1$  (respectively  $-\mathcal{P}_1$ ) can be embedded in the algebras of linear difference operators of infinite order:

$$+\psi_1 = \left\{ \sum_{\alpha \geq 0} c_\alpha X^\alpha \right\}, \quad -\psi_1 = \left\{ \sum_{\alpha \leq 0} c_\alpha X^\alpha \right\}.$$

The domain of an operator in  $\pm\psi_1$  is  $\mathcal{F}_1^0$ , the space of functions of finite (=compact) support:

$$\mathcal{F}_1^0 = \{f : Z \rightarrow \mathbb{C}; f = 0 \text{ except at a finite number of points}\}.$$

An element of  ${}_{\pm}\mathcal{P}_1$  has an inverse in  ${}_{\pm}\psi_1$  (assume  $c_0 \neq 0$ ) iff  $c_0(m) \neq 0$  for all  $m$ ; then

$$(1.5) \quad \left( \sum_{0 \leq \alpha < M} c_{\alpha} X^{\alpha} \right)^{-1} = \left( 1 + \sum_{1 \leq \alpha < M} \frac{c_{\alpha}}{c_0} X^{\alpha} \right)^{-1} c_0^{-1} \\ = \sum_{k=0}^{\infty} (-1)^k \left( \sum_{1 \leq \alpha < M} \frac{c_{\alpha}}{c_0} X^{\alpha} \right)^k c_0^{-1}.$$

Each term on the right-hand side is evaluated according to (1.4), and since the lowest order term in  $(\sum (c_{\alpha}/c_0)X^{\alpha})^k$  is

$$\left( \frac{c_1}{c_0} X \right)^k = \left( \frac{c_1}{c_0} \right) \left( X \frac{c_1}{c_0} \right) \left( X^2 \frac{c_1}{c_0} \right) \cdots \left( X^k \frac{c_1}{c_0} \right) X^k,$$

we see that the sum in (1.5) is well defined, since the coefficients of any  $X^k$  are finite sums. This is a generalization of taking the reciprocal in the algebra of formal power series, the latter corresponding to the case where the  $c_{\alpha}$ 's are all constants.

The above formalism can be applied to solve a general linear difference equation

$$(1.6) \quad \sum_{\alpha=0}^M c_{\alpha}(m) f(m - \alpha) = 0, \quad m \geq M,$$

in terms of the initial values  $f(0), \dots, f(M - 1)$ . Of course there is a unique solution iff  $c_0(m) \neq 0$  for all  $m$ , and we can write (1.6) as

$$(1.6') \quad \left( \sum_{\alpha=0}^M c_{\alpha} X^{-\alpha} \right) f = 0 \quad \text{in } m \geq M.$$

Extending  $f$  by 0 in  $\{m < 0\}$ , we get

$$(1.6'') \quad \left( \sum_{\alpha=0}^M c_{\alpha} X^{-\alpha} \right) f = g \quad \text{in } Z,$$

where  $g$  is supported in  $\{0 \leq m \leq M - 1\}$  and each of its values is a linear combination of  $f(0), \dots, f(M - 1)$ .

From (1.6'') we get

$$f = \left( \sum_{\alpha=0}^M c_{\alpha} X^{-\alpha} \right)^{-1} g,$$

which is an *explicit* expression for  $f$ , in spite of its "formal" appearance.

**2. The algebra of partial difference operators.** The discussion in § 1 can easily be generalized to several discrete variables. For  $n$  a positive integer, consider the vector space of functions

$$\mathcal{F}_n = \{f: Z^n \rightarrow \mathbb{C}\}.$$

A linear partial difference operator is a mapping  $P: \mathcal{F}_n \rightarrow \mathcal{F}_n$  of the form

$$(2.1) \quad Pf(m) = \sum_{|\alpha| < M} c_{\alpha}(m) f(m + \alpha), \quad f \in \mathcal{F}_n.$$

Where multiindex notation is used;  $m, \alpha \in Z^n$ ,  $m = (m_1, \dots, m_n)$ ,  $\alpha = (\alpha_1, \dots, \alpha_n)$ ,  $|\alpha| = \sum_{i=1}^n |\alpha_i|$ , and  $M$  is a positive integer. The coefficients  $c_{\alpha}(m)$  are elements of  $\mathcal{F}_n$ . In case all the  $c_{\alpha}$ 's are constants (polynomials), we have a linear partial difference equation with constant (polynomial) coefficients. We introduce the shift operators  $X_i f(m_1, \dots, m_i, \dots, m_n) = f(m_1, \dots, m_i + 1, \dots, m_n)$ ,  $i = 1, \dots, n$ ; ( $X_i$  is the unit

shift in the  $m_i$  coordinate). We can write (2.1) in the form

$$(2.2) \quad P = \sum_{|\alpha| < M} c_\alpha X^\alpha,$$

because  $f(m + \alpha) = X_1^{\alpha_1} \cdots X_n^{\alpha_n} f(m) = X^\alpha f(m)$ . The set of all linear partial difference operators on  $Z^n$  will be denoted by  $\mathcal{P}_n$ . The operators  $(X_1, \dots, X_n)$  satisfy the simple ‘‘Leibnitz rule’’,

$$(2.3) \quad X^\alpha (fg) = (X^\alpha f)(X^\alpha g),$$

which shows that  $\mathcal{P}_n$  is an algebra:

$$(2.4) \quad \left( \sum c_\alpha X^\alpha \right) \left( \sum d_\beta X^\beta \right) = \sum_\alpha \sum_\beta c_\alpha (X^\alpha d_\beta) X^{\alpha+\beta}.$$

Let  $\Gamma_1, \dots, \Gamma_{2^n}$  be the  $2^n$  orthants in  $Z^n$ ; then if  $\Gamma$  is such an orthant,

$$\Gamma \mathcal{P}_n = \left\{ \sum_{\alpha \in \Gamma} c_\alpha X^\alpha; \text{ only finitely many } c_\alpha \text{'s are nonzero} \right\}$$

is a subalgebra of  $\mathcal{P}_n$ , since  $\Gamma$  is a cone.

$\Gamma \mathcal{P}_n$  can be extended to the algebras of linear partial difference operators of infinite order  $\Gamma \psi_n = \{ \sum_{\alpha \in \Gamma} c_\alpha X^\alpha \}$ .

The domain of an operator  $\Gamma \psi_n$  is  $\mathcal{F}_n^0$ , the space of functions of finite (=compact) support,

$$\mathcal{F}_n^0 = \{ f: Z^n \rightarrow \mathbb{C}; f = 0 \text{ except at a finite number of points} \}.$$

As a matter of fact, if  $P \in \Gamma \psi_n$  involves all  $X_1, \dots, X_n$  (i.e., it is not of lower dimension),  $P$  can be applied to functions whose support is a union of ‘‘hyperstrips’’, i.e., functions whose support is a subset of a set of the form  $\bigcup_{i=1}^n \{ -M_i \leq m_i \leq M_i \}$ , where the  $M_i$  are positive integers. The space of such functions will be denoted by  $\mathcal{F}_n^\hat{0}$ . Of course, it may happen that the domain is even larger.

An element  $P$  of  $\Gamma \mathcal{P}_n$  has an inverse in  $\Gamma \psi_n$  (assume  $c_0 \neq 0$ ) iff  $c_0(m) \neq 0$  for all  $m$ , and then

$$(2.5) \quad \left( \sum_{|\alpha| < M} c_\alpha X^\alpha \right)^{-1} = \left( 1 + \sum_{1 < |\alpha| < M} \frac{c_\alpha}{c_0} X^\alpha \right)^{-1} c_0^{-1} \\ = \sum_{k=0}^\infty (-1)^k \left( \sum_{1 \leq |\alpha| \leq M} \frac{c_\alpha}{c_0} X^\alpha \right)^k c_0^{-1}.$$

Each term on the right-hand side is evaluated according to (2.4); since the lowest order terms in  $(\sum_{1 \leq |\alpha| \leq M} (c_\alpha/c_0) X^\alpha)^k$  have order  $k$ , it is seen that the sum in (2.5) is well defined, since the coefficient of any  $X^\beta$  takes contributions only from the first  $|\beta|$  terms in (2.5), and thus consists of a finite sum. This generalizes the taking of the reciprocal in the algebra of formal power series of several variables, the latter corresponding to the case where all the  $c_\alpha$ ’s are constants.

The above formalism can be applied to solve a general linear partial difference equation in an orthant  $\Gamma$ ,

$$(2.6) \quad \sum_{\substack{|\alpha| < M \\ \alpha \in \Gamma}} c_\alpha(m) f(m - \alpha) = 0 \quad \text{in } \{m; m - \alpha \in \Gamma\},$$

in terms of the boundary values of  $f$ , that is, values of  $f$  near the axes. Of course (2.6) has

a unique solution iff  $c_0(m) \neq 0$  for all  $m$ , and we can write (2.6) in the form

$$(2.6') \quad \left( \sum_{|\alpha| < M} c_\alpha X^{-\alpha} \right) f = 0 \quad \text{in } \{m; m - \alpha \in \Gamma\}.$$

Let us extend  $f$  by zero outside  $\Gamma$ . Then (2.6') can be written as

$$(2.6'') \quad \left( \sum_{|\alpha| < M} c_\alpha X^{-\alpha} \right) f = g \quad \text{in } Z^n,$$

where  $g$  is supported in a neighborhood of the axes, i.e., is an element of the function space  $\mathcal{F}_n^0$ , discussed above, and each of the values of  $g$  is a linear combination of values of  $f$  near the axes.

From (2.6'') we get

$$(2.7) \quad f = \left( \sum_{|\alpha| < M} c_\alpha X^{-\alpha} \right)^{-1} g,$$

which is an *explicit* expression for  $f$  in  $\Gamma$ , in terms of its values near the axes. Suppose  $\underline{P}(z)$  is a polynomial in  $n$  variables,  $z_1, \dots, z_n$ ; then if

$$\underline{P}(z)^{-1} = \sum_{m \geq 0} c(m) z^m, \quad (z^m = z_1^{m_1} \cdots z_n^{m_n}),$$

and  $\delta$  is the discrete Dirac delta function:

$$\delta(0) = 1, \quad \delta(m) = 0, \quad m \neq 0,$$

then

$$c(m) = \left( \sum c(\alpha) X^{-\alpha} \right) \delta(m), \quad m \in Z^n_+.$$

So

$$(2.8) \quad c = \underline{P}(X^{-1})^{-1} \delta,$$

and  $c$  satisfies the partial difference equation  $\underline{P}(X^{-1})c = \delta$ , so  $c$  is a fundamental solution of the operator  $\underline{P}(X^{-1})$ .

**3. Elimination in the algebra of linear partial difference operators with constant coefficients.**

**3.1** The algebra of linear partial difference operators with constant coefficients in  $n$  variables,  $c\mathcal{P}_n$ , is isomorphic to the algebra of polynomials  $\mathbb{C}[z_1, \dots, z_n]$ , and the procedure in Van der Waerden [8, §§ 27, 77, 78] can be used to derive equations of lower dimensions from a system of equations.

*Example.* Solve the system

- (i)  $f(m + 1, n + 1) - 2f(m, n + 1) + f(m + 1, n) = 0,$
- (ii)  $f(m, n) + if(m + 1, n) - f(m + 1, n + 1) - if(m, n + 1) = 0, \quad m, n \geq 0.$

Setting  $Xf(m, n) = f(m + 1, n), Yf(m, n) = f(m, n + 1)$ , we can write the above equations in shorthand as

$$(3.1) \quad \begin{aligned} (XY - 2Y + X)f &= 0, \\ (I + iX - XY - iY)f &= 0. \end{aligned}$$



Let us eliminate  $Y$  from

$$P_1(X, Y) = XY - 2Y + X,$$

$$P_2(X, Y) = I + iX - XY - iY.$$

We would like to get an operator involving  $X$  only, so we write

$$P_1(X, Y) = Y(X - 2) + X,$$

$$P_2(X, Y) = -Y(X + i) + (I + iX),$$

and we get that

$$Q(X, Y) = (X + i)P_1(X, Y) + (X - 2)P_2(X, Y) = -2I + (1 - i)X + (1 + i)X^2,$$

(3.2a)  $(-2I + (1 - i)X + (1 + i)X^2)f = 0.$

Similarly, it also satisfies

(3.2b)  $((i + 2)Y^2 - (i + 1)Y - I)f = 0.$

(3.2a), (3.2b) immediately yield  $f$ , given  $f(0, 0), f(0, 1), f(1, 0), f(1, 1).$

In general, given  $n$  partial difference operators with constant coefficients, on  $Z^n$ , we can use the elimination procedure to obtain  $n$  ordinary difference operators  $Q_1(X_1), \dots, Q_n(X_n).$  If this is the case, we say that the ring  $(P_1, \dots, P_n)$  is “a complete intersection”. The elimination algorithm not only yields  $n$  ordinary difference operators  $Q_1(X_1), \dots, Q_n(X_n)$  in case of “complete intersection”, but also tells us when the algorithm “breaks down”, whenever it is not.

**3.2** The elimination procedure in Van der Waerden [8] can be applied in any polynomial ring  $R[x_1, \dots, x_n],$  where  $R$  is a commutative ring (there it is assumed that  $R$  is a field, but for our purpose it is enough that  $R$  is a commutative ring), in particular, if  $R$  is the ring of partial difference operators with constant coefficients. Let  $x_1, \dots, x_n$  be  $n$  indeterminates, and suppose we have  $n + 1$  operators with constant coefficients,

$$P_j(X_1, \dots, X_n; x_1, \dots, x_n), \quad j = 1, \dots, n + 1,$$

where the dependence on  $x_1, \dots, x_n$  is polynomial. In other words, the  $P_j$ 's are polynomials in  $X_1, \dots, X_n, x_1, \dots, x_n.$  In general it is possible to obtain, by elimination, an operator  $Q(X_1, \dots, X_n),$  independent of  $x_1, \dots, x_n,$  which is in the ring generated by  $P_1, \dots, P_{n+1};$  i.e., there exist  $Q_1(X, x), \dots, Q_{n+1}(X, x)$  such that  $Q_1P_1 + \dots + Q_{n+1}P_{n+1}$  is independent of  $x_1, \dots, x_n.$

**3.3 A new proof to MacMahon's Master Theorem.** MacMahon's Master Theorem (MacMahon [7], see also Andrews [1, p. 214]) asserts that the coefficient of  $x_1^{m_1} \dots x_n^{m_n}$  in

$$\prod_{i=1}^n \left( \sum_{j=1}^n a_{ij}x_j \right)^{m_i}$$

is equal to the coefficient of  $z_1^{m_1} \dots z_n^{m_n}$  in the power series expansion of  $[\det(\delta_{ij} - a_{ij}z_i)]^{-1}.$

Setting

$$F(m_1, \dots, m_n; x_1, \dots, x_n) = \prod_{i=1}^n \left( \sum_{j=1}^n a_{ij} \frac{x_j}{x_i} \right)^{m_i},$$

we are interested in  $F_0(m_1, \dots, m_n) = \text{const. term in } F(m_1, \dots, m_n; x_1, \dots, x_n).$  Now

$$X_i F = \left( \sum_{j=1}^n a_{ij} \frac{x_j}{x_i} \right) F, \quad i = 1, \dots, n,$$

and  $F$  satisfies the  $n$  partial difference equations

$$\sum_{j=1}^n x_j(-a_j + \delta_{ij}X_i)F = 0, \quad i = 1, \dots, n.$$

These are  $n$  linear homogeneous equations in  $x_1, \dots, x_n$  and (in this case Gaussian) elimination yields

$$(3.3) \quad P(X)F = 0, \quad m_1, \dots, m_n \geq 0,$$

where  $P(X) = \det(-a_{ij} + \delta_{ij}X_i)$ . Since  $P(X)$  is independent of  $x_1, \dots, x_n$ , and is a linear operator, we also have

$$(3.4) \quad P(X)F_0 = 0.$$

We now claim that

$$(3.5) \quad \begin{aligned} F_0 &= [X_1^{-1} \cdots X_n^{-1}P(X)]^{-1}\delta \\ &= X_1 \cdots X_n P(X)^{-1}\delta \\ &= [\det(\delta_{ij} - a_{ij}X_i^{-1})]^{-1}\delta. \end{aligned}$$

This follows from the fact that both sides are solutions of (3.3), and the boundary values match by the inductive hypothesis. By the remarks at the end of § 2 it follows that (3.5) implies the MacMahon Master Theorem. We prefer, however, to preserve the theorem in the form (3.5), because, as will be seen later, it yields a generalization.

**3.4 Good's proof to Dyson's conjecture.** In 1962, Dyson [2] made the following conjectures:

*Conjecture B.*

$$\text{The constant term in } \left( \prod_{\substack{i \neq j \\ 1 \leq i, j \leq n}} \left( 1 - \frac{x_i}{x_j} \right) \right)^a \text{ is } (na)!/(a!)^n$$

and its generalization

*Conjecture C.*

$$\text{The constant term in } \prod_{\substack{i \neq j \\ 1 \leq i, j \leq n}} \left( 1 - \frac{x_i}{x_j} \right)^{a_i} \text{ is } \frac{(a_1 + \dots + a_n)!}{a_1! \cdots a_n!}.$$

This was proved by Gunson [5], Wilson [9], and Good [4]. Good's proof is the proof that got us started in this business. Because of its importance, and also because of its elegance, we shall repeat it, in our notation. Set

$$F(a_1, \dots, a_n; x_1, \dots, x_n) = \prod_{\substack{i \neq j \\ 1 \leq i, j \leq n}} \left( 1 - \frac{x_i}{x_j} \right)^{a_i}.$$

$F$  satisfies the  $n$  partial difference equations,

$$X_i^{-1}F = \left[ \prod_{j \neq i} \left( 1 - \frac{x_i}{x_j} \right) \right]^{-1} F, \quad i = 1, \dots, n, \quad a_i > 0.$$

Elimination of  $x_1, \dots, x_n$  yields

$$(3.6) \quad (I - X_1^{-1} - \dots - X_n^{-1})F = 0 \quad \text{in } a_1, \dots, a_n > 0.$$

This equation is independent of  $x_1, \dots, x_n$ , and therefore (3.6) is also satisfied by  $F_0(a_1, \dots, a_n)$ , the constant term of  $F$ . So

$$(3.7) \quad (I - X_1^{-1} - \dots - X_n^{-1})F_0 = 0.$$

$G(a_1, \dots, a_n) = (a_1 + \dots + a_n)! / (a_1! \dots a_n!)$  is also a solution of (3.6) and the boundary values match, by the inductive hypothesis. So it follows that  $F_0(a_1, \dots, a_n) = (a_1 + \dots + a_n)! / (a_1! \dots a_n!)$ .

**3.5 There are many generalizations to Dyson’s conjecture B.** In Dyson’s conjecture C the factors  $(1 - (x_i/x_1)), \dots, (1 - (x_i/x_n))$ ,  $(i = 1, \dots, n)$  were grouped together, but we can take any  $n$  subsets of the  $n(n - 1)$  factors  $(1 - (x_i/x_j)) (i \neq j, 1 \leq i, j \leq n)$  and group them together, forming a function of  $a_1, \dots, a_n, x_1, \dots, x_n$ . Then we can use elimination to find a partial difference equation independent of  $x_1, \dots, x_n$ , satisfied by that function, and therefore also satisfied by the constant term (or any other coefficient, for that matter). Let us illustrate it by the following.

FACT.  $F_0(a, b, c) = \text{constant term in } F(x_1, x_2, x_3, a, b, c)$

$$= \left[ \left(1 - \frac{x_1}{x_2}\right) \left(1 - \frac{x_2}{x_1}\right) \right]^a \left[ \left(1 - \frac{x_1}{x_3}\right) \left(1 - \frac{x_3}{x_1}\right) \right]^b \left[ \left(1 - \frac{x_2}{x_3}\right) \left(1 - \frac{x_3}{x_2}\right) \right]^c,$$

is given by,

$$G(a, b, c) = (2a)!(2b)!(2c)!(a + b + c)! / [a!b!c!(a + b)!(a + c)!(b + c)!].$$

Proof.  $F$  satisfies the following linear partial difference equations,

$$X_1 F = \left(2 - \frac{x_1}{x_2} - \frac{x_2}{x_1}\right) F,$$

$$X_2 F = \left(2 - \frac{x_1}{x_3} - \frac{x_3}{x_1}\right) F,$$

$$X_3 F = \left(2 - \frac{x_2}{x_3} - \frac{x_3}{x_2}\right) F.$$

Eliminating  $x_1, x_2, x_3$  we get that  $F$  satisfies the partial difference equation

$$(3.8) \quad (X_1^2 + X_2^2 + X_3^2 - 2X_1X_2 - 2X_1X_3 - 2X_2X_3 + X_1X_2X_3)F = 0.$$

Since this is a linear equation and is independent of  $x_1, x_2, x_3$ , it is also satisfied by  $F_0(a, b, c)$ . It is straightforward (albeit rather long) to check that this equation is also satisfied by  $G(a, b, c)$ . It is trivial to check that  $F_0 = G$  on the boundary of  $Z_+^3, \{a = 0\} \cup \{b = 0\} \cup \{c = 0\}$ .

The special case  $a = b = c$  yields conjecture B for  $n = 3$ . The above proof was given for pedagogical reasons, because the Fact is equivalent to

$$\sum (-1)^k \binom{2a}{a+k} \binom{2b}{b+k} \binom{2c}{c+k} = \frac{(2a)!(2b)!(2c)!(a + b + c)!}{a!b!c!(a + b)!(a + c)!(b + c)!};$$

this is equivalent to the terminating form of Dixon’s theorem, which in turn is equivalent to Dyson’s conjecture C for  $n = 3$  (Andrews [1, p. 214]).

**4. A generalization of MacMahon’s Master Theorem.**

4.1 In § 6 we shall describe how to generalize the elimination procedure to systems of linear partial difference operators with variable coefficients. In this case the task is much harder, since the ring of linear partial difference operators is not commutative.

However, in some cases of variable coefficients operators the elimination method generalizes right away. This happens when the operators to be eliminated are pairwise commutative. In particular, let us prove the following generalization of MacMahon’s Master Theorem.

**THEOREM.** *Let  $(f_{ij}(m_i)), 0 \leq i, j \leq n$ , be a matrix of discrete functions (where the  $i$ th row has functions depending only on  $m_i$ ), and for a discrete function  $G$ , let*

$$G^{(m)} = G(0)G(1) \cdots G(m-1).$$

*Then  $F_0(m_1, \dots, m_n)$  = the constant term of*

$$F(m_1, \dots, m_n; x_1, \dots, x_n) = \prod_{i=1}^n \left( \sum_{j=1}^n f_{ij} \frac{x_j}{x_i} \right)^{(m_i)},$$

*is given by*

$$(4.1) \quad G(m_1, \dots, m_n) = [\det (\delta_{ij} - f_{ij}X_i^{-1})]^{-1} \delta,$$

*where  $\delta$  is the discrete Dirac delta function.  $\det (\delta_{ij} - f_{ij}X_i^{-1})$  is a linear partial difference operator with variable coefficients, and its inverse is calculated by (2.5).*

*Proof.* The proof is along the same lines as the one given in subsection 3.3.  $F = F(m_1, \dots, m_n; x_1, \dots, x_n)$  satisfies

$$X_i F = \left( \sum_{j=1}^n f_{ij}(m_i) \frac{x_j}{x_i} \right) F, \quad i = 1, \dots, n;$$

*i.e.,*

$$\sum_{j=1}^n x_j [\delta_{ij} X_i - f_{ij}(m_i)] F = 0, \quad i = 1, \dots, n.$$

Gaussian elimination still works here because any two entries in the matrix  $(\delta_{ij} X_i - f_{ij}(m_i))$  which are in different rows, commute, since  $X_i$  acts only on functions independent of  $m_i$ . Consequently  $F(m_i, \dots, m_n; x_i, \dots, x_n)$  satisfies

$$(4.2) \quad \det (\delta_{ij} X_i - f_{ij}(m_i)) F = 0.$$

This partial difference equation is independent of  $x_1, \dots, x_n$ , and since it is linear it follows that it is also satisfied by  $F_0(m_1, \dots, m_n)$ , the constant term of  $F$ .

So

$$(4.3) \quad \det (\delta_{ij} X_i - f_{ij}(m_i)) F_0 \equiv 0.$$

But we also have

$$(4.4) \quad \det (\delta_{ij} X_i - f_{ij}(m_i)) G \equiv 0.$$

$F_0 = G$  on the boundary of  $Z_+^n$ , that is on  $\cup_{i=1}^n \{m_i = 0\}$ , by the inductive hypothesis, and thus the theorem follows:  $F_0 \equiv G$  throughout  $Z_+^n$ .

**4.2 A  $q$ -analogue of MacMahon’s Master Theorem.** The above theorem answers, in particular, a question raised by Andrews [1, p. 213] about a  $q$ -analogue to MacMahon’s Master Theorem. The  $q$ -analogue of  $(a + b)^m$  is  $(a + b)(a + qb)(a + q^2b) \cdots (a + q^{m-1}b)$ ; and naturally the  $q$ -analogue of  $(a + b + c)^m$  would be  $(a + b + c)(a + qb + q^2c)(a + q^2b + q^4c) \cdots (a + q^{m-1}b + q^{2m-2}c)$ , and in general, a  $q$ -analogue of  $(a_1 + \cdots + a_n)^m$  would be

$$(a_1 + \cdots + a_n)(a_1 + qa_2 + \cdots + q^{n-1}a_n) \\ \times (a_1 + q^2a_2 + \cdots + q^{2(n-1)}a_n) \cdots (a_1 + q^{m-1}a_2 + \cdots + q^{(m-1)(n-1)}a_n).$$

It is seen that a  $q$ -analogue of the Master Theorem is obtained by putting  $f_{ij}(m_i) = a_{ij}q^{m_i(j-1)}$ , where the  $a_{ij}$ 's are constants.

**5. About the possibility of proving a  $q$ -analogue of Dyson's conjecture.** In [1, p. 216], Andrews conjectured that the constant term of

$$F = \prod_{\substack{i \neq j \\ 1 \leq i, j \leq n}} \left( \frac{\varepsilon_{ij} x_i}{x_j} \right)_{a_i}, \quad \varepsilon_{ij} = \begin{cases} 1, & i < j, \\ q, & j > i, \end{cases}$$

is  $q_{a_1+\dots+a_n}/(q_{a_1} \cdots q_{a_n})$ , where  $(x)_a = (1-x)(1-qx) \cdots (1-q^{a-1}x)$ . This is the  $q$ -analogue of Dyson's conjecture C. Let us try to generalize Good's proof. We have the  $n$  equations

$$\begin{aligned} X_1 F &= \left(1 - q^{a_1} \frac{x_1}{x_2}\right) \cdots \left(1 - q^{a_1} \frac{x_1}{x_n}\right) F, \\ X_2 F &= \left(1 - q^{a_2+1} \frac{x_2}{x_1}\right) \left(1 - q^{a_2} \frac{x_2}{x_3}\right) \cdots \left(1 - q^{a_n} \frac{x_2}{x_n}\right) F, \\ &\vdots \\ X_n F &= \left(1 - q^{a_n+1} \frac{x_n}{x_1}\right) \cdots \left(1 - q^{a_n+1} \frac{x_n}{x_{n-1}}\right) F. \end{aligned}$$

By finding the expressions for  $X_1^{\alpha_1} \cdots X_n^{\alpha_n}$  for high enough  $\alpha_1, \dots, \alpha_n$ , it is possible in principle (for a fixed  $n$ ), to eliminate  $x_1, \dots, x_n$  from these equations, and get a linear partial difference operator  $P(a_1, \dots, a_n; X_1, \dots, X_n)$  such that  $PF \equiv 0$ . Then it would be possible to check that  $P(q_{a_1+\dots+a_n}/(q_{a_1} \cdots q_{a_n})) = 0$ , and equate boundary values. Details will appear elsewhere.

However, this process is very complicated to do by hand (a symbolic computer will help here), and we were unable to find such an equation even for  $n = 3$ .

**6. Elimination in the algebra of linear partial difference operators.**

**6.1. Gaussian elimination in the ring of linear ordinary difference operators.** Since the process of elimination in Van der Waerden [8] is based on Gaussian elimination in a commutative ring, we shall first describe how to modify Gaussian elimination to the noncommutative ring of partial difference operators. To begin with, let us consider the ring of linear ordinary difference operators.

Suppose we have the two operators

(6.1a)  $Q_1 = (aX + b)\lambda_1 + P_2(m, X)\lambda_2 + \cdots + P_n(m, X)\lambda_n,$

(6.1b)  $Q_2 = (a'X + b')\lambda_1 + P'_2(m, X)\lambda_2 + \cdots + P'_n(m, X)\lambda_n,$

where  $P_2, \dots, P_n, P'_2, \dots, P'_n$  are any linear difference operators and the coefficients of  $\lambda_1$  are first order. If  $a, b, a', b'$  were constants we could have just multiplied the first equation by  $(a'X + b')$ , the second by  $(aX + b)$  and subtracted, thus getting rid of  $\lambda_1$ . But since  $a, a', b, b'$  are functions,  $(aX + b)$  and  $(a'X + b')$  do not commute in general.

We first form

(6.2a)  $b'Q_1 - bQ_2 = \phi(m)X\lambda_1 + \cdots,$

(6.2b)  $a'Q_1 - aQ_2 = -\phi(m)\lambda_1 + \cdots.$

Now we apply  $X$  to (6.2b) and get

(6.2b')  $X(a'Q_1 - aQ_2) = -[X\phi(m)]X\lambda_1 + \cdots.$

Since  $\phi$  and  $X\phi$  are functions, they commute, and  $(X\phi)(6.2a) + \phi(6.2b')$  yields an operator independent of  $\lambda_1$ .

The above process can be described as reducing two equations in which the coefficients of  $\lambda_1$  are first order difference operators, to two equations in  $(X\lambda_1, \lambda_2, \dots, \lambda_n)$  in which the coefficients of  $X\lambda_1$  are zero order difference operators, i.e. functions.

The same method can be applied to the case in which the coefficients of  $\lambda_1$  are of any given order  $k$ .

$$(6.3a) \quad Q_1 = (a_0X^k + a_1X^{k-1} + \dots + a_k)\lambda_1 + \dots,$$

$$(6.3b) \quad Q_2 = (a'_0X^k + a'_1X^{k-1} + \dots + a'_k)\lambda_1 + \dots.$$

Now

$$(6.4a) \quad a'_0Q_1 - a_0Q_2 = (b_1X^{k-1} + \dots + b_k)\lambda_1 + \dots,$$

$$(6.4b) \quad a'_kQ_1 - a_kQ_2 = (c_0X^{k-1} + \dots + c_{k-1})X\lambda_1 + \dots.$$

Now applying  $X$  to (6.4a) we get two operators, linear in  $(X\lambda_1), \lambda_2, \dots, \lambda_n$ , in which the coefficients of  $(X\lambda_1)$  are operators of order  $k-1$ . Continuing in this manner, we get rid of  $\lambda_1$  altogether.

If we have  $n$  operators,  $Q_1, \dots, Q_n$  linear in  $\lambda_1, \dots, \lambda_n$ , we can use the above procedure to get  $n-1$  operators linear in  $\lambda_2, \dots, \lambda_n$ . Continuing in this manner, we end up with one operator linear in  $\lambda_n$ , and since it is possible to divide by  $\lambda_n$ , we end up with an operator  $Q$  independent of  $\lambda_1, \dots, \lambda_n$ . This operator is in the ideal generated by  $\{Q_1, \dots, Q_n\}$ , and so  $Q_1f = 0, \dots, Q_nf = 0 \Rightarrow Qf = 0$ .

**6.2 Gaussian elimination in the ring of linear partial difference operators.** Let us consider a special case first,

$$Q_1 = [(aY + b)X + (cY + d)]\lambda_1 + P_2(m, X, Y)\lambda_2 + \dots + P_n(m, X, Y)\lambda_n,$$

$$Q_2 = [(a'Y + b')X + (c'Y + d')]\lambda_1 + P'_2(m, X, Y)\lambda_2 + \dots + P_n(m, X, Y)\lambda_n.$$

Let us write it as follows,

$$Q_1 = (aY + b)(X\lambda_1) + (cY + d)\lambda_1 + \dots,$$

$$Q_2 = (a'Y + b')(X\lambda_1) + (c'Y + d')\lambda_1 + \dots.$$

Using the process of subsection 6.1 we get rid of  $X\lambda_1$ , to get an operator

$$Q'_1 = P_1(Y)\lambda_1 + \dots.$$

Once again using the above process, this time to get rid of  $\lambda_1$ , yields

$$Q'_2 = P_2(Y)(X\lambda_1) + \dots.$$

$XQ'_1$  and  $Q'_2$  are two operators, linear in  $(X\lambda_1), \lambda_2, \dots, \lambda_n$ , but with the advantage that the coefficients of  $(X\lambda_1)$  are just ordinary difference operators. In this form it is possible to use the method in the previous subsection, to get rid of  $(X\lambda_1)$  altogether.

In general, suppose that we know how to do Gaussian elimination for partial difference operators of dimension  $K-1$ ; let us describe how to perform Gaussian elimination for partial difference operators of dimension  $K$ . Consider the two operators

$$(6.5a) \quad A = P_1(X_1, \dots, X_K)\lambda_1 + P_2(X_1, \dots, X_K)\lambda_2 + \dots + P_n(X_1, \dots, X_K)\lambda_n,$$

$$(6.5b) \quad B = P'_1(X_1, \dots, X_K)\lambda_1 + P'_2(X_1, \dots, X_K)\lambda_2 + \dots + P'_n(X_1, \dots, X_K)\lambda_n.$$

Let us write

$$P_1(X_1, \dots, X_K) = \sum_{i=0}^L Q_i(X_2, \dots, X_K) X_1^i,$$

$$P'_1(X_1, \dots, X_K) = \sum_{i=0}^L Q'_i(X_2, \dots, X_K) X_1^i.$$

Substituting in (6.5) we get that  $A, B$  are linearly dependent on  $\lambda_1, (X_1\lambda_1), \dots, (X_1^L\lambda_1), \lambda_2, \dots, \lambda_n$ . Using the algorithm for dimension  $K - 1$ , we can get rid of  $\lambda_1$ , yielding

$$C = C_1(X_1\lambda_1) + \dots + C_L(X_1^L\lambda_1),$$

and using it another time, we can get rid of  $(X_1^L\lambda_1)$ :

$$D = D_1\lambda_1 + \dots + D_{L-1}(X_1^{L-1}\lambda_1).$$

$C, X_1D$  are two linear operators in  $(X_1\lambda_1), \dots, (X_1^L\lambda_1)$ . Continuing the process, we can dispose of  $(X_1\lambda_1), (X_1^2\lambda_1), \dots, (X_1^L\lambda_1)$  successively, and finally get rid of  $\lambda_1$  altogether. This algorithm eventually yields an operator which is independent of  $\lambda_1, \dots, \lambda_n$ .

**6.3 General elimination.** Let us recall that if  $P_1, P_2 \in R[x_1, \dots, x_n]$ , where  $R$  is a commutative ring, we get an element  $Q \in \{P_1, P_2\}$ , which is in  $R[x_2, \dots, x_n]$ , by expressing  $P_1, P_2$  as polynomials in  $x_1$ , with coefficients in  $R[x_2, \dots, x_n]$ , and by forming  $x_1^l P_1, x_1^l P_2$  for sufficiently large  $l$ , thus getting linear equations in the powers of  $x_1$ , with coefficients in  $R[x_2, \dots, x_n]$ ; then we use Gaussian elimination. The same method can be used in  $\mathcal{P}_n[x_1, \dots, x_n]$ , where  $\mathcal{P}_n$  is the ring of partial difference operators on  $Z^n$ . This is so because we know how to perform Gaussian elimination in that ring.

In general, if we have  $N + 1$  operators  $P_1(x_1, \dots, x_N), \dots, P_{N+1}(x_1, \dots, x_N)$ , where the dependence on  $x_1, \dots, x_N$  is polynomial, it is possible to get an operator which belongs to the ideal  $\{P_1, \dots, P_{N+1}\}$ , and which is independent of  $x_1, \dots, x_N$ . The present algorithm is a generalization to the ring of linear partial difference operators, of the process described in subsection 3.2 for the ring of partial difference operators with constant coefficients.

**6.4 Overdetermined systems of linear partial difference operators.** In subsection 3.1 we saw that two linear partial difference operators with constant coefficients usually give rise to an operator of lower dimension. The same is true for general linear partial difference operators. Let  $P, Q$  be two such operators on  $Z^n$ , and write them as follows,

$$P = \sum P_i(X_2, \dots, X_n) X_1^i,$$

$$Q = \sum Q_i(X_2, \dots, X_n) X_1^i.$$

By considering  $P, X_1P, \dots, X_1^L P, Q, X_1Q, \dots, X_1^M Q$  for sufficiently large  $L, M$  we get linear dependence on the powers of  $X_1$  and using the process of Gaussian elimination described in subsections 6.1, 6.2, we obtain an operator involving only  $X_2, \dots, X_n$ , which is in the ideal  $\{P, Q\}$ . In general if we have  $n$  operators  $P_1(X_1, \dots, X_n), \dots, P_n(X_1, \dots, X_n)$  we should get  $n$  ‘‘ordinary’’ operators  $Q_1(X_1), \dots, Q_n(X_n)$ . If this is the case, the ideal  $\{P_1, \dots, P_n\}$  is called ‘‘complete intersection’’. If this is not the case, then the algorithm will tell us so by breaking down.

*Example.* Find 2 ordinary difference operators satisfied by every solution of the system

$$\begin{aligned} \text{(i)} \quad & mf(m+1, n+1) + nf(m+1, n) + 2mf(m, n+1) - mnf(m, n) = 0, \\ \text{(ii)} \quad & f(m+1, n+1) + (n+3)f(m+1, n) + mf(m, n+1) - 3mf(m, n) = 0, \\ & m, n \geq 0. \end{aligned}$$

In our notation

$$\begin{aligned} P_1(X, Y)f &= (mXY + nX + 2mY - mnI)f = 0, \\ P_2(X, Y)f &= (XY + (n+3)X + mY - 3mI)f = 0. \end{aligned}$$

We have to eliminate  $P_1, P_2$ . Writing

$$\begin{aligned} P_1 &= mXY + 2mY + (nX - mnI), \\ P_2 &= XY + mY + (n+3)X - 3mI, \end{aligned}$$

yields

$$\begin{aligned} Q_1 &= P_1 - mP_2 = (2m - m^2)Y + (nX - mnI - m(n+3)X + 3m^2I), \\ Q_2 &= P_1 - 2P_2 = (m-2)XY + (-(n+6)X + 6m - mn). \end{aligned}$$

Simplifying, we get

$$\begin{aligned} Q_1 &= m(2-m)Y + (n - mn - 3m)X + m(3m - n)I, \\ Q_2 &= (m-2)XY + (-(n+6)X + (6-n)mI). \end{aligned}$$

So,

$$\begin{aligned} XQ_1 &= (1-m^2)XY - (mn+3m+3)X^2 + (m+1)(3m-n+3)X, \\ Q_2 &= (m-2)XY + (-(n+6)X + (6-n)m). \end{aligned}$$

$(m-2)XQ_1 - (1-m^2)Q_2$  will be an ordinary difference operator in the  $m$ -direction, i.e., an operator in  $X$ . Similarly, it is possible to find an ordinary difference operator in the  $n$ -direction.

*Remark.* In Even and Gillis [3] the authors mention that they were unable to find a combinatorial proof of the ordinary difference equation (34) there, which is satisfied by  $P_{rst}$ . However they exhibit 3 partial difference equations satisfied by  $P_{rst}$ , namely (9), (32), and one obtained from (32) by replacing  $r$  by  $s$ . Thus our (completely elementary) algorithm should yield their ordinary difference equation (34).

#### REFERENCES

- [1] GEORGE E. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in Theory and Applications of Special Functions, Richard A. Askey, ed., Academic Press, New York, 1975, pp. 191–224.
- [2] F. J. DYSON, *Statistical theory of the energy levels of complex systems I*, J. Math. Phys., 3 (1962), pp. 140–156.
- [3] S. EVEN AND J. GILLIS, *Derangements and Laguerre polynomials*, Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 135–143.
- [4] I. J. GOOD, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.
- [5] J. GUNSON, *Proof of a conjecture by Dyson in the statistical theory of energy levels*, J. Math. Phys., 3 (1962), pp. 752–753.
- [6] LARS HÖRMANDER, *Linear Partial Differential Operators*, Springer, Berlin, 1963.



- [7] P. A. MACMAHON, *Combinatory Analysis*, Vol. 1, Cambridge University Press, Cambridge, 1915, (reprinted by Chelsea, New York, 1960).
- [8] B. L. VAN DER WAERDEN, *Modern Algebra*, Vols. 1, 2, F. Blum, transl., Frederick Ungar Publishing Co., New York, 1953.
- [9] K. WILSON, *Proof of a conjecture by Dyson*, J. Math. Phys., 3 (1962), pp. 1040–1043.

## UNIQUENESS OF WALSH SERIES WHICH SATISFY AN AVERAGED GROWTH CONDITION\*

WILLIAM R. WADE†

**Abstract.** Let  $S$  be a Walsh series and denote its  $2^N$ th partial sums by  $S_{2^N}$ , for  $N = 0, 1, \dots$ . We show that if  $f$  is a finite-valued function belonging to  $L^p$ ,  $p > 1$ , if  $S_{2^N}$  converges to  $f$  as  $N \rightarrow \infty$  (for all but countably many points in the interval  $[0, 1]$ ), and if

$$\int_0^1 \left( \sum_{k=1}^{\infty} [S_{2^k} - S_{2^{k-1}}]^2 \right)^{1/2} < \infty,$$

then  $S$  is the Walsh-Fourier series of  $f$ .

**1. Introduction.** Let  $\psi_0, \psi_1, \dots$  denote the Walsh-Paley functions (see [3]). Thus, for each nonnegative integer  $k$  and each point  $x$  belonging to the unit interval  $[0, 1]$ , we have that

$$(1) \quad \psi_k(x) = \prod_{j=0}^{\infty} \exp(\pi i x_{j+1} k_j),$$

where the numbers  $x_j$  and  $k_j$  are either 0 or 1 and come from the binary expansions of  $x$  and  $k$ :

$$x = \sum_{j=1}^{\infty} x_j 2^{-j}, \quad k = \sum_{j=0}^{\infty} k_j 2^j.$$

When  $x \in [0, 1)$  is a dyadic rational, the finite binary expansion is used.

Let  $N \geq 0$  be an integer. We shall denote the  $2^N$ th partial sum of a Walsh series  $S = \sum_{k=0}^{\infty} a_k \psi_k$  by

$$(2) \quad S_{2^N}(x) = \sum_{k=0}^{2^N-1} a_k \psi_k(x), \quad x \in [0, 1].$$

In case the series  $S$  is the Walsh-Fourier series of some integrable function  $f$ , i.e.,

$$a_k = \int_0^1 f(t) \psi_k(t) dt, \quad k = 0, 1, \dots,$$

we shall denote the  $2^N$ th partial sums of  $S$  by  $S_{2^N}(f, x)$ .

In certain types of applications, especially those in communications, the signals received or the data recorded are discrete values of the  $2^N$ th partial sums of the Walsh-Fourier series of a certain unknown function  $f$  whose values are being sought (see, e.g., [1] and [5]). The integer  $2^N$  often corresponds to the number of readings or samples taken and, presumably, as  $N$  gets larger the prediction for the values of  $f$  gets more precise. In fact (see [3]), if  $f$  is continuous, then  $S_{2^N}(f)$  converges uniformly to  $f$ , as  $N \rightarrow \infty$ .

However, noise and other interference can alter the signal so much that the source  $f$  is impossible to recover. This situation can prevail despite the fact that the signals are converging to a continuous function. For example, the Walsh series

$$(3) \quad T \equiv \sum_{k=0}^{\infty} \psi_k(x)$$

---

\* Received by the editors September 14, 1979, and in revised form January 26, 1980. This research was supported in part by the National Science Foundation under Grant MCS 78-00902.

† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37916.

has  $2^N$ th partial sums which converge to zero *everywhere* on the interval  $(0, 1)$ , but  $T$  is certainly not the Walsh–Fourier series of zero. In particular, an additional criterion is needed to conclude that a Walsh series, converging to an integrable function  $f$ , is the Walsh–Fourier series of that function.

Fine [3] was the first to address this problem. He showed that if  $f$  is a finite-valued, integrable function, if  $S$  is a Walsh series whose coefficients,  $a_k$ , tend to zero as  $k \rightarrow \infty$ , and if for every  $x \in [0, 1]$ ,  $S_{2^N}(x) \rightarrow f(x)$  as  $N \rightarrow \infty$ , then  $S$  is the Walsh–Fourier series of  $f$ .

Crittenden and Shapiro [2] obtained a far stronger result containing a theorem which generalizes Fine’s result in two directions. First, convergence of the partial sums,  $S_{2^N}$ , can be relaxed at countably many points. And, secondly, the growth condition “ $\lim_{k \rightarrow \infty} a_k = 0$ ” can be replaced by

$$(4) \quad \lim_{N \rightarrow \infty} 2^{-N} S_{2^N}(x) = 0, \quad \text{for all } x \in [0, 1].$$

That this growth condition must hold for every point  $x \in (0, 1]$  can be verified by observing that the series (3) satisfies (4) at all points except  $x = 0$ .

Neither of these growth conditions is particularly suited for applications. On the one hand, condition (4) uses the data as collected, but it can only be checked at the sample points, i.e., at finitely many points  $x \in [0, 1]$ . On the other hand, in order to obtain the coefficients  $a_k$ , one must use the formula

$$a_k = \int_0^1 \psi_k(t) S_{2^N}(t) dt, \quad k < 2^N,$$

and then check to see whether  $a_k$  is converging to zero as  $k \rightarrow \infty$ . Notice that since  $S_{2^N}$  is a step function, the integral above can be computed exactly.

A growth condition using unprocessed data  $S_{2^N}$ , but involving average values rather than pointwise conditions, would be more useful. We propose the following condition. A Walsh series  $S$  is said to satisfy *condition H* if

$$(5) \quad \int_0^1 \sqrt{\sum_{k=0}^{\infty} [S_{2^{k+1}}(x) - S_{2^k}(x)]^2} dx < \infty.$$

Notice that when truncated to  $0 \leq k \leq N$ , this integral also involves step functions, and thus can be computed exactly. Also, it requires one integral per layer  $S_{2^{k+1}} - S_{2^k}$  rather than  $2^k$  integrals.

Recall [4] that the Walsh–Fourier series of any function belonging to dyadic  $H^1$  necessarily satisfies condition  $H$ . Since dyadic  $H^1$  contains the spaces  $L^p[0, 1]$  for  $p > 1$ , condition (5) is not too narrow to encompass most applications.

The following result will be proved in § 3.

**THEOREM.** *Let  $S = \sum_{k=0}^{\infty} a_k \psi_k$  be a Walsh series which satisfies condition  $H$ , and suppose that  $f$  is a finite-valued function belonging to dyadic  $H^1$ . If  $\lim_{N \rightarrow \infty} S_{2^N}(x) = f(x)$  for all but countably many points  $x \in [0, 1]$ , then  $S$  is the Walsh–Fourier series of  $f$ .*

It should be pointed out that the same proof can be used to establish the Haar series analogue of this theorem.

**2. The fundamental lemma.** Suppose that

$$(6) \quad I_0 = [(p_0 - 1)/2^{k_0}, p_0/2^{k_0})$$

is a half-open interval, where  $p_0$  and  $k_0$  are nonnegative integers. A sequence of pairs of intervals  $\{(I_j, \tilde{I}_j)\}_{j=0}^{\infty}$  is said to be *properly nested* in the interval  $I_0$  if for each integer  $j \geq 1$

the following four properties are satisfied:

- (7) the intervals  $I_j$  and  $\tilde{I}_j$  are half-open on the right and have dyadic rational endpoints;
- (8) the length of both  $I_j$  and  $\tilde{I}_j$  is  $2^{-k_0-j}$ ;
- (9)  $I_j \cup \tilde{I}_j = I_{j-1}$  and  $I_j \cap \tilde{I}_j = \emptyset$ ;
- (10)  $I_j \subset I_{j-1}$ .

Observe that conditions (7) through (10) guarantee two useful properties. Let  $j \geq 1$  be an integer. If  $k$  is any integer which satisfies  $2^{k_0+j-1} \leq k < 2^{k_0+j}$ , then the Walsh function  $\psi_k$  is constant on both intervals  $I_j$  and  $\tilde{I}_j$ , and changes signs from one interval to the other:

$$(11) \quad \psi_k(x) = -\psi_k(y) \quad \text{for } x \in I_j, y \in \tilde{I}_j.$$

The following lemma reveals the true nature of condition  $H$ . It prevents the partial sums from building up before vanishing locally, thereby ruling out series such as example (3).

LEMMA. *Suppose that  $S$  is a Walsh series which satisfies condition  $H$  and that  $I_0$  is an interval of the form (6). Suppose further, that  $S_{2^{k_0}}$  is never zero on  $I_0$  and that the sequence  $\{(I_j, \tilde{I}_j)\}_{j=0}^\infty$  is properly nested in  $I_0$ . Then there exists an integer  $k_1 > k_0$  such that  $S_{2^{k_1}}$  is never zero on  $\tilde{I}_{k_1-k_0}$ .*

To prove this lemma we begin by supposing to the contrary that there is a properly nested sequence  $\{(I_j, \tilde{I}_j)\}_{j=0}^\infty$  in  $I_0$  such that  $S_{2^{k_0+j}}$  has a zero in  $\tilde{I}_j$  for each integer  $j > 0$ . Since  $S_{2^{k_0+j}}$  is constant on  $I_j$  and  $\tilde{I}_j$  for  $j > 0$ , we are assuming that

$$(12) \quad S_{2^{k_0+j}}(x) \equiv 0, \quad x \in \tilde{I}_j, \quad j > 0.$$

By hypothesis, there exists a nonzero constant  $d$ , such that

$$(13) \quad S_{2^{k_0}}(x) \equiv d, \quad x \in I_0.$$

Thus by (12) and (10) ( $j = 1$ ), it follows that

$$S_{2^{k_0+1}}(x) - S_{2^{k_0}}(x) \equiv -d, \quad x \in \tilde{I}_1.$$

Applying (11), we conclude that

$$S_{2^{k_0+1}}(x) - S_{2^{k_0}}(x) \equiv d, \quad x \in I_1.$$

This identity, together with (10) and (13), implies that  $S_{2^{k_0+1}}(x) \equiv 2d$  when  $x \in I_1$ . Continuing by induction we have that

$$(14) \quad |S_{2^{k_0+j+1}}(x) - S_{2^{k_0+j}}(x)| \equiv 2^j |d|, \quad x \in \tilde{I}_j,$$

for each integer  $j > 0$ .

It remains to see that (14) and condition  $H$  are incompatible. We shall verify this fact by showing that (14) leads to an inequality of the form

$$\int_0^1 \Sigma_n(x) dx \geq n \cdot c$$

for large  $n$ , where  $c$  is a certain nonzero constant and where

$$\Sigma_n^2(x) \equiv \sum_{k=1}^n (S_{2^k}(x) - S_{2^{k-1}}(x))^2$$

for  $x \in [0, 1]$ ,  $n = 1, 2, \dots$ .

Toward this fix positive integers  $n$  and  $j$  with  $n > k_0 + j$ , and suppose that  $k_0 \leq k < k_0 + j$ . From (9) and (10) it follows that  $\tilde{I}_j \subseteq I_{k-k_0}$ . By (14), then, we have that

$$(15) \quad |S_{2^k}(x) - S_{2^{k-1}}(x)| \equiv 2^{k-k_0}|d|, \quad x \in \tilde{I}_j.$$

Moreover, the condition  $n > k_0 + j$  implies that

$$\Sigma_n^2 \geq \sum_{k=k_0}^{k_0+j-1} (S_{2^k} - S_{2^{k-1}})^2.$$

In particular, (15) yields the following estimate for  $j = 1, 2, \dots, n - k_0$ :

$$(16) \quad \Sigma_n(x) \geq 2^{j-1}|d|, \quad x \in \tilde{I}_j.$$

We are now prepared to estimate the integral of  $\Sigma_n$ . By conditions (9) and (10), the intervals  $\tilde{I}_j$  are pairwise disjoint and their union equals  $I_0$ . Consequently, we have that

$$\int_0^1 \Sigma_n(x) dx \geq \sum_{j=1}^{n-k_0} \int_{\tilde{I}_j} \Sigma_n(x) dx$$

for  $n > k_0$ . Applying (16) to each of these integrals, we conclude that

$$\int_0^1 \Sigma_n(x) dx \geq |d| \sum_{j=1}^{n-k_0} 2^{j-1} |\tilde{I}_j|.$$

However, according to (8), this inequality reduces to the following inequality, thereby completing the proof of the lemma:

$$\int_0^1 \Sigma_n(x) dx \geq (n - k_0) |d| 2^{-k_0-1}.$$

**3. A proof of the theorem.** Let  $\Delta$  denote the Walsh series whose coefficients are given by  $a_k - a_k(f)$ , for  $k = 0, 1, \dots$ . We have, then, that  $\Delta$  satisfies condition  $H$ , and that

$$(17) \quad \lim_{N \rightarrow \infty} \Delta_{2^N}(x) = 0 \quad \text{a.e. } x \in [0, 1].$$

Suppose that the theorem is false, i.e., that  $\Delta$  is not the zero series. Then there exists an integer  $k_0 \geq 0$  such that  $\Delta_{2^{k_0}}$  is not zero on an interval  $I_0$  of the form (6). If  $x_1$  is any given point, we can apply the lemma (by chasing down any properly nested sequence in  $I_0$  which eventually avoids  $x_1$ ) to choose an interval  $\tilde{I}_{k_1-k_0}$  such that  $\Delta_{2^{k_1}}$  is not zero on  $\tilde{I}_{k_1-k_0}$ . Thus, if we avoided  $x_1$  and both endpoints of  $I_0$ , we could choose an integer  $k_0^*$  and an interval  $I_0^* = [(p^* - 1)/2^{k_0^*}, p^*/2^{k_0^*}]$  whose closure  $F$  satisfied  $x_1 \notin F \subset I_0$  such that  $\Delta_{2^{k_0^*}}(x) \neq 0$  for  $x \in I_0^*$ .

It is our aim to carry this construction one step further. Specifically, we claim that given a constant  $M$ , there exists an interval  $I_1 = [(q_1 - 1)/2^{m_1}, q_1/2^{m_1}]$  such that  $I_1 \subset I_0^*$  and such that

$$(18) \quad |S_{2^{m_1}}(x)| > M, \quad x \in I_1.$$

Suppose for a moment that this claim has been established. If we denote the set of points

$x$  at which  $S_{2^N}(x)$  fails to converge to  $f(x)$  by  $E = \{x_1, x_2, \dots\}$  and apply the above construction countably many times, we generate intervals  $I_1, I_2, \dots$  and integers  $m_1, m_2, \dots$ , such that for each integer  $j > 0$ ,  $x_j$  does not belong to the closure of  $I_j$ , the closure of  $I_j$  is contained in  $I_{j-1}$ , and

$$(19) \quad |S_{2^{m_j}}(x)| > j, \quad x \in I_j.$$

Let  $\xi$  be a point which belongs to all the intervals  $I_j$ . Then since  $\xi \neq x_j$ ,  $j = 1, 2, \dots$ , it must be the case that  $S_{2^N}(\xi) \rightarrow f(\xi)$ , as  $N \rightarrow \infty$ . In particular, (19) implies that  $f(\xi) = \pm\infty$ , which contradicts the fact that  $f$  is finite-valued.

It suffices, therefore, to establish the claim which ends with inequality (18). Toward this, suppose to the contrary that there exists a number  $M_0$  such that for all intervals  $I \subset I_0^*$  of the form  $[(q-1)/2^m, q/2^m)$  there exists a point in  $I$  at which the absolute value of  $S_{2^m}$  does not exceed  $M_0$ . Since  $S_{2^m}$  is constant on each of the intervals  $I$ , and since  $I_0^*$  can be written as a union of such intervals for each fixed  $m \geq k_0^*$ , it follows that

$$(20) \quad |S_{2^m}(x)| \leq M_0, \quad x \in I_0^*, \quad m \geq k_0^*.$$

Recall that  $\Delta_{2^{k_0^*}}(x) \equiv d^* \neq 0$  for  $x \in I_0^*$ . This means that

$$(21) \quad 0 \neq d^*/2^{k_0^*} \equiv \int_{I_0^*} \Delta_{2^m}(t) dt, \quad m \geq k_0^*,$$

since the integral over  $I_0^*$  of each Walsh function  $\psi_j$  with  $j \geq 2^{k_0^*}$  is identically zero. However, (20) together with the bounded convergence theorem, implies that  $S_{2^m}$  converges to  $f$  in  $L^1(I_0^*)$ , as  $m \rightarrow \infty$ . Also, since  $f$  is integrable, we have that  $S_{2^m}[f]$  converges to  $f$  in  $L^1(I_0^*)$ , as  $m \rightarrow \infty$ . Consequently,  $\Delta_{2^m}$  converges to zero in  $L^1(I_0^*)$ , as  $m \rightarrow \infty$ . Since this conclusion and (21) are incompatible, we have arrived at a contradiction. In particular, the claim is established and the proof of the theorem is complete.

#### REFERENCES

- [1] K. G. BEAUCHAMP, *Walsh Functions and their Applications*, Academic Press, New York, 1975.
- [2] R. B. CRITTENDEN AND V. L. SHAPIRO, *Sets of uniqueness on the group  $2^\omega$* , *Ann. of Math.*, 81 (1965), pp. 550-564.
- [3] N. J. FINE, *On the Walsh functions*, *Trans. Amer. Math. Soc.*, 65 (1949), pp. 372-414.
- [4] A. M. GARSIA, *Martingale Inequalities*, W. A. Benjamin, Reading, MA, 1973.
- [5] H. F. HARMUTH, *Transmission of Information by Orthogonal Functions*, Springer-Verlag, New York, 1972.

## SOME BASIC HYPERGEOMETRIC EXTENSIONS OF INTEGRALS OF SELBERG AND ANDREWS\*

RICHARD ASKEY†

**Abstract.** A. Selberg evaluated an important multivariable extension of the beta function integral. Andrews found a related integral and evaluated it using a result of Dyson, Gunson and Wilson. Basic hypergeometric, or  $q$ -series, extensions of these integrals are considered and evaluated in the two-dimensional case. Conjectures are given for the values of these integrals in the  $n$ -dimensional case.

**1. Introduction.** In an almost unknown paper, A. Selberg [35] evaluated an important multivariable integral of the beta function type. He showed that

$$(1.1) \quad \int_0^1 \cdots \int_0^1 \prod_{i=1}^n t_i^{x-1} (1-t_i)^{y-1} \left| \prod_{1 \leq i < j \leq n} (t_i - t_j) \right|^{2z} dt_1 \cdots dt_n \\ = \prod_{j=1}^n \frac{\Gamma(x + (j-1)z)\Gamma(y + (j-1)z)\Gamma(jz + 1)}{\Gamma(x + y + (n+j-2)z)\Gamma(z + 1)},$$

$\text{Re } x > 0, \text{Re } y > 0, \text{Re } z > -[1/n, \text{Re } x/(n-1), \text{Re } y/(n-1)]$ .

Many important integrals are special or limiting cases of this integral. One is an integral considered by Mehta and Dyson [28] (see also Mehta [27]).

$$(1.2) \quad \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \sum_{i=1}^n t_i^2\right) \left| \prod_{1 \leq i < j \leq n} (t_i - t_j) \right|^{2z} dt_1 \cdots dt_n \\ = \prod_{j=1}^n \frac{\Gamma(jz + 1)}{\Gamma(z + 1)}.$$

This can be obtained by setting  $y = x$  and  $2t_i = 1 + s_i(2x)^{-1/2}$  and using Stirling's formula. Another integral is

$$(1.3) \quad \int_0^{\infty} \cdots \int_0^{\infty} \prod_{i=1}^n t_i^{x-1} e^{-t_i} \left| \prod_{1 \leq i < j \leq n} (t_i - t_j) \right|^{2z} dt_1 \cdots dt_n \\ = \prod_{j=1}^n \frac{\Gamma(x + (j-1)z)\Gamma(jz + 1)}{\Gamma(z + 1)}.$$

Just let  $t_i \rightarrow t_i/y$  and take  $y \rightarrow \infty$ .

Letting  $t_i \rightarrow t_i^2, x = \frac{1}{2}$ , in (1.3) gives

$$(1.4) \quad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\sum_{i=1}^n t_i^2\right) \left| \prod_{1 \leq i < j \leq n} (t_i^2 - t_j^2) \right|^{2z} dt_1 \cdots dt_n \\ = \prod_{j=1}^n \frac{\Gamma(\frac{1}{2} + (j-1)z)\Gamma(jz + 1)}{\Gamma(z + 1)}.$$

These integrals have been used to prove some conjectures of Macdonald [26]. Macdonald has some other conjectures that could be proven if an appropriate extension of Selberg's integral could be found. Three conjectured extensions will be stated and

---

\* Received by the editors March 3, 1980. This research was supported in part by the National Science Foundation under Grant MCS-78-07244.

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

proven in the two-dimensional case. Before stating these the appropriate one-dimensional extension of the beta function will be given. The first is just a disguised form of the  $q$ -binomial theorem.

Fix  $q$ ,  $0 < q < 1$ . Define the  $q$ -gamma function of F. H. Jackson by

$$(1.5) \quad \Gamma_q(x) = \frac{(q; q)_\infty}{(q^x; q)_\infty} (1-q)^{1-x},$$

where

$$(1.6) \quad (a; q)_\infty = \prod_{n=0}^{\infty} (1 - aq^n).$$

If

$$(1.7) \quad (a; q)_n = \frac{(a; q)_\infty}{(aq^n; q)_\infty},$$

then the  $q$ -binomial theorem can be given as

$$(1.8) \quad \frac{(ax; q)_\infty}{(x; q)_\infty} = \sum_{n=0}^{\infty} \frac{(a; q)_n}{(q; q)_n} x^n.$$

See [9, p. 66] and [7] for simple proofs.

Following F. H. Jackson, define

$$(1.9) \quad \int_0^1 f(x) d_q x = (1-q) \sum_{n=0}^{\infty} f(q^n) q^n.$$

The  $q$ -binomial theorem can then be stated as

$$(1.10) \quad \int_0^1 t^{x-1} \frac{(tq; q)_\infty}{(tq^y; q)_\infty} d_q t = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}, \quad \operatorname{Re} x > 0.$$

Since  $\lim_{q \rightarrow 1} \Gamma_q(x) = \Gamma(x)$ , [6], and

$$\begin{aligned} \lim_{q \rightarrow 1} \frac{(tq; q)_\infty}{(tq^y; q)_\infty} &= \lim_{q \rightarrow 1} \sum_{n=0}^{\infty} \frac{(q^{1-y}; q)_n}{(q; q)_n} (tq^y)^n \\ &= \sum_{n=0}^{\infty} \frac{(1-y)_n}{n!} t^n = (1-t)^{y-1}, \end{aligned}$$

formula (1.10) extends Euler's formula

$$(1.11) \quad \int_0^1 t^{x-1} (1-t)^{y-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad \operatorname{Re} x > 0, \quad \operatorname{Re} y > 0.$$

As far as I know the  $q$ -binomial theorem was first stated by Rothe [32].

Ramanujan found two other extensions of the beta function. After changing variables by  $t \rightarrow t/(1+t)$  in (1.11), the resulting integral is

$$(1.12) \quad \int_0^\infty \frac{t^{x-1}}{(1+t)^{x+y}} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad \operatorname{Re} x > 0, \quad \operatorname{Re} y > 0.$$



Ramanujan extended this to

$$(1.13) \quad \int_0^\infty t^{x-1} \frac{(-tq^{x+y}; q)_\infty}{(-t; q)_\infty} dt = \frac{\Gamma(x)\Gamma(1-x)\Gamma_q(y)}{\Gamma_q(x+y)\Gamma_q(1-x)}, \quad \text{Re } x > 0, \quad \text{Re } y > 0.$$

A simple proof is given in [7]. His other extension uses

$$(1.14) \quad \int_0^\infty f(t) d_q t = (1-q) \sum_{-\infty}^\infty f(q^n) q^n.$$

Ramanujan also found that [31, p. 196, #17]

$$(1.15) \quad \sum_{-\infty}^\infty \frac{(a; q)_n}{(b; q)_n} x^n = \frac{(ax; q)_\infty \left(\frac{q}{ax}; q\right)_\infty (q; q)_\infty \left(\frac{b}{a}; q\right)_\infty}{(x; q)_\infty \left(\frac{b}{ax}; q\right)_\infty (b; q)_\infty \left(\frac{q}{a}; q\right)_\infty}, \quad \left|\frac{b}{a}\right| < |x| < 1.$$

A proof of this is also given in [7]. References to other proofs are given there. This implies that

$$(1.16) \quad \int_0^\infty \frac{(-cq^{x+y}t; q)_\infty}{(-ct; q)_\infty} t^{x-1} d_q t = \frac{\Gamma_q(x)\Gamma_q(y)(-cq^x; q)_\infty(-1^{1-x}c^{-1}; q)_\infty}{\Gamma_q(x+y)(-c; q)_\infty(-qc^{-1}; q)_\infty},$$

Re  $x$ , Re  $y$  > 0.

When  $q \rightarrow 1$  and  $c > 0$ , this converges to

$$\int_0^\infty \frac{t^{x-1}}{(1+ct)^{x+y}} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \frac{(1+c^{-1})^x}{(1+c)^x} = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} c^{-x}.$$

**2. The first extension of Selberg’s integral.** Selberg [35] first evaluated (1.1) when  $z = k$ , a positive integer, and then obtained the general result by use of a uniqueness theorem for analytic functions (a special case of Carlson’s theorem [11] can be used). At present I know how to obtain some results only when  $z$  is a positive integer. An extension of  $t_i^{x-1}(1-t_i)^{y-1}$  is given by  $t_i^{x-1}(t_i q; q)_\infty / (t_i q^y; q)_\infty$ . The problem comes from trying to extend  $\prod_{i < j} (t_i - t_j)^{2k}$ . We want a nonnegative function of order  $2k$  which vanishes when  $q \rightarrow 1$ . This is provided by

$$(2.1) \quad \prod_{1 \leq i < j \leq n} t_i^{2k} \left(\frac{t_j q^{-k}}{t_i}; q\right)_{2k},$$

which vanishes when  $t_i = t_j$  and on  $k$  lines on one side of this line and  $k - 1$  on the other side. By symmetry we could take

$$(2.2) \quad \prod_{1 \leq i < j \leq n} t_i^{2k} \left(\frac{t_j q^{1-k}}{t_i}; q\right)_{2k},$$

which also vanishes on  $k$  lines on one side of  $t_i = t_j$  and  $k - 1$  on the other side, but the sides are interchanged.

One two-dimensional extension of Selberg’s integral is

$$(2.3) \quad I_2(x, y, k) = \int_0^1 \int_0^1 t_1^{x-1} t_2^{y-1} \frac{(t_1 q; q)_\infty}{(t_1 q^y; q)_\infty} \frac{(t_2 q; q)_\infty}{(t_2 q^y; q)_\infty} t_1^{2k} \left(\frac{t_2 q^{-k}}{t_1}; q\right)_{2k} d_q t_1 d_q t_2.$$

Expand  $((t_2q^{-k}/t_1); q)_{2k}$  by the  $q$ -binomial theorem to get

$$\begin{aligned} I_2(x, y, k) &= \sum_{j=0}^{2k} \frac{(q^{-2k}; q)_j}{(q; q)_j} q^{jk} \int_0^1 t_1^{x+2k-j-1} \frac{(t_1q; q)_\infty}{(t_1q^y; q)_\infty} d_q t_1 \int_0^1 t_2^{x+j-1} \frac{(t_2q; q)_\infty}{(t_2q^y; q)_\infty} d_q t_2 \\ &= \sum_{j=0}^{2k} \frac{(q^{-2k}; q)_j q^{jk} \Gamma_q(x+2k-j) \Gamma_q(y) \Gamma_q(x+j) \Gamma_q(y)}{(q; q)_j \Gamma_q(x+y+2k-j) \Gamma_q(x+y+j)} \\ &= \frac{\Gamma_q(x) \Gamma_q(x+2k) \Gamma_q(y) \Gamma_q(y)}{\Gamma_q(x+y) \Gamma_q(x+y+2k)} {}_3\phi_2 \left( \begin{matrix} q^{-2k}, q^x, q^{1-x-y-2k} \\ q^{1-x-2k}, q^{x+y} \end{matrix}; q, q^{k+y} \right). \end{aligned}$$

The basic hypergeometric series  ${}_{p+1}\phi_p$  is defined by

$$(2.4) \quad {}_{p+1}\phi_p \left( \begin{matrix} a_0, \dots, a_p \\ b_1, \dots, b_p \end{matrix}; q, x \right) = \sum_{n=0}^{\infty} \frac{(a_0; q)_n \dots (a_p; q)_n}{(q; q)_n (b_1; q)_n \dots (b_p; q)_n} x^n.$$

Jackson [19] evaluated a series which is equivalent to

$$(2.5) \quad {}_3\phi_2 \left( \begin{matrix} q^{-2k}, a, b \\ q^{1-2k}, q^{1-2k} \end{matrix}; q, \frac{q^{1-k}}{ab} \right) = \frac{(q^{k+1}; q)_k (abq^k; q)_k}{(aq^k; q)_k (bq^k; q)_k} q^{-k}.$$

Carlitz [12] gave another proof, first proving a quadratic transformation between two basic hypergeometric series, then obtaining (2.5) as one term of this series identity. He also obtained the identity

$$(2.6) \quad {}_3\phi_2 \left( \begin{matrix} q^{-2k}, a, b \\ q^{1-2k}, q^{1-2k} \end{matrix}; q, \frac{q^{2-k}}{ab} \right) = \frac{(q^{k+1}; q)_k (abq^k; q)_k}{(aq^k; q)_k (bq^k; q)_k}.$$

Using Jackson's sum gives

$$\begin{aligned} I_2(x, y, k) &= \frac{\Gamma_q(x) \Gamma_q(x+2k) \Gamma_q(y) \Gamma_q(y) q^{-k} (q^{k+1}; q)_k (q^{1-y-k}; q)_k}{\Gamma_q(x+y) \Gamma_q(x+y+2k) (q^{x+k}; q)_k (q^{1-x-y-k}; q)_k} \\ (2.7) \quad &= \frac{\Gamma_q(x) \Gamma_q(x+k) \Gamma_q(y) \Gamma_q(y) \Gamma_q(2k+1)}{\Gamma_q(x+y) \Gamma_q(x+y+2k) \Gamma_q(k+1)} q^{(x-1)k} \frac{(q^y; q)_k}{(q^{x+y}; q)_k} \\ &= \frac{\Gamma_q(x) \Gamma_q(x+k) \Gamma_q(y) \Gamma_q(y+k) \Gamma_q(k+1) \Gamma_q(2k+1)}{\Gamma_q(x+y+k) \Gamma_q(x+y+2k) [\Gamma_q(k+1)]^2} q^{(x-1)k}. \end{aligned}$$

Either by symmetry, or from Carlitz's second sum (2.6), one obtains

$$\begin{aligned} (2.8) \quad &\int_0^1 \int_0^1 t_1^{x-1} t_2^{x-1} \frac{(t_1q; q)_\infty}{(t_1q^y; q)_\infty} \frac{(t_2q; q)_\infty}{(t_2q^y; q)_\infty} t_1^{2k} \left( \frac{t_2q^{1-k}}{t_1}; q \right)_{2k} d_q t_1 d_q t_2 \\ &= \frac{\Gamma_q(x) \Gamma_q(x+k) \Gamma_q(y) \Gamma_q(y+k) \Gamma_q(k+1) \Gamma_q(2k+1)}{\Gamma_q(x+y+k) \Gamma_q(x+y+2k) [\Gamma_q(k+1)]^2} q^{xk}. \end{aligned}$$

Selberg's proof of (1.1) is very ingenious. Unfortunately it does not seem to work in the more general setting being considered. He used the symmetry of (1.1) in  $x$  and  $y$  in an essential way. That is easy to see by replacing  $t_i$  by  $1 - s_i$ . The symmetry of (1.10) in  $x$  and  $y$  is not obvious, though it is true. This symmetry fails in the two-dimensional case.

Thus Selberg’s proof does not extend. No substitute proof has been found, so we will have to settle for a conjecture in dimensions three or more.

*Conjecture 1.*

$$\begin{aligned}
 (2.9) \quad & \int_0^1 \cdots \int_0^1 \prod_{i=1}^n t_i^{x-1} \frac{(t_i q; q)_\infty}{(t_i q^y; q)_\infty} \prod_{1 \leq i < j \leq n} t_i^{2k} \left( \frac{t_j q^{1-k}}{t_i}; q \right)_{2k} d_q t_1 \cdots d_q t_n \\
 & = q^{a(k,n)x + b(k,n)} \prod_{j=1}^n \frac{\Gamma_q(x + (j-1)k) \Gamma_q(y + (j-1)k) \Gamma_q(jk + 1)}{\Gamma_q(x + y + (n+j-2)k) \Gamma_q(k + 1)}.
 \end{aligned}$$

To show that it is likely that  $a(k, n) = k \binom{n}{2}$  and  $b(k, n) = 2k^2 \binom{n}{3}$ , we need to consider further extensions of Selberg’s integral.

**3. Further extensions of Selberg’s integral.** The first published statement of Selberg’s integral is of an integral on  $(0, \infty)$  which arises from (1.1) by the change of variables  $t_i = s_i(1 + s_i)^{-1}$ , [34].

$$\begin{aligned}
 (3.1) \quad & \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^n \frac{s_i^{x-1}}{(1 + s_i)^{x+y+2(n-1)z}} \Big|_{1 \leq i < j \leq n} (s_i - s_j) \Big|^{2z} ds_1 \cdots ds_n \\
 & = \prod_{j=1}^n \frac{\Gamma(x + (j-1)z) \Gamma(y + (j-1)z) \Gamma(jz + 1)}{\Gamma(x + y + (n+j-2)z) \Gamma(z + 1)}.
 \end{aligned}$$

A two-dimensional  $q$ -extension of the case  $z = k$  can be found as follows.

$$\begin{aligned}
 & \int_0^\infty \int_0^\infty s_1^{x-1} s_2^{x-1} \frac{(-s_1 q^{x+y+2k}; q)_\infty (-s_2 q^{x+y+2k}; q)_\infty}{(-s_1; q)_\infty (-s_2; q)_\infty} s_1^{2k} \left( \frac{s_2 q^{1-k}}{s_1}; q \right)_{2k} ds_1 ds_2 \\
 & = \sum_{j=0}^{2k} \frac{(q^{-2k}; q)_j}{(q; q)_j} q^{(k+1)j} \int_0^\infty s_1^{x+2k-j-1} \frac{(-s_1 q^{x+y+2k}; q)_\infty}{(-s_1; q)_\infty} ds_1 \\
 & \quad \cdot \int_0^\infty s_2^{x+j-1} \frac{(-s_2 q^{x+y+2k}; q)_\infty}{(-s_2; q)_\infty} ds_2 \\
 & = \sum_{j=0}^{2k} \frac{(q^{-2k}; q)_j}{(q; q)_j} q^{(k+1)j} \frac{\Gamma(x + 2k - j) \Gamma(1 - x - 2k + j) \Gamma_q(y + j)}{\Gamma_q(1 - x - 2k + j) \Gamma_q(x + y + 2k)} \\
 & \quad \cdot \frac{\Gamma(x + j) \Gamma(1 - x - j) \Gamma_q(y - j + 2k)}{\Gamma_q(1 - x - j) \Gamma_q(x + y + 2k)} \\
 & = \frac{\Gamma_q(y) \Gamma_q(y + 2k)}{\Gamma_q(1 - x - 2k) \Gamma_q(x + y) \Gamma_q(1 - x) \Gamma_q(x + y)} \sum_{j=0}^{2k} \frac{(q^{-2k}; q)_j}{(q; q)_j} q^{(k+1)j} \\
 & \quad \cdot \frac{\pi}{\sin \pi(x + 2k - j)} \frac{\pi}{\sin \pi(x + j)} \frac{(q^y; q)_j}{(q^{1-x-2k}; q)_j} \frac{(q^x; q)_j}{(q^{1-y-2k}; q)_j} \frac{q^{(1-x)j}}{q^{(y+2k)j}} \\
 & = \frac{[\Gamma(x) \Gamma(1-x)]^2 \Gamma_q(y) \Gamma_q(y + 2k)}{\Gamma_q(x + y) \Gamma_q(x + y + 2k) \Gamma_q(1-x) \Gamma_q(1-x-2k)} \\
 & \quad \cdot {}_3\phi_2 \left( \begin{matrix} q^{-2k}, q^y, q^x \\ q^{1-y-2k}, q^{1-x-2k} \end{matrix}; q, q^{2-k-x-y} \right) \\
 & = \frac{[\Gamma(x) \Gamma(1-x)]^2 \Gamma_q(y) \Gamma_q(y + 2k)}{\Gamma_q(x + y + 2k) \Gamma_q(x + y + 2k) \Gamma_q(1-x) \Gamma_q(1-x-2k)} \frac{(q^{k+1}; q)_k (q^{k+x+y}; q)_k}{(q^{k+x}; q)_k (q^{k+y}; q)_k}
 \end{aligned}$$

$$= \frac{\Gamma_q(x)\Gamma_q(x+k)\Gamma_q(y)\Gamma_q(y+k)\Gamma_q(k+1)\Gamma_q(2k+1)}{\Gamma_q(x+y+k)\Gamma_q(x+y+2k)\Gamma_q(k+1)\Gamma_q(k+1)} \cdot \frac{[\Gamma(x)\Gamma(1-x)]^2}{\Gamma_q(x)\Gamma_q(1-x)\Gamma_q(x+2k)\Gamma_q(1-x-2k)}.$$

This suggests the following conjecture.

*Conjecture 2.*

$$(3.2) \quad \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^n s_i^{x-1} \frac{(-s_i q^{x+y+2(n-1)k}; q)_\infty}{(-s_i; q)_\infty} \prod_{1 \leq i < j \leq n} s_i^{2k} \left( \frac{s_j q^{1-k}}{s_i}; q \right)_{2k} ds_1 \cdots ds_n$$

$$= \prod_{j=1}^n \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1)\Gamma(x)\Gamma(1-x)}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)\Gamma_q(x+2(j-1)k)\Gamma_q(1-x-2(j-1)k)}.$$

One trouble with the integral in this conjecture is that the integrand changes sign on the support of the measure. This drawback can be removed by using Ramanujan's sum (1.16) rather than his integral (1.13).

In two dimensions the result is

$$(3.3) \quad \int_0^\infty \int_0^\infty s_1^{x-1} s_2^{y-1} \frac{(-cs_1 q^{x+y+2k}; q)_\infty (-cs_2 q^{x+y+2k}; q)_\infty}{(-cs_1; q)_\infty (-cs_2; q)_\infty} s_1^{2k} \left( \frac{q^{1-k} s_2}{s_1}; q \right)_{2k} d_q s_1 d_q s_2$$

$$= \frac{[\Gamma_q(x)\Gamma_q(x+k)\Gamma_q(y)\Gamma_q(y+k)\Gamma_q(k+1)\Gamma_q(2k+1)(-cq^x; q)_\infty \cdot (-cq^{x+2k}; q)_\infty (-q^{1-x}/c; q)_\infty (-q^{1-x-2k}/c; q)_\infty]}{\Gamma_q(x+y+k)\Gamma_q(x+y+2k)\Gamma_q(k+1)\Gamma_q(k+1)(-c; q)_\infty^2 (-q/c; q)_\infty^2}.$$

The proof is identical to those given before, so it will not be repeated. It is now natural to make the following conjecture.

*Conjecture 3.*

$$(3.4) \quad \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^n s_i^{x-1} \frac{(-cs_i q^{x+y+2(n-1)k}; q)_\infty}{(-cs_i; q)_\infty} \prod_{1 \leq i < j \leq n} s_i^{2k} \left( \frac{s_j q^{1-k}}{s_i}; q \right)_{2k} d_q s_1 \cdots d_q s_n$$

$$= \prod_{j=1}^n \frac{[\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1) \cdot (-cq^{x+2(j-1)k}; q)_\infty (-q^{1-x-2(j-1)k} c^{-1}; q)_\infty]}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)(-c; q)_\infty (-qc^{-1}; q)_\infty}$$

$$= \prod_{j=1}^n \frac{[\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1) \cdot (-cq^x; q)_\infty (-q^{1-x} c^{-1}; q)_\infty q^{k(j-1)(1-2x)}]}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)(-c; q)_\infty (-qc^{-1}; q)_\infty c^{2(j-1)k} q^{2k^2(j-1)^2}}.$$

To complete the first conjecture take  $-ca^{x+y+2(n-1)k} = q$  to determine  $y$  and then take  $-c = q^y$  (a different  $y$ ). The left-hand side of (3.4) is then the left-hand side of (2.9). The right-hand side of (3.4) is

$$\prod_{j=1}^n \frac{[\Gamma_q(x+(j-1)k)\Gamma_q(1-x-y-(2n-j-1)k)\Gamma_q(jk+1) \cdot (q^{x+y+2(j-1)k}; q)_\infty (q^{1-x-y-2(j-1)k}; q)_\infty]}{\Gamma_q(1-y-nk+jk)\Gamma_q(k+1)(q^y; q)_\infty (q^{1-y}; q)_\infty}$$

$$= \prod_{j=1}^n \frac{[\Gamma_q(x+(j-1)k)\Gamma_q(1-x-y-(n+j-2)k)\Gamma_q(jk+1) \cdot (q^{x+y+2(j-1)k}; q)_\infty (q^{1-x-y-2(j-1)k}; q)_\infty]}{\Gamma_q(1-y-(j-1)k)\Gamma_q(k+1)(q^y; q)_\infty (q^{1-y}; q)_\infty}$$

$$= A \prod_{j=1}^n \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1)}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)},$$

where

$$\begin{aligned}
 A &= \prod_{j=1}^n \frac{[\Gamma_q(1-x-y-(n+j-2)k)\Gamma_q(x+y+(n+j-2)k)]}{(q^{x+y+2(j-1)k}; q)_\infty (q^{1-x-y-2(j-1)k}; q)_\infty} \\
 &= \prod_{j=1}^n \frac{(q^{y+(j-1)k}; q)_\infty (q^{1-y-(j-1)k}; q)_\infty (q^{x+y+2(j-1)k}; q)_\infty (q^{1-x-y-2(j-1)k}; q)_\infty}{(q^y; q)_\infty (q^{1-y}; q)_\infty (q^{x+y+(n+j-2)k}; q)_\infty (q^{1-x-y-(n+j-2)k}; q)_\infty} \\
 &= \prod_{j=1}^n \frac{(q^{1-y-(j-1)k}; q)_{(j-1)k} (q^{x+y+2(j-1)k}; q)_{(n-j)k}}{(q^y; q)_{(j-1)k} (q^{1-x-y-(n+j-2)k}; q)_{(n-j)k}} \\
 &= \prod_{j=1}^n \frac{(-1)^{(j-1)k} (-1)^{(n-j)k}}{q^{y(j-1)k} q^{((j-1)k)}} q^{[x+y+2(j-1)k](n-j)k} q^{\binom{n-j}{2}k} \\
 &= q^{kx\binom{n}{2} + 2k\binom{n}{3}}.
 \end{aligned}$$

A tedious calculation shows that Conjecture 1 is true with this value of  $A$  when  $n = 3, k = 1$ .

**4. A  $q$ -extension of Andrews' integral.** In addition to Selberg's integral and its special cases, there is a second identity that suggests that further interesting beta function integrals in several variables can be evaluated. This comes from another integral Dyson encountered. In an attempt to evaluate

$$(4.1) \quad \int_0^{2\pi} \cdots \int_0^{2\pi} \prod_{1 \leq i < j \leq n} |e^{i\theta_i} - e^{i\theta_j}|^z d\theta_1 \cdots d\theta_n,$$

Dyson was led to conjecture:

If  $a_i$  are nonnegative integers and if C.T.  $f(t_1, \dots, t_n)$  is the constant term in the Laurent expansion of  $f$ , then

$$(4.2) \quad \text{C.T.} \prod_{1 \leq i \neq j \leq n} \left(1 - \frac{t_i}{t_j}\right)^{a_i} = \frac{(a_1 + \cdots + a_n)!}{a_1! \cdots a_n!}.$$

He proved this when  $n = 2$  (easy),  $n = 3$  (equivalent to Dixon's sum of a terminating well-poised  ${}_3F_2$ ),  $n = 4$  and  $n = 5$ , and  $a_i \equiv 1, 2$  or  $4$  for general  $n$ . This conjecture was completely proven by Gunson [17] and Wilson [40], and finally a very elegant proof was found by Good [16]. Many of these papers have been reprinted in Porter [30], and a nice summary of much of the work that led to this problem and the Mehta-Dyson conjecture is given in Mehta [27]. Andrews [3] found an integral similar to Selberg's integral that contains this conjecture. Selberg's integral gives Dyson's conjecture when  $a_i$  are all equal, as will appear in a more general context in the next section. In a preliminary version of his integral, Andrews considered

$$(4.3) \quad \int_0^1 \cdots \int_0^1 \prod_{i=j}^n t_i^{x-1} (1-t_i)^{a_i-x} \prod_{1 \leq i \neq j \leq n} \left(1 - \frac{t_i}{t_j}\right)^{a_i} dt_1 \cdots dt_n.$$

Unfortunately this integral does not converge for any  $x$  when  $n \geq 3$ , so he had to use a more complicated double circuit integral representation for the beta function. (See [3] for details.) In the  $q$ -case it will be possible to extend (4.3).

The first  $q$ -extension in several variables was Andrews' extension of (4.2). (See [1].)

$$(4.4) \quad (\text{Conjecture}) \quad \text{C.T.} \prod_{1 \leq i \neq j \leq n} \left(\frac{t_i}{t_j} \varepsilon_{ij}; q\right)_{a_i} = \frac{(q; q)_{a_1 + \cdots + a_n}}{(q; q)_{a_1} \cdots (q; q)_{a_n}},$$

with  $\varepsilon_{ij} = 1$  when  $i < j$  and  $\varepsilon_{ij} = q$  when  $i > j$ . He proved this when  $n = 2$  and  $n = 3$  and Macdonald [26] proved it when  $a_i \equiv 1, a_i \equiv 2$  and  $a_i \equiv \infty$  for all  $n$ .

The same type of argument used above gives

$$(4.5) \quad \int_0^1 \int_0^1 t_1^{x-1} t_2^{x-1} \frac{(t_1 q; q)_\infty}{(t_1 q^{a_1+1-x}; q)_\infty} \frac{(t_2 q; q)_\infty}{(t_2 q^{a_2+1-x}; q)_\infty} \left(\frac{t_1}{t_2}; q\right)_{a_1} \left(\frac{t_2 q}{t_1}; q\right)_{a_2} d_q t_1 d_q t_2 \\ = \frac{\Gamma_q(a_1 + a_2 + 1 - x) [\Gamma_q(x) \Gamma_q(1 - x)]^2}{\Gamma_q(a_1 + 1) \Gamma_q(a_2 + 1) \Gamma_q(1 - x)}.$$

This suggests the following conjecture.

*Conjecture 4.* If  $\text{Re } x$  is sufficiently large and  $a_i$  are nonnegative integers, then

$$(4.6) \quad \int_0^1 \cdots \int_0^1 \prod_{i=1}^n t_i^{x-1} \frac{(t_i q; q)_\infty}{(t_i q^{a_i+1-x}; q)_\infty} \prod_{i \neq j \leq n} \left(\frac{t_i}{t_j} \varepsilon_{ij}; q\right)_{a_i} d_q t_1 \cdots d_q t_n \\ = \frac{\Gamma_q(a_1 + \cdots + a_n + 1 - x) [\Gamma_q(x) \Gamma_q(1 - x)]^n}{\Gamma_q(a_1 + 1) \cdots \Gamma_q(a_n + 1) \Gamma_q(1 - x)}.$$

Andrews' conjecture (4.4) follows from (4.6). To see this, multiply both sides by  $((1 - q^x)/(1 - q))^n$ .

The resulting identity can be written as

$$\left(\frac{1 - q^x}{1 - q}\right)^n \sum_{i_1, \dots, i_n} a(i_1, \dots, i_n) \prod_{j=1}^n \frac{\Gamma_q(x + \alpha_j) \Gamma_q(a_j + 1 - x)}{\Gamma_q(\alpha_j + a_j + 1)} \\ = \frac{\Gamma_q(a_1 + \cdots + a_n + 1 - x) [\Gamma_q(x + 1) \Gamma_q(1 - x)]^n}{\Gamma_q(a_1 + 1) \cdots \Gamma_q(a_n + 1) \Gamma_q(1 - x)},$$

with  $\alpha_1 + \cdots + \alpha_n = 0$ , because  $(t_i \varepsilon_{ij}/t_j; q)_{a_i}$  is homogeneous of degree zero. The only terms on the left-hand side that survive when  $x \rightarrow 0$  are those that come from the constant term in  $\prod_{i \neq j} (t_i \varepsilon_{ij}/t_j; q)_{a_i}$ , for  $\lim_{x \rightarrow 0} (1 - q^x) \Gamma_q(x + \alpha_j)$  exists for all  $\alpha_j$  and is zero when  $\alpha_j > 0$ .

It is natural to ask if Andrews' integral has other  $q$ -extensions, in the same way Selberg's integral was extended in § 3. In the two-dimensional case a different  ${}_3F_2$  arises, which probably cannot be summed. When a change of variables is done on (4.3) in the case  $n = 2$  and the integral is evaluated by expanding  $(t_1 - t_2)^{a_1 + a_2}$ , the resulting series has the form  ${}_3F_2\left(\begin{matrix} a, 1 - a, c \\ d, e \end{matrix}; 1\right)$  with  $2c + 1 = d + e$ . Such series were summed by Whipple. (See Bailey [9, p. 16].) They can be summed because Dixon's sum of the well-poised  ${}_3F_2$ ,

$$(4.7) \quad {}_3F_2\left(\begin{matrix} a, b, c \\ a + 1 - b, a + 1 - c \end{matrix}; 1\right) = \frac{\Gamma\left(1 + \frac{a}{2}\right) \Gamma(1 + a - b) \Gamma(1 + a - c) \Gamma\left(1 + \frac{a}{2} - b - c\right)}{\Gamma(1 + a) \Gamma\left(1 + \frac{a}{2} - b\right) \Gamma\left(1 + \frac{a}{2} - c\right) \Gamma(1 + a - b - c)}$$

can be transformed to Whipple's sum using a transformation of Kummer,

$$(4.8) \quad {}_3F_2\left(\begin{matrix} a, b, c \\ d, e \end{matrix}; 1\right) = \frac{\Gamma(e) \Gamma(d + e - a - b - c)}{\Gamma(e - c) \Gamma(d + e - a - b)} {}_3F_2\left(\begin{matrix} d - a, d - b, c \\ d, d + e - a - b \end{matrix}; 1\right).$$

Basic hypergeometric extensions of (4.7) and (4.8) exist. The extension of (4.8) is

$$(4.9) \quad {}_3\phi_2\left(a, b, c; d, e; q, \frac{de}{abc}\right) = \frac{\left(\frac{e}{c}; q\right)_\infty \left(\frac{de}{ab}; q\right)_\infty}{(e; q)_\infty \left(\frac{de}{abc}; q\right)_\infty} {}_3\phi_2\left(\frac{d}{a}, \frac{d}{b}, c; d, \frac{de}{ab}; q, \frac{e}{c}\right).$$

Two extensions of (4.7) when it terminates are (2.5) and (2.6). The power series variables in these cases are  $(de/(abc))^{1/2}$  and  $(deq^2/(abc))^{1/2}$  rather than  $de/(abc)$ , so these two identities cannot be combined. A  $q$ -extension of Whipple’s sum as a  ${}_4\phi_3$  was found by Andrews [2]. I have, however, been unable to find an integral that extends (4.3) that leads to this sum, or to evaluate any of the sums that arose in other attempts. There may be an extension, but if so it is still hidden.

**5. Connection between some of the conjectures.** The  $q$ -analogue of Andrews’ integral has a connection with the first  $q$ -extension of Selberg’s integral contained in Conjecture 1. To see this, observe that

$$(5.1) \quad t_i^{2k} \left(\frac{q^{1-k}t_j}{t_i}; q\right)_{2k} = (-1)^k (t_i t_j)^k q^{-\binom{k}{2}} \left(\frac{t_i}{t_j}; q\right)_k \left(\frac{t_j q}{t_i}; q\right)_k,$$

so Conjecture 1 can be written as

$$(5.2) \quad \int_0^1 \cdots \int_0^1 \prod_{1 \leq i \neq j \leq n} \left(\frac{t_i}{t_j} \varepsilon_{ij}; q\right)_k \prod_{i=1}^n t_i^{x+(n-1)k-1} \frac{(t_i q; q)_\infty}{(t_i q^y; q)_\infty} d_q t_i \\ = (-1)^k \binom{n}{2} \binom{n}{2} \binom{k}{2} + kx \binom{n}{2} + 2k^2 \binom{n}{3} \prod_{j=1}^n \frac{\Gamma_q(x+(j-1)k) \Gamma_q(y+(j-1)k) \Gamma_q(jk+1)}{\Gamma_q(x+y+(n+j-2)k) \Gamma_q(k+1)}.$$

When  $x$  is replaced by  $x - (n - 1)k$  and  $y$  is replaced by  $k + 1 - x$ , then (5.2) can be shown to reduce to the special case  $a_1 = \cdots = a_n = k$  of (4.6). Since Conjecture 4 implies Andrews’ conjecture (4.4), Conjecture 1 would imply the special case  $a_i = k$  of Andrews’ conjecture.

Macdonald obtained his  $q = 1$  conjectures for  $B_n, C_n, D_n, B'_n, C'_n$  and  $BC_n$  from Selberg’s integral using the change of variables  $t_i = \sin^2 \theta_i$ . Such a change of variables cannot be done in these conjectures. It is, however, possible to change variables quadratically in a  $q$ -integral, for

$$(5.3) \quad \int_0^1 f(t) d_{q^2} t = (1+q) \int_0^1 f(t^2) t d_q t = (1-q^2) \sum_{i=0}^\infty f(q^{2i}) q^{2i}.$$

Then

$$(5.4) \quad t_i^{4k} \left(\frac{q^{2-2k}t_j^2}{t_i^2}; q^2\right)_{2k} = t_i^{2k} \left(\frac{q^{1-k}t_j}{t_i}; q\right)_{2k} t_i^{2k} \left(\frac{-q^{1-k}t_j}{t_i}; q\right)_{2k}$$

is a natural extension of  $(t_i^2 - t_j^2)^{2k}$ , so it may be possible to obtain more of Macdonald’s conjectures once Conjecture 1 is proven.

**6. Another consequence of Conjecture 1.** There is a second way to let  $q \rightarrow 1$  in Conjecture 1 which leads to an identity different from Selberg’s integral. Multiply both sides of (2.9) by  $((q^y; q)_\infty^n / (q; q)_\infty^n) (1 - q)^{2k} \binom{n}{2}^{-n}$ , take  $q^x = a$  and  $y = -N$ , and let  $q \rightarrow 1$ .

The result is

*Conjecture 5.*

$$(6.1) \quad \sum_{m_1, \dots, m_n=0}^N \prod_{1 \leq i < j \leq n} (1 - k + m_j - m_i)_{2k} \prod_{i=1}^n a^{m_i} \frac{(-N)_{m_i}}{m_i!}$$

$$= (1 - a)^{Nn - 2k} a^{k \binom{n}{2}} \prod_{j=1}^n (-N)_{(j-1)k} \frac{(jk)!}{k!}.$$

This can be written as

*Conjecture 5'.*

$$(6.2) \quad \sum_{m_1, \dots, m_n=0}^N \prod_{1 \leq i < j \leq n} (1 - k + m_j - m_i)_{2k} \prod_{i=1}^n \binom{N}{m_i} p^{m_i} (1 - p)^{N - m_i}$$

$$= \prod_{j=1}^n [p(1 - p)]^{(j-1)k} \frac{(jk)!}{k!} \frac{N!}{[N - (j-1)k]!}.$$

In a similar fashion, it is possible to take a limit in Conjecture 3. In a form similar to (6.2), it is

*Conjecture 6.*

$$(6.3) \quad \sum_{m_1, \dots, m_n=0}^{\infty} \prod_{1 \leq i < j \leq n} (1 - k + m_j - m_i)_{2k} \prod_{i=1}^n \frac{(\beta)_{m_i}}{m_i!} c^{m_i} (1 - c)^{\beta}$$

$$= \prod_{j=1}^n \left[ \frac{c}{(1 - c)^2} \right]^{(j-1)k} \frac{(jk)!}{k!} (\beta)_{(j-1)k}.$$

Because Conjecture 1 has been proven when  $n = 2$ , Conjectures 5 and 6 also hold when  $n = 2$ .

There is a possible extension of Conjecture 5' which would contain the case  $z = k$  of Selberg's integral. It uses the hypergeometric distribution

$$(6.4) \quad \sum_{x=0}^N \binom{x + \alpha - 1}{x} \binom{N - x + \beta - 1}{N - x} = \binom{N + \alpha + \beta - 1}{N}.$$

A reasonable conjecture is then

*Conjecture 7.*

$$(6.5) \quad \sum_{x_1, \dots, x_n=0}^N \prod_{1 \leq i < j \leq n} (1 - k + x_j - x_i)_{2k} \prod_{i=1}^n \binom{x_i + \alpha - 1}{x_i} \binom{N - x_i + \beta - 1}{N - x_i}$$

$$= \prod_{j=1}^n \frac{(\alpha)_{(j-1)k} (\beta)_{(j-1)k} (\alpha + \beta)_{N + (j-1)k} (jk)!}{(\alpha + \beta)_{(n+j-2)k} (1)_{N - (j-1)k} k!}.$$

Finally, it is possible that Selberg's integral extends to the most general extension of the beta function using the  $q$ -integral of Jackson that I know. This extension of the beta function is

$$(6.6) \quad \int_{-c}^d \frac{\left(\frac{-qx}{c}; q\right)_{\infty} \left(\frac{qx}{d}; q\right)_{\infty}}{\left(\frac{-q^{\alpha}x}{c}; q\right)_{\infty} \left(\frac{q^{\beta}x}{d}; q\right)_{\infty}} d_q x = \frac{\Gamma_q(\alpha)\Gamma_q(\beta)}{\Gamma_q(\alpha + \beta)} \frac{cd}{(c + d)} \frac{\left(\frac{-c}{d}; q\right)_{\infty} \left(\frac{-d}{c}; q\right)_{\infty}}{\left(\frac{-q^{\beta}c}{d}; q\right)_{\infty} \left(\frac{-q^{\alpha}d}{c}; q\right)_{\infty}}.$$



(See [4].) On the basis of very scanty evidence I close this set of conjectures with *Conjecture 8*.

$$\int_{-c}^d \cdots \int_{-c}^d \prod_{1 \leq i < j \leq n} t_i^{2k} \left( \frac{t_j q^{1-k}}{t_i}; q \right)_{2k} \prod_{i=1}^n \frac{\left( \frac{-qt_i}{c}; q \right)_\infty \left( \frac{qt_i}{d}; q \right)_\infty}{\left( \frac{-q^x t_i}{c}; q \right)_\infty \left( \frac{q^y t_i}{d}; q \right)_\infty} d_q t_i$$

$$= \prod_{j=1}^n \frac{\Gamma_q(x + (j-1)k) \Gamma_q(y + (j-1)k) \Gamma_q(jk + 1) \left( \frac{-c}{d}; q \right)_\infty \left( \frac{-d}{c}; q \right)_\infty (cd)^{1+(j-1)k}}{\Gamma_q(x + y + (n+j-2)k) \Gamma_q(k + 1) \left( \frac{-c}{d} q^{y+(j-1)k}; q \right)_\infty \left( \frac{-d}{c} q^{x+(j-1)k}; q \right)_\infty (c+d)}$$

The symmetry in Selberg’s integral that is missing in Conjecture 1 has been restored. Also, Conjecture 8 is more general than Conjectures 1 and 7 and contains a direct extension of Conjecture 7 when  $-c = q^N d$ . Unfortunately I have not yet figured out how to expand  $t_1^{2k} (q^{1-k} t_2 / t_1; q)_{2k}$  in an appropriate series so as to prove the two-dimensional case of this conjecture. It has been checked in one nontrivial case,  $k = 1, n = 2$ .

**7. Connection with other topics.** Multivariate versions of the gamma function and associated integrals have arisen in a number of different contexts. Wishart [41] introduced a multivariate normal distribution and evaluated its integral. This is closely related to the case  $z = \frac{1}{2}$  in (1.2). Statisticians have primarily considered matrices with real entries, and so have encountered the special case of Selberg’s integral when  $z = \frac{1}{2}$ . (See Wilks [39, Chap. 18] for some examples.) Hermitian and symplectic matrices correspond to  $z = 1$  and  $z = 2$ , respectively. (See Mehta [27, p. 38].) The normal distribution cases were considered by Mehta and Dyson [28]. Dyson [13] also gave one reason why it is necessary to extend the results to other  $z$ . He wished to differentiate an identity. The earliest special case of Selberg’s integral that was considered is the limiting case when  $z \rightarrow \infty$ . Stieltjes considered the problem of finding the maximum value of

$$\prod_{i=1}^n t_i^a (1 - t_i)^b \prod_{1 \leq i < j \leq n} |t_i - t_j|,$$

for  $0 \leq t_i \leq 1, a, b > 0$ . He showed that the maximum occurs at the zeros of

$$p_n(t) = {}_2F_1 \left( \begin{matrix} -n, n + 2a + 2b - 1 \\ 2a \end{matrix}; t \right),$$

polynomials which are orthogonal on  $[0, 1]$  with respect to  $t^{2a-1} (1-t)^{2b-1}$ . (See Szegő [37, Chap. VI].) Stieltjes also evaluated the maximum value of this function. If  $I_n(x, y, z)$  is the integral in (1.1), then the maximum of this function can be computed using Stirling’s formula, for it is  $\lim_{z \rightarrow \infty} [I(2az, 2bz, z)]^{1/(2z)}$ . Since  $\prod_{i=1}^n t_i^a (1 - t_i)^b$  can be computed easily when  $t_i$  are the zeros of  $p_n(t)$ , this gives a new evaluation of the discriminant of the polynomial  ${}_2F_1$ . See Szegő [37, Chap. VI] for another derivation of this result and references to the earlier treatments of Stieltjes, Hilbert and Schur.

Another occurrence of a special case of Selberg’s integral is in work on moment spaces.  $I_n(1, 1, 2)$  was used by Karlin and Shapley [22, Thm. 15.1] to find the  $n$ -dimensional volume of a moment space. (Also see Karlin and Studden [23, p. 129] and Schoenberg [33].)

There is another connection with orthogonal polynomials that promises to be very important in the future. This concerns polynomials of several variables which are

orthogonal with respect to the integrand in (1.1). When  $z = \frac{1}{2}$  in (1.1), this is treated by James and Constantine [21]. See James [20] for more on this and limiting cases that correspond to (1.2) and (1.3). He also gives further references to other work. The general case of (1.1) in two dimensions has led to a set of orthogonal polynomials in two variables. See Koornwinder [24, Case VI, p. 450] and references to earlier work of his. Later work is surveyed by Sprinkhuizen [36]. Other references there are to Berezin and Karpelevic [10] and Maass [25]. Herz [18] should also be consulted.

Another important paper is Gindikin [15]. He gives only his multidimensional integrals as integrals over groups, rather than over the underlying Euclidean spaces, but there is a connection between Selberg's integral and his work. Multidimensional gamma and beta functions have also been considered in number theory. Some of this will be surveyed in Terras [38].

Macdonald's paper [26] contains many interesting conjectures. One of these can be extended in the following way.

$$(7.1) \quad \text{C.T.} \left(1 - \frac{x_1}{x_2}\right)^a \left(1 - \frac{x_2}{x_1}\right)^a \left(1 - \frac{x_1}{x_3}\right)^a \left(1 - \frac{x_3}{x_1}\right)^a \left(1 - \frac{x_2}{x_3}\right)^b \left(1 - \frac{x_3}{x_2}\right)^b \cdot \left(1 - \frac{x_1^2}{x_2 x_3}\right)^c \left(1 - \frac{x_2 x_3}{x_1^2}\right)^c = \frac{(2a + b + c)!(2a)!(2b)!(2c)!}{(a + b + c)!(a + b)!(a + c)!a!b!c!}.$$

When  $a = b = c$ , this is his conjecture (2.1) for  $B_2$ . The identity in (6.1) is true and can be proven by two uses of Dixon's well-poised sum (4.7) applied to the constant term in the expansion of the left-hand side. First adjacent terms are grouped

$$\left(1 - \frac{x_1}{x_2}\right)^a \left(1 - \frac{x_2}{x_1}\right)^a = \left(-\frac{x_2}{x_1}\right)^a \left(1 - \frac{x_1}{x_2}\right)^{2a},$$

the constant term is picked out, and the resulting double series is summed one series at a time. The details are easy and so will not be given here. Morris [29] has found a  $q$ -extension of (6.1), but this is as far as we have gotten in adding extra parameters to Macdonald's conjectures. Recall that Dyson added many extra parameters to get his conjecture (4.2). Initially he conjectured this when  $a_i \equiv k$ . As Macdonald has pointed out, Dyson's conjecture comes from the root systems of  $A_n$ . Until Selberg's integral came to light, it seemed impossible to prove results of this type without the extra degrees of freedom provided by Dyson's conjecture. This is no longer true. However, with more freedom comes more information, so it is still interesting to try to extend Macdonald's conjectures in this way.

Gindikin [15] says that the right way to extend hypergeometric functions to several variables is to use integral representations which generalize Euler's:

$$(7.2) \quad {}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix}; x\right) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 (1-xt)^{-a} t^{b-1} (1-t)^{c-b-1} dt,$$

rather than extensions of Barnes' integral representation:

$$(7.3) \quad {}_2F_1\left(\begin{matrix} a, b \\ c \end{matrix}; x\right) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{\Gamma(a+s)\Gamma(b+s)\Gamma(-s)}{\Gamma(c+s)} (-z)^s ds,$$

when  $|\arg(-z)| < \pi$  and the path of integration separates the poles of  $\Gamma(a+s)$  and  $\Gamma(b+s)$  from those of  $\Gamma(-s)$ . He may be right, and a test question is to see if Selberg's integral can be extended to Barnes type integrals. There is a  $q$ -extension of Mellin-Barnes integrals which can be considered. The most general extension of the beta

function integral I know is given in [8]. It uses the  $q$ -extension of the Mellin-Barnes integrals, and in the absolutely continuous case it is

$$\frac{1}{2\pi} \int_{-1}^1 \prod_{n=0}^{\infty} \frac{(1-2xq^n + q^{2n})(1-2xq^{n+(1/2)} + q^{2n+1})(1+2xq^n + q^{2n})(1+2xq^{n+(1/2)} + q^{2n+1})}{(1-2axq^n + a^2q^{2n})(1-2bxq^n + b^2q^{2n})(1+2cxq^n + c^2q^{2n})(1+2dxq^n + d^2q^{2n})} \cdot (1-x^2)^{-1/2} dx$$

$$(7.4) \quad = \frac{(abcd; q)_{\infty}}{(ab; q)_{\infty}(-ac; q)_{\infty}(-ad; q)_{\infty}(-bc; q)_{\infty}(-bd; q)_{\infty}(cd; q)_{\infty}(q; q)_{\infty}},$$

when  $\max(|a|, |b|, |c|, |d|, |q|) < 1$ .

This is a natural place to look for such an extension.

Another integral that seems similar to (1.2) is

$$(7.5) \quad \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-a^2 \sum_{i \neq j} \frac{1}{(t_i - t_j)^2} - \frac{1}{2} \sum_{i=1}^n t_i^2\right) dt_1 \cdots dt_n = e^{-a \binom{n}{2}}.$$

See Gallavotti and Marchioro [14] for a proof. It would be interesting to extend this.

**Acknowledgments.** I would like to thank F. J. Dyson for a translation of the proof in [35], I. G. Macdonald for extensive correspondence on these topics two years ago and for a copy of [26] and G. Andrews for a copy of [3] and for informing me about Selberg’s integral. Kevin Kadell was a patient listener who forced me to revise one of these conjectures.

*Note added in proof.* W. Morris observed that (7.1) and the  $q$ -extension of it are essentially contained in two papers of L. Carlitz, *Summation of some double series*, I, II, *Glasnik Matem.* 10, 30 (1979), pp. 73–81 and 11, 31 (1976), pp. 199–203.

There are other types of extensions of Selberg’s integral that directly relate to  $B_n$ ,  $C_n$ ,  $BC_n$  and  $D_n$ . These will appear in another paper.

REFERENCES

[1] G. E. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in [5], pp. 191–224.  
 [2] ———, *On  $q$ -analogues of the Watson and Whipple summations*, *SIAM J. Math. Anal.*, 7 (1976), 332–336.  
 [3] ———, *Notes on the Dyson conjecture*, *SIAM J. Math. Anal.*, to appear.  
 [4] G. E. ANDREWS AND R. ASKEY, *Another  $q$ -extension of the beta function*, to appear.  
 [5] R. ASKEY (editor), *Theory and Application of Special Functions*, Academic Press, New York, 1975.  
 [6] ———, *The  $q$ -gamma and  $q$ -beta functions*, *Applicable Analysis*, 8 (1978), pp. 125–141.  
 [7] ———, *Ramanujan’s extensions of the gamma and beta functions*, *Amer. Math. Monthly*, 87 (1980), pp. 346–359.  
 [8] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, in preparation.  
 [9] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, 1935, reprinted by Hafner, New York, 1964.  
 [10] F. A. BEREZIN AND F. I. KARPELEVIC, *Zonal spherical functions and Laplace operators on some symmetric spaces*, *Dokl. Akad. Nauk. SSSR (N.S.)*, 118 (1958), pp. 9–12. (In Russian.)  
 [11] R. P. BOAS, *Entire Functions*, Academic Press, New York, 1954.  
 [12] L. CARLITZ, *Some formulas of F. H. Jackson*, *Monat. für Math.*, 73 (1969), pp. 193–198.  
 [13] F. J. DYSON, *Statistical theory of the energy levels of complex systems: I*, *J. Math. Phys.*, 3 (1962), pp. 140–156.  
 [14] G. GALLAVOTTI AND C. MARCHIORO, *On the calculation of an integral*, *J. Math. Anal. Appl.*, 44 (1973), pp. 661–675.

- [15] S. G. GINDIKIN, *Analysis in homogeneous domains*, Uspehi Mat. Nauk., 19 (1965), #4, pp. 3–92 (in Russian); Russian Mathematical Surveys, 19 (1964), #4, pp. 1–90 (in English).
- [16] I. J. GOOD, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.
- [17] J. GUNSON, *Proof of a conjecture by Dyson in the statistical theory of energy levels*, J. Math. Phys., 3 (1962), pp. 752–753.
- [18] C. S. HERZ, *Bessel functions of matrix argument*, Ann. of Math., 61 (1955), pp. 474–523.
- [19] F. H. JACKSON, *Certain  $q$ -identities*, Quart. J. Math., 12 (1941), pp. 167–172.
- [20] A. T. JAMES, *Special functions of matrix and single argument in statistics*, in [5], pp. 497–520.
- [21] A. T. JAMES AND A. G. CONSTANTINE, *Generalized Jacobi polynomials as spherical functions of the Grassmann manifold*, Proc. London Math. Soc. (3), 29 (1974), pp. 174–192.
- [22] S. KARLIN AND L. S. SHAPLEY, *Geometry of Moment Spaces*, Memoirs of the American Mathematical Society 12, Providence, 1953.
- [23] S. KARLIN AND W. STUDDEN, *Tchebycheff Systems*, Interscience, New York, 1966.
- [24] T. KOORNWINDER, *Two-variable analogues of the classical orthogonal polynomials*, in [5], pp. 435–495.
- [25] H. MAASS, *Zur Theorie der Kugelfunktionen einer Matrixvariablen*, Math. Ann., 135 (1958), pp. 391–416.
- [26] I. G. MACDONALD, *Some conjectures for root systems and finite reflection groups*, to appear.
- [27] M. L. MEHTA, *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.
- [28] M. L. MEHTA AND F. J. DYSON, *Statistical theory of the energy levels of complex systems: V*, J. Math. Phys., 4 (1963), pp. 713–719.
- [29] W. MORRIS, Private communication.
- [30] C. E. PORTER, *Statistical Theories of Spectral Fluctuations*, Academic Press, New York, 1965.
- [31] S. RAMANUJAN, *Notebooks of Srinivasa Ramanujan*, Vol. II, Tata Institute of Fundamental Research, Bombay, 1957.
- [32] H. A. ROTHE, *Systematisches Lehrbuche der Arithmetik*, Leipzig, 1811.
- [33] I. J. SCHOENBERG, *An isoperimetric inequality for closed curves convex in even-dimensional Euclidean spaces*, Acta Math., 91 (1954), pp. 143–164.
- [34] A. SELBERG, *Über einen Satz von A. Gelfand*, Arch. Math. Naturvid., 44 (1941), pp. 159–170.
- [35] ———, *Bemerkninger om et Multipelt Integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71–78.
- [36] I. G. SPRINKHUIZEN-KUYPER, *Koornwinder Polynomials*, Thesis, University of Amsterdam, 1979.
- [37] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Coll. Publ. 23, fourth edition, American Mathematical Society, Providence, RI, 1975.
- [38] A. TERRAS, *Fourier Analysis on Symmetric Spaces and Applications to Number Theory*, to appear.
- [39] S. S. WILKS, *Mathematical Statistics*, Wiley, New York, 1962.
- [40] K. WILSON, *Proof of a conjecture of Dyson*, J. Math. Phys., 3 (1962), pp. 1040–1043.
- [41] J. WISHART, *The generalized product moment distribution in samples from a normal multivariate population*, Biometrika 20A, (1928), pp. 32–43.

## ESTIMATES FOR THE DERIVATIVES OF SOLUTIONS TO WEAKLY SINGULAR FREDHOLM INTEGRAL EQUATIONS\*

JUHANI PITKÄRANTA†

**Abstract.** The paper deals with the differential properties of a function  $\phi(x)$  defined over a bounded domain in  $R^m$ ,  $m \geq 1$ , as a solution to a weakly singular Fredholm integral equation of the second kind. Estimates near the boundary of the domain for the derivatives of  $\phi(x)$  are given, and an explicit local singular resolution is derived for  $\phi(x)$  near a smooth portion of the boundary, assuming that the kernel of the integral equation is of a specific form.

**1. Introduction.** Let  $\Omega$  be an open, bounded domain in  $R^m$ ,  $m \geq 1$ . We consider the differentiability properties of a function  $\phi(x)$  defined on  $\Omega$  as a solution to a Fredholm integral equation of the second kind,

$$(1.1) \quad \phi(x) - \int_{\Omega} K(x, y)\phi(y) dy = q(x), \quad x \in \Omega,$$

where the kernel is at most weakly singular; i.e.,  $|K(x, y)| \leq C|x - y|^{-\lambda}$ ,  $0 > \lambda > -m$ .

Weakly singular integral equations of the form (1.1) arise in many physical applications. We mention the integral equation formulations of various elliptic boundary value problems [5, pp. 137-272], [1], [3], and the particle transport problems of astrophysics and reactor theory [2].

If the kernel  $K(x, y)$  has a singularity at  $x = y$ , the solution of (1.1) is generally not a smooth function, even if  $q(x)$  is smooth. Instead, we find that the derivatives of  $\phi(x)$ , starting from a certain order, are unbounded on  $\partial\Omega$ . In the one-dimensional case this is known from previous studies; in fact, the nature of the singularities is known in detail for many special kernels, see [4], [6], [7], [8], [9].

In the present paper we study the behavior of the solution to (1.1) in a general multidimensional setting. In the first case we allow  $\Omega$  to an arbitrary open and bounded set. Assuming certain differentiability properties for  $K(x, y)$  and  $q(x)$ , we estimate the growth rates of the derivatives of  $\phi(x)$  to arbitrary order as  $x$  approaches the boundary  $\partial\Omega$ . Such estimates are especially relevant for the design of efficient numerical algorithms for the solution of (1.1): knowing the growth rates of the derivatives, one can design numerical algorithms with local approximation properties fitted so as to achieve any desired global convergence rate.

In the second case we make the more detailed assumptions that  $\partial\Omega$  contains an open subset of a smooth surface and that  $q(x)$  is a smooth function in the vicinity of this surface. We also assume some further properties for  $K(x, y)$ . We then derive a local singular resolution showing explicitly the behavior of  $\phi(x)$  near the smooth portion of the boundary. The singular resolution consists of a linear combination of known singularly behaved functions, with unknown smooth functions acting as coefficients. This may be regarded as a generalization of the known expansions in one dimension.

The notation and the main results of the paper are presented in § 2. Sections 3 and 4 are devoted to the proofs.

**2. Notation and the main results.** In what follows,  $\Omega$  denotes a bounded, open subset of  $R^m$ ,  $m \geq 1$ , with boundary  $\partial\Omega$ . If  $x \in \Omega$ , we let  $\rho(x) = \inf_{y \in \partial\Omega} |x - y|$ , where  $|\cdot|$  denotes the Euclidean norm. For a real-valued function defined on  $\Omega$ ,  $D^\alpha \phi$  denotes a

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

† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland. This work was supported by the Finnish National Research Council for Technical Sciences.

partial derivative of  $\phi(x)$ , with  $\alpha$  a multi-index,  $\alpha = (\alpha_1, \dots, \alpha_m)$ ,  $\alpha_i \geq 0$ . The order of the derivative is denoted by  $|\alpha|$ ,  $|\alpha| = \alpha_1 + \dots + \alpha_m$ . For a function  $\varphi(x, y)$  defined over  $\Omega \times \Omega$  we use the symbols  $D_x^\alpha$  and  $D_y^\alpha$  for the partial derivatives with respect to the Cartesian coordinates of  $x$  and  $y$ , respectively. As usual,  $L_p(\Omega)$ ,  $1 \leq p \leq \infty$ , denotes the space of real functions  $\varphi(x)$  such that  $\|\varphi\|_{L_p(\Omega)} = [\int_\Omega |\varphi|^p dx]^{1/p} < \infty$ . We let finally  $C$  denote a constant which may take different values in different usages and may depend on  $\Omega$  and on any parameters to be introduced, unless noted otherwise.

Let  $s \geq 0$  be a real number such that  $s = k + \sigma$ ,  $k = \text{integer}$ ,  $k \geq 0$ , and  $0 \leq \sigma < 1$ . Then as usual,  $C^s(\bar{\Omega})$  denotes the space of real functions  $\varphi(x)$  on  $\Omega$  such that  $\|\varphi\|_{C^s(\bar{\Omega})} < \infty$ , where

$$\|\varphi\|_{C^s(\bar{\Omega})} = \max_{|\alpha| \leq k} \sup_{x \in \Omega} |D^\alpha \varphi(x)| + \max_{\substack{|\alpha|=k \\ x, y \in \Omega \\ x \neq y}} \sup \frac{|(D^\alpha \varphi)(x) - (D^\alpha \varphi)(y)|}{|x - y|^\sigma}.$$

For irregular domains, we modify the definition of  $C^s(\bar{\Omega})$  as follows. First, if  $x \in \Omega$ , denote by  $B(x)$  the  $\rho(x)$ -neighborhood of  $x$ . Then we define  $\bar{C}^s(\Omega)$  to be the space of functions  $\varphi(x)$  such that

$$\|\varphi\|_{\bar{C}^s(\Omega)} = \sup_{x \in \Omega} \|\varphi|_{B(x)}\|_{C^s(\overline{B(x)})} < \infty.$$

The spaces  $C^s(\bar{\Omega})$  and  $\bar{C}^s(\Omega)$  coincide for integral  $s$ , and even for nonintegral  $s$  if  $\Omega$  is sufficiently regular. For example, it is possible to show that if  $\Omega$  is a simply connected Lipschitzian domain having a piecewise smooth boundary, then there is a constant  $C$  depending on  $\Omega$  such that  $\|\varphi\|_{C^s(\bar{\Omega})} \leq C \|\varphi\|_{\bar{C}^s(\Omega)}$  for all  $\varphi \in \bar{C}^s(\Omega)$ , i.e.,  $\bar{C}^s(\Omega) = C^s(\bar{\Omega})$ .

We also need some spaces with norms containing weight functions. Let  $k \geq 0$  be an integer and  $\mu \leq k$  a real number, and let  $\mu_j = \max\{0, \mu - k + j\}$ ,  $j = 0, \dots, k$ . Then we say  $\varphi \in \bar{C}^{k, \mu}(\Omega)$  if  $\|\varphi\|_{\bar{C}^{k, \mu}(\Omega)} < \infty$ , where

$$\|\varphi\|_{\bar{C}^{k, \mu}(\Omega)} = \max_{j=0, \dots, k} \left\{ \max_{|\alpha|=j} \left\{ \sup_{x \in \Omega} \rho(x)^{\mu_j} |D^\alpha \varphi(x)| \right\} \right\}.$$

We need finally some special classes of functions, which will be used as kernels in (1.1). Let  $\lambda$  be a real parameter, and let  $\mathcal{K}_{\lambda, \Omega}$  be the linear manifold of functions  $K(x, y)$  defined on the set  $\{(x, y) \in \Omega \times \Omega, x \neq y\}$  and such that

$$(i) \quad |D_x^\alpha K(x, y)| \leq C(1 + |x - y|^{\lambda - |\alpha|}), \quad x, y \in \Omega, x \neq y, |\alpha| \geq 0, C = C(K, \alpha) < \infty,$$

and

$$(ii) \quad \text{if } x \in \Omega \text{ and if } u, t \in \Omega, u \neq t, \text{ are such that } |u - x| < \rho(x), |t - x| < \rho(x), \text{ then}$$

$$K(u, t) = A_x(u, r, \theta), \quad r = |u - t|, \quad \theta = \frac{1}{r}(u - t),$$

where  $A_x$  is defined on  $R^m \times \{r \in R^1; r > 0\} \times R^m$  and satisfies

$$\left| D_u^\alpha D_\theta^\beta \left( \frac{\partial}{\partial r} \right)^l A_x(u, r, \theta) \right| \leq C(1 + r^{\lambda - l}), \quad u \in R^m, r > 0, \theta \in R^m, |\theta| = 1,$$

$$|\alpha|, |\beta|, l \geq 0, \quad C = C(K, \alpha, \beta, l) < \infty.$$

It is obvious that if  $\lambda \geq \mu$ , then  $\mathcal{H}_{\lambda,\Omega} \subset \mathcal{H}_{\mu,\Omega}$ . We also point out that if  $K(x, y) \in \mathcal{H}_{\lambda,\Omega}$ , then  $D_x^\alpha K(x, y) \in \mathcal{H}_{\lambda-|\alpha|,\Omega}$ .

If  $K \in \mathcal{H}_{\lambda,\Omega}$ ,  $\lambda > -m$ , we may define a bounded integral operator  $T : L_p(\Omega) \rightarrow L_p(\Omega)$ ,  $1 \leq p \leq \infty$ , through the formula

$$(2.1) \quad (T\phi)(x) = \int_{\Omega} K(x, t)\phi(t) dt.$$

We denote by  $\mathcal{T}_{\lambda,\Omega}$  the set of integral operators  $T$  defined by (2.1) with  $K \in \mathcal{H}_{\lambda,\Omega}$ ,  $\lambda > -m$ .

Now consider in  $L_1(\Omega)$  the integral equation

$$(2.2) \quad \phi(x) - (T\phi)(x) = q(x), \quad x \in \Omega,$$

where  $T \in \mathcal{T}_{\lambda,\Omega}$ ,  $\lambda > -m$ . We have:

**THEOREM 2.1.** *Let  $\phi \in L_1(\Omega)$  be a solution (not necessarily unique) to (2.2), where  $T \in \mathcal{T}_{\lambda,\Omega}$ ,  $\lambda > -m$ ,  $\lambda \neq \text{integer}$ , and let  $\phi_n = \phi - \sum_{k=0}^{n-1} T^k q$ ,  $n \geq 1$ . Then we have the inequalities*

$$\begin{aligned} \|\phi_n\|_{\bar{C}^{m+\lambda}(\Omega)} &\leq C \|\phi\|_{L_1(\Omega)}, & \text{if } \lambda + (n-2)(m+\lambda) > 0, \\ \|\phi_n\|_{\bar{C}^{k,k-m-\lambda}(\Omega)} &\leq C \|\phi\|_{L_1(\Omega)}, & \text{if } \lambda + (n-1)(m+\lambda) > k, k > m+\lambda, \\ \|\phi\|_{\bar{C}^{m+\lambda}(\Omega)} &\leq C [\|\phi\|_{L_1(\Omega)} + \|q\|_{\bar{C}^{m+\lambda}(\Omega)}], & q \in \bar{C}^{m+\lambda}(\Omega), \\ \|\phi\|_{\bar{C}^{k,k-m-\lambda}(\Omega)} &\leq C [\|\phi\|_{L_1(\Omega)} + \|q\|_{\bar{C}^{k,k-m-\lambda}(\Omega)}], & q \in \bar{C}^{k,k-m-\lambda}(\Omega), \\ & & k > m+\lambda. \end{aligned}$$

We make next some more detailed assumptions on  $T$ ,  $\Omega$  and  $q$ . First, we assume that the kernel of the operator  $T$  in (2.2) has the more specific form

$$(2.3) \quad K(x, y) = r^\lambda A_1(x, r, \theta) + A_2(x, r, \theta), \quad r = |x - y|, \quad \theta = \frac{1}{r}(x - y),$$

where  $A_1, A_2 \in \bar{C}^\infty(\mathbb{R}^m \times \mathbb{R}^1 \times \mathbb{R}^m)$  and  $\lambda > -m$  ( $\lambda$  may also be integral). Obviously then  $T \in \mathcal{T}_{\lambda,\Omega}$ . We assume further that  $\partial\Omega$  contains an open subset of a smooth surface, with  $\Omega$  locally on one side of the surface. Then there exists a coordinate system  $\{x_1, \dots, x_m\}$  and a sphere  $B \subset \mathbb{R}^m$  with center at  $z \in \partial\Omega$  and with radius  $\rho(B) > 0$ , such that

$$B \cap \Omega = \{x \in B; x_m > \psi(x_1, \dots, x_{m-1})\},$$

where  $\psi \in C^\infty(\mathbb{R}^{m-1})$ . We then assume that  $q$  in (2.2) satisfies  $q|_{B \cap \Omega} \in C^\infty(\overline{B \cap \Omega})$ .

Let  $B_1$  be another sphere also centered at  $z$  and with radius  $\rho(B_1) < \rho(B)$ . Then we have the following result for the behavior of the solution to (2.2) in  $B_1 \cap \Omega$ :

**THEOREM 2.2.** *Let  $\phi \in L_1(\Omega)$  be a solution to (2.2) under the above assumptions on  $T$ ,  $\Omega$  and  $q$ . Then if  $x \in B_1 \cap \Omega$ , we have, for arbitrary  $k \geq 0$ , the representation*

$$(2.3) \quad \phi(x) = \sum_{i=1}^{i_0} \sum_{j=0}^{j_0(i)} g(x)^{i(m+\lambda)} [\log g(x)]^j c_{ijk}(x) + \zeta_k(x),$$

where  $c_{ijk}, \zeta_k \in C^k(\overline{B_1 \cap \Omega})$ ,  $i_0(m+\lambda) > k$ ,  $j_0(i)$  is the number of integers in the set  $\{\nu(m+\lambda); \nu = 1, 2, \dots, i\}$ , and  $g(x) = x_m - \psi(x_1, \dots, x_{m-1})$ .

It could be proven (see [8] for the one-dimensional case) that under the assumptions made, the expansion (2.3) is the most that one can say about the smoothness of  $\phi(x)$  in general, i.e., the coefficients  $c_{ijk}(x)$  are generally nonzero at the boundary. In

particular this shows that if  $\lambda \neq$  integer and  $q \in C^\infty(\bar{\Omega})$ , the inclusions  $\phi \in \bar{C}^{m+\lambda}(\Omega)$  and  $\phi \in \bar{C}^{k, k-m-\lambda}(\Omega)$ ,  $k > m + \lambda$ , as given by Theorem 2.1, cannot be improved.

With appropriate changes of variables, the results of Theorems 2.1 and 2.2 could be formulated also in the case where (1.1) is defined over the boundary  $\partial\Omega$  of a bounded domain  $\Omega \subset R^{m+1}$ . As an example of this type of problem, consider the integral equation arising from the use of single and double layer potential representations in solving the Laplace equation in  $\Omega$  [5]. This integral equation is of weakly singular type, provided that  $\partial\Omega$  is regular to a degree  $C^s$  with  $s > 1$ . For example, if  $\partial\Omega$  consists of smooth sections that are joined together so that the unit normal vector to  $\partial\Omega$  is continuous everywhere on  $\partial\Omega$ , then  $\partial\Omega$  is of class  $C^s$  with  $s < 2$ , and Theorems 2.1 and 2.2 could be applied to determine the behavior of the solution near the irregular points of  $\partial\Omega$ . If instead  $\partial\Omega$  contains corners, then the integral equation is of singular type, and the above results do not apply.

We finally point out that in the cases where (1.1) is defined over a closed surface, the kernel is most often such that the solution is smooth, except at points where either  $\partial\Omega$  or  $q$  is not infinitely smooth.

**3. Proof of Theorem 2.1.** We establish first some boundedness results for weakly singular integral operators.

**THEOREM 3.1.** *Let  $T \in \mathcal{T}_{\lambda, \Omega}$ ,  $\lambda > -m$ ,  $\lambda \neq$  integer. Then  $T$  is a bounded map  $T: L_\infty(\Omega) \rightarrow \bar{C}^{m+\lambda}(\Omega)$ .*

*Proof.* Let  $f \in L_\infty(\Omega)$  be given and let  $\phi(x) = (Tf)(x) = \int_\Omega K(x, t)f(t) dt$ . We let  $\nu$  be an integer such that  $m + \lambda - \nu = \sigma \in (0, 1)$ . Then if  $\alpha$  is a multi-index such that  $|\alpha| \leq \nu$ , we have

$$\begin{aligned} |D^\alpha \phi(x)| &\leq \int_\Omega |D_x^\alpha K(x, t)| |f(t)| dt \\ (3.1) \qquad &\leq C \int_\Omega |x-t|^{-m+\sigma} \|f\|_{L_\infty(\Omega)} dt \\ &\leq C_1 \|f\|_{L_\infty(\Omega)}, \end{aligned}$$

so  $T$  is a bounded map from  $L_\infty(\Omega)$  to  $\bar{C}^\nu(\Omega)$ .

Now let  $\alpha$  be such that  $|\alpha| = \nu$ , and let  $K^\alpha(x, y) = D_x^\alpha K(x, y)$ . Then  $K^\alpha \in \mathcal{H}_{-m+\sigma, \Omega}$ . Let  $x, y \in \Omega$  be such that  $|x-y| < \rho(x)$ , and denote by  $B$  the  $|x-y|$ -neighborhood of  $x$ . We write

$$\begin{aligned} D^\alpha \phi(u) &= \int_B K^\alpha(u, t)f(t) dt + \int_{\Omega \setminus B} K^\alpha(u, t)f(t) dt \\ &= \psi_1(u) + \psi_2(u), \quad u \in \bar{B}, \end{aligned}$$

and we estimate  $|\psi_i(x) - \psi_i(y)|$ ,  $i = 1, 2$ . First, we easily conclude that

$$\begin{aligned} |\psi_1(x) - \psi_1(y)| &\leq C \|f\|_{L_\infty(\Omega)} \int_B (|x-t|^{-m+\sigma} + |y-t|^{-m+\sigma}) dt \\ &\leq C_1 \|f\|_{L_\infty(\Omega)} \int_0^{2|x-y|} r^{-1+\sigma} dr \\ &\leq C_2 |x-y|^\sigma \|f\|_{L_\infty(\Omega)}. \end{aligned}$$

To estimate  $|\psi_2(x) - \psi_2(y)|$ , note first that if  $u \in B$  and  $\beta$  is a multi-index such that



$|\beta| = 1$ , we have

$$\begin{aligned} |D^\beta \psi_2(u)| &\leq C \|f\|_{L_\infty(\Omega)} \int_{\Omega \setminus B} |u-t|^{-m+\sigma-1} dt \\ &\leq C_1 \delta(u)^{\sigma-1} \|f\|_{L_\infty(\Omega)}, \end{aligned}$$

where  $\delta(u) = \inf_{v \in \partial B} |u-v|$ . Using this and observing that for  $u = u(s) = sx + (1-s)y$ ,  $s \in (0, 1)$ ,  $\delta(u(s)) > s|x-y|$ , we get

$$\begin{aligned} |\psi_2(x) - \psi_2(y)| &= \left| \int_0^1 \frac{d}{ds} \psi_2(sx + (1-s)y) ds \right| \\ &\leq C \|f\|_{L_\infty(\Omega)} |x-y| \int_0^1 [|x-y|s]^{\sigma-1} ds \\ &\leq C_1 |x-y|^\sigma \|f\|_{L_\infty(\Omega)}. \end{aligned}$$

Combining the above inequalities we have, for  $|\alpha| = \nu$  and for  $x, y \in \Omega$  such that  $|x-y| < \rho(x)$ ,

$$|(D^\alpha Tf)(x) - (D^\alpha Tf)(y)| \leq C |x-y|^\sigma \|f\|_{L_\infty(\Omega)}.$$

Together with (3.1), this completes the proof.  $\square$

**THEOREM 3.2.** *Let  $T \in \mathcal{T}_{\lambda, \Omega}$ ,  $\lambda = \nu - m + \sigma$ ,  $\nu = \text{integer}$ ,  $\nu \geq 0$ ,  $\sigma \in (0, 1)$ . Then if  $k \geq 1$ ,  $T$  is a bounded map  $T: \bar{C}^{k, k-\sigma}(\Omega) \rightarrow \bar{C}^{k+\nu, k-\sigma}(\Omega)$ .*

*Proof.* Let  $f \in \bar{C}^{k, k-\sigma}(\Omega)$ ,  $k \geq 1$ . By Theorem 3.1:

$$(3.2) \quad \|D^\alpha Tf\|_{L_\infty(\Omega)} \leq C \|f\|_{L_\infty(\Omega)}, \quad |\alpha| \leq \nu.$$

To estimate the higher derivatives of  $Tf$ , take any  $\alpha$  such that  $|\alpha| = \nu$  and write

$$\begin{aligned} (D^\alpha Tf)(u) &= \int_B K^\alpha(u, t) f(t) dt + \int_{\Omega \setminus B} K^\alpha(u, t) f(t) dt \\ &= \psi_1(u) + \psi_2(u), \quad u \in B, \end{aligned}$$

where  $K^\alpha(u, t) = D_u^\alpha K(u, t)$  and  $B$  is the  $\frac{1}{2}\rho(x)$ -neighborhood of  $x$ . Take  $\beta$  to be another multi-index such that  $1 \leq |\beta| \leq k$ . Then, using repeatedly the identity  $(\partial/\partial u_i) K^\alpha(u, t) = -(\partial/\partial t_i) K^\alpha(u, t) + (\partial/\partial u_i + \partial/\partial t_i) K^\alpha(u, t)$ , and integrating by parts, we obtain

$$(3.3) \quad (D^\beta \psi_1)(x) = \sum D_x^{\beta'} \int_{\partial B} K_{\beta''}^\alpha(x, t) \frac{x_i - t_i}{|x-t|} D_t^{\beta''} f(t) dt + \sum \int_B K_{\gamma'}^\alpha(x, t) D_t^{\gamma''} f(t) dt;$$

here the sums are over finite sets of indices and multi-indices satisfying  $|\beta'| + |\beta''| + |\beta'''| = |\beta| - 1$  and  $|\gamma'| + |\gamma''| = |\beta|$ , and we have used the abbreviation

$$(3.4) \quad K_\beta^\alpha(x, t) = \left(\frac{\partial}{\partial x_1} + \frac{\partial}{\partial t_1}\right)^{\beta_1} \cdots \left(\frac{\partial}{\partial x_m} + \frac{\partial}{\partial t_m}\right)^{\beta_m} K^\alpha(x, t), \quad t \in B.$$

By our assumptions, we have  $K^\alpha \in \kappa_{-m+\sigma, \Omega}$ . Hence, if  $u, t \in B$ , we may write

$$(3.5) \quad K^\alpha(u, t) = A(u, r, \theta), \quad r = |u-t|, \quad \theta = \frac{1}{r}(u-t),$$

where  $A$  satisfies

$$|D_u^\beta A(u, r, \theta)| \leq Cr^{-m+\sigma}, \quad |\beta| \geq 0, \quad C = C(\beta) < \infty.$$

Now, since (3.4) and (3.5) imply that  $K_\beta^\alpha(x, t) = D_u^\beta A(u, r, \theta)|_{u=x}$  for  $t \in B$ , we have

$$(3.6) \quad |D_x^\gamma K_\beta^\alpha(x, t)| \leq C |x - t|^{-m+\sigma-|\gamma|}, \quad t \in B.$$

Using (3.6) and the inequalities

$$\begin{aligned} \frac{1}{2}\rho(x) &\leq \rho(t) \leq 2\rho(x), & t \in \bar{B}, \\ |D^\gamma f(t)| &\leq \rho(t)^{\sigma-|\gamma|} \|f\|_{\bar{C}^{k,k-\sigma}(\Omega)}, & t \in \bar{B}, \quad 0 < |\gamma| \leq k, \\ |\beta'| + |\beta''| &\leq |\beta| - 1, & |\gamma''| \leq |\beta|, \end{aligned}$$

in (3.3), we get after some computation:

$$|D^\beta \psi_1(x)| \leq C \rho(x)^{\sigma-|\beta|} \|f\|_{\bar{C}^{k,k-\sigma}(\Omega)}.$$

Noting finally that

$$\begin{aligned} |D^\beta \psi_2(x)| &\leq C \int_{C \setminus B} |x - t|^{-m+\sigma-|\beta|} |f(t)| dt \\ &\leq C_1 \rho(x)^{\sigma-|\beta|} \|f\|_{L^\infty(\Omega)}, \end{aligned}$$

we have proved:

$$\begin{aligned} |(D^\beta D^\alpha T f)(x)| &\leq C \rho(x)^{\sigma-|\beta|} \|f\|_{\bar{C}^{k,k-\sigma}(\Omega)}, \\ |\alpha| = \nu, \quad 0 &\leq |\beta| \leq k, \quad x \in \Omega. \end{aligned}$$

Together with (3.2), this completes the proof.  $\square$

Suppose next that we are given the integral operators  $T_i, i = 1, \dots, n$ , with  $T_i \in \mathcal{T}_{\lambda_i}, \lambda_i > -m$ . Then if  $K_i$  is the kernel of  $T_i$ , the compounded operator  $\prod_{i=1}^n T_i$  is an integral operator with the kernel  $M_n(x, y)$  defined recursively as

$$(3.7) \quad \begin{aligned} M_i(x, y) &= \int_\Omega K_i(x, t) M_{i-1}(t, y) dt, \quad i \geq 2, \\ M_1 &= K_1. \end{aligned}$$

Our next theorem, fundamental to the subsequent proofs, is concerned with the differential properties of  $M_n(x, y)$ .

**THEOREM 3.3.** *Let  $n \geq 2$  be an integer, and let  $\lambda_i, i = 1, \dots, n$  be real numbers,  $\lambda_i > -m$ . Further let  $K_i \in \mathcal{K}_{\lambda_i, \Omega}$  be given, and let  $M_n$  be defined by (3.7). Then if  $x, y \in \Omega$ , we have the estimates*

$$(3.8) \quad |D_x^\alpha M_n(x, y)| \leq \begin{cases} C(1 + |x - y|^{\mu_n - |\alpha| - \varepsilon}) & \text{if } m + \lambda_n - |\alpha| > 0 \\ C[(1 + |\log \rho(x)|)(1 + |x - y|^{\mu_n - 1 - \varepsilon}) + |x - y|^{\mu_n - |\alpha| - \varepsilon}] & \text{if } m + \lambda_n - |\alpha| = 0 \\ C[\rho(x)^{m + \lambda_n - |\alpha|} (1 + |x - y|^{\mu_n - 1 - \varepsilon}) + |x - y|^{\mu_n - |\alpha| - \varepsilon}] & \text{if } m + \lambda_n - |\alpha| < 0, \end{cases}$$

where  $\mu_1 = \lambda_1, \mu_i = \mu_{i-1} + (m + \lambda_i), i \geq 2, \varepsilon$  is an arbitrary positive number, and  $C$  is a constant depending on  $M_n, |\alpha|, \Omega$  and  $\varepsilon$ .

*Remark.* For  $\alpha = 0$ , this result is well known, see [5, Thm. 1, p. 59].

*Proof.* Assume first that  $m + \lambda_n - |\alpha| > 0$ . Then since the assertion is true for  $\alpha = 0$  (see the above remark), we get

$$|D_x^\alpha M_n(x, y)| \leq C \int_\Omega (1 + |x - t|^{\lambda_n - |\alpha|})(1 + |y - t|^{\mu_n - 1 - \varepsilon}) dt.$$

Since  $\lambda_n - |\alpha| > -m$  and  $\mu_{n-1} - \varepsilon > -m$  for  $\varepsilon$  small enough, we may apply the estimates for compounded weakly singular integrals [5, p. 59] to obtain

$$\begin{aligned} |D_x^\alpha M_n(x, y)| &\leq C(1 + |x - y|^{m + \lambda_n - |\alpha| + \mu_{n-1} - \varepsilon}) \\ &= C(1 + |x - y|^{\mu_n - |\alpha| - \varepsilon}), \end{aligned}$$

which was to be proved.

For  $m + \lambda_n - |\alpha| \leq 0$  we consider two cases separately. First, let  $x, y \in \Omega$  be such that  $\rho(x) \leq 4|x - y|$ . We make the induction hypothesis that either  $n = 2$ , or  $n \geq 3$  and (3.8) holds with  $n$  replaced by  $n - 1$ .

Let  $B_i = \{u \in R^m; |u - x| < \rho_i\}$ ,  $i = 1, \dots, 4$ , where  $\rho_1 = \frac{1}{8}\rho(x)$ ,  $\rho_2 = \frac{1}{2}|x - y|$ ,  $\rho_3 = 2|x - y|$ , and  $\rho_4 = \text{diam}(\Omega)$ . We write

$$\begin{aligned} M_n(u, y) &= \int_{B_1} K_n(u, t)M_{n-1}(t, y) dt \\ &\quad + \sum_{i=2}^4 \int_{\Omega \cap B_i \setminus B_{i-1}} K_n(u, t)M_{n-1}(t, y) dt \\ &= \sum_{i=1}^4 \psi_i(u, y), \quad u \in B_1, \end{aligned}$$

and we estimate  $|D_u^\alpha \psi_i(u, y)|_{u=x}$  for each  $i$ .

For  $i = 1$  we choose the multi-indices  $\gamma$  and  $\beta$  so that  $\gamma + \beta = \alpha$  and  $-m < \lambda_n - |\gamma| \leq 0$ , and we integrate by parts to obtain

$$\begin{aligned} (3.9) \quad D_u^\alpha \psi_1(u, y)|_{u=x} &= \sum D_x^{\gamma + \alpha'} \int_{\partial B_1} K_{\alpha''}(x, t) \frac{t_i - x_i}{|x - t|} D_t^{\alpha'''} M_{n-1}(t, y) ds \\ &\quad + \sum \int_{B_1} D_x^\gamma K_{\beta'}(x, t) D_t^{\beta''} M_{n-1}(t, y) dt, \end{aligned}$$

where

$$K_\alpha(x, t) = \left(\frac{\partial}{\partial x_1} + \frac{\partial}{\partial t_1}\right)^{\alpha_1} \dots \left(\frac{\partial}{\partial x_m} + \frac{\partial}{\partial t_m}\right)^{\alpha_m} K_n(x, t), \quad t \in B_1,$$

and the (finite) sums are such that  $|\gamma| + |\alpha'| + |\alpha''| + |\alpha'''| = |\alpha| - 1$  and  $|\gamma| + |\beta'| + |\beta''| = |\alpha|$ . By the reasoning familiar from the proof of Theorem 3.2, we have

$$|D_x^\gamma K_\alpha(x, t)| \leq C|x - t|^{\lambda_n - |\gamma|}, \quad x, t \in B_1, \quad \lambda_n - |\gamma| \leq 0.$$

Also, by the induction hypothesis and since  $\frac{7}{8}\rho(x) \leq \rho(t) \leq \frac{9}{8}\rho(x)$  and  $\frac{1}{2}|x - y| \leq |t - y| \leq \frac{3}{2}|x - y|$  for  $t \in \bar{B}_1$ , we have

$$|D_t^\alpha M_{n-1}(t, y)| \leq \begin{cases} C(1 + \rho(x)^{m + \lambda_{n-1} - |\alpha| - \varepsilon}) + \rho(x)^{m + \lambda_{n-1} - |\alpha| - \varepsilon} |x - y|^{\mu_{n-2} - \varepsilon} \\ \quad + |x - y|^{\mu_{n-1} - |\alpha| - \varepsilon}, & n \geq 3 \\ C(1 + |x - y|^{\mu_{n-1} - |\alpha|}), & n = 2, \quad t \in B_1, \quad \varepsilon > 0. \end{cases}$$

Take  $\varepsilon$  small enough so that  $m + \lambda_{n-1} - \varepsilon > 0$ . Then  $\rho(x)^{m + \lambda_{n-1} - \varepsilon} \leq C|x - y|^{m + \lambda_{n-1} - \varepsilon}$ , and we have

$$|D_t^\alpha M_{n-1}(t, y)| \leq C\rho(x)^{-|\alpha|}(1 + |x - y|^{\mu_{n-1} - 2\varepsilon}), \quad t \in B_1, \quad n \geq 2.$$

Apply this and the inequalities

$$|\gamma| + |\alpha'| + |\alpha''| \leq |\alpha| - 1, \quad |\gamma| + |\beta''| \leq |\alpha|,$$

in (3.9) to finally obtain

$$|D_u^\alpha \psi_1(u, y)|_{|u=x}| \leq C \rho(x)^{m+\lambda_n-|\alpha|} (1+|x-y|^{\mu_{n-1}-2\epsilon}).$$

For  $i = 2$  we use the inequality

$$\begin{aligned} M_{n-1}(t, y) &\leq C(1+|t-y|^{\mu_{n-1}-\epsilon}) \\ &\leq C_1(1+|x-y|^{\mu_{n-1}-\epsilon}), \quad t \in B_2, \end{aligned}$$

to obtain

$$\begin{aligned} |D_u^\alpha \psi_2(u, y)|_{|u=x}| &\leq C(1+|x-y|^{\mu_{n-1}-\epsilon}) \int_{B_2 \setminus B_1} |D_x^\alpha K_n(x, t)| dt \\ &\leq C_1(1+|x-y|^{\mu_{n-1}-\epsilon}) \int_{B_2 \setminus B_1} |x-t|^{\lambda_n-|\alpha|} dt \\ &\leq C_2(1+|x-y|^{\mu_{n-1}-\epsilon}) \int_{\rho(x)/2}^{|x-y|/2} r^{m-1+\lambda_n-|\alpha|} dr \\ &\leq \begin{cases} C_3(1+|\log \rho(x)|)(1+|x-y|^{\mu_{n-1}-\epsilon}) & \text{if } m+\lambda_n-|\alpha|=0, \\ C_3 \rho(x)^{m+\lambda_n-|\alpha|} (1+|x-y|^{\mu_{n-1}-\epsilon}) & \text{if } m+\lambda_n-|\alpha|<0. \end{cases} \end{aligned}$$

Similarly, using the inequality

$$|D_x^\alpha K_n(x, t)| \leq C|x-t|^{\lambda_n-|\alpha|} \leq C_1|x-y|^{\lambda_n-|\alpha|}, \quad t \in B_3 \setminus B_2,$$

we get

$$\begin{aligned} |D_u^\alpha \psi_3(u, y)|_{|u=x}| &\leq C|x-y|^{\lambda_n-|\alpha|} \int_{B_3 \setminus B_2} |M_{n-1}(t, y)| dt \\ &\leq C_1|x-y|^{\lambda_n-|\alpha|} \int_{B_3 \setminus B_2} (1+|t-y|^{\mu_{n-1}-\epsilon}) dt \\ &\leq C_2|x-y|^{\lambda_n-|\alpha|} \int_0^{3|x-y|} r^{m-1}(1+r^{\mu_{n-1}-\epsilon}) dr. \end{aligned}$$

Since  $|x-y|^{m+\lambda_n-|\alpha|} \leq C \rho(x)^{m+\lambda_n-|\alpha|}$  and  $1+|\log|x-y|| \leq C(1+|\log \rho(x)|)$ , we obtain the same estimate as for  $i = 2$ .

Finally, if  $i = 4$  we have

$$\begin{aligned} |D_u^\alpha \psi_4(u, y)|_{|u=x}| &\leq C \int_{B_4 \setminus B_3} |x-t|^{\lambda_n-|\alpha|} (1+|y-t|^{\mu_{n-1}-\epsilon}) dt \\ &\leq C_1 \int_{B_4 \setminus B_3} (|x-t|^{\lambda_n-|\alpha|} + |x-t|^{\lambda_n+\mu_{n-1}-|\alpha|-\epsilon}) dt \\ &\leq \begin{cases} C_2(1+|\log \rho(x)|+|x-y|^{\mu_n-|\alpha|-\epsilon'}), & m+\lambda_n-|\alpha|=0, \quad \epsilon'>\epsilon, \\ C_2(\rho(x)^{m+\lambda_n-|\alpha|}+|x-y|^{\mu_n-|\alpha|-\epsilon'}), & m+\lambda_n-|\alpha|<0, \quad \epsilon'>\epsilon. \end{cases} \end{aligned}$$

Upon combining the estimates for  $|D_u^\alpha \psi_i(u, y)|_{|u=x}|$ ,  $i = 1, \dots, 4$ , we obtain (3.8).

It remains to consider the case  $\rho(x) > 4|x-y|$ . We will show that for each  $x \in \Omega$ ,  $M_n$  can be split as

$$(3.10) \quad M_n(u, y) = U_{n,x}(u, y) + V_{n,x}(u, y), \quad x \in \Omega, \quad y \in \Omega, \quad |u-x| < \rho(x),$$

where  $U_{n,x}(u, y)$  is defined for all  $u, y \in R^m$ ,  $u \neq y$ , and satisfies

$$(3.11) \quad |D_u^\alpha U_{n,x}(u, y)|_{|u=x}| \leq C(1+|x-y|^{\mu_n-|\alpha|-\epsilon}), \quad x, y \in \Omega, \quad x \neq y, \quad \epsilon > 0,$$

and  $V_{n,x}$  satisfies

$$(3.12) \quad \begin{aligned} & |D_u^\alpha V_{n,x}(u, y)|_{u=x} \\ & \leq \begin{cases} C(1 + |\log \rho(x)| + \rho(x)^{\mu_n - |\alpha| - \epsilon}), & m + \lambda_n - |\alpha| = 0 \\ C(\rho(x)^{m + \lambda_n - |\alpha|} + \rho(x)^{\mu_n - |\alpha| - \epsilon}), & m + \lambda_n - |\alpha| < 0, \end{cases} \quad x, y \in \Omega, \quad \epsilon > 0. \end{aligned}$$

It is clear that if (3.10) through (3.12) hold for  $\rho(x) > 4|x - y|$ , then (3.8) holds as well.

We prove (3.10) through (3.12) by induction. First, we recall from our assumptions that we have

$$(3.13) \quad K_n(u, y) = P_{n,x}(u, y) + Q_{n,x}(u, y), \quad x, y \in \Omega, \quad |u - x| < \rho(x), \quad y \neq u,$$

where  $P_{n,x}(u, y)$  is defined for all  $u, y \in R^m$ ,  $u \neq y$ , and satisfies

$$(3.14) \quad |D_u^\alpha P_{n,x}(u, y)| \leq C(1 + |u - y|^{\lambda_n - |\alpha|}),$$

and  $Q_{n,x}$  satisfies

$$(3.15) \quad \begin{aligned} & Q_{n,x}(u, y) = 0, \quad \text{if } |u - x| < \rho(x), \quad |y - x| < \rho(x), \\ & |D_u^\alpha Q_{n,x}(u, y)|_{u=x} \leq C(1 + |x - y|^{\lambda_n - |\alpha|}) \quad \text{if } y \in \Omega, \quad |y - x| \geq \rho(x). \end{aligned}$$

In view of (3.13) through (3.15), the asserted splitting is true for  $n = 1$ .

Next assume that (3.10) through (3.12) hold when  $n$  is replaced by  $n - 1$ . Let  $B_i = \{u \in R^m; |u - x| < \rho_i\}$ ,  $i = 1, 2$ , where  $\rho_1 = \frac{1}{2}\rho(x)$  and  $\rho_2 = \text{diam}(\Omega) + C$ ,  $C = \text{const.}$ ,  $C > 0$ . Let  $e(u)$  be a smooth function defined on  $R^m$  such that  $e(u) = 1$  for  $u \in \Omega$  and  $e(u) = 0$  for  $u \in R^m \setminus B_2$ . We claim that we can define

$$U_{n,x}(u, y) = e(u) \int_{R^m} P_{n,x}(u, t)e(t)U_{n-1,x}(t, y) dt, \quad u, y \in R^m, \quad u \neq y.$$

Let us first check that  $V_{n,x} = M_n - U_{n,x}$  satisfies (3.12). We have

$$\begin{aligned} V_{n,x}(u, y) &= \int_{\Omega \setminus B_1} K_n(u, t)M_{n-1}(t, y) dt + \int_{R^m \setminus B_1} P_{n,x}(u, t)e(t)U_{n-1,x}(t, y) dt \\ &\quad + \int_{B_1} P_{n,x}(u, t)V_{n-1,x}(t, y) dt. \end{aligned}$$

Here the first two terms satisfy the desired inequalities, as can be verified by direct differentiation. For the third term we obtain the same estimates from a partial integration formula similar to (3.9) and from the induction hypotheses. We omit the details.

It remains to check that  $U_{n,x}(u, y)$  satisfies (3.11). We start from the formula

$$D_u^\alpha U_{n,x}(u, y) = \sum [D_u^{\beta'} e(u)] D_u^{\alpha'} \int_{R^m} P_{\alpha''}(u, t) [D_t^{\beta''} e(t)] D_t^{\alpha'''} U_{n-1,x}(t, y) dt,$$

where  $P_\alpha(u, t) = (\partial/\partial u_1 + \partial/\partial t_1)^{\alpha_1} \cdots (\partial/\partial u_m + \partial/\partial t_m)^{\alpha_m} P_{n,x}(u, t)$ , and the sum is over a finite set of multi-indices such that  $\alpha' + \alpha'' + \alpha''' + \beta' + \beta'' = \alpha$  and either  $\alpha' = 0$ ,  $\mu_{n-1} - |\alpha''| > 0$  or  $-m < \mu_{n-1} - |\alpha''| \leq 0$ . If  $\mu_n - |\alpha''| > 0$  (and  $\alpha' = 0$ ) or if  $m + \lambda_n - |\alpha'| > 0$ , we

have

$$\begin{aligned}
 | [D_u^{\beta'} e(u)] D_u^{\alpha'} \int_{R^m} P_{\alpha''}(u, t) [D_t^{\beta''} e(t)] D_t^{\alpha''} U_{n-1,x}(t, y) dt_{|u=x} | \\
 \leq C \int_{B_2} (1 + |x - t|^{\lambda_n - |\alpha'|}) (1 + |t - y|^{\mu_{n-1} - |\alpha''| - \varepsilon}) dt \\
 \leq C_1 (1 + |x - y|^{\mu_n - |\alpha| - \varepsilon}), \quad \varepsilon > 0.
 \end{aligned}$$

In the remaining cases, i.e.,  $m + \lambda_n - |\alpha'| \leq 0$  and  $-m < \mu_{n-1} - |\alpha''| \leq 0$ , we write

$$\begin{aligned}
 \int_{R^m} K_{\alpha''}(u, t) [D_t^{\beta''} e(t)] D_t^{\alpha''} U_{n-1,x}(t, y) dt = \int_{B_0} \{ \cdot \cdot \cdot \} dt + \int_{R^m \setminus B_0} \{ \cdot \cdot \cdot \} dt \\
 = \psi_1(u, y) + \psi_2(u, y),
 \end{aligned}$$

where  $B_0$  is the  $\frac{1}{2}|x - y|$ -neighborhood of  $x$ . Upon integrating by parts in the first term, we can now verify that

$$|D_u^{\alpha'} \psi_i(u, y)|_{|u=x} \leq C |x - y|^{\mu_n - |\alpha| - \varepsilon}, \quad i = 1, 2, \quad \varepsilon > 0.$$

Hence,  $U_{n,x}(u, y)$  satisfies (3.11), and the proof of Theorem 3.3 is complete.  $\square$

*Remark.* For fixed  $M_n, \alpha$ , and  $\varepsilon$ , the constant  $C$  in (3.8) depends only on  $\text{diam}(\Omega)$ . Moreover,  $C$  remains bounded as  $\text{diam}(\Omega) \rightarrow 0$ .

As a consequence of Theorem 3.3, we obtain the following.

**COROLLARY 3.1.** *Let  $T \in \mathcal{T}_{\lambda, \Omega}$ ,  $\lambda > -m$ ,  $\lambda \neq \text{integer}$ . Then if  $\lambda + (n - 2)(m + \lambda) > 0$ ,  $T^n$  is a bounded map  $T^n : L_1(\Omega) \rightarrow \bar{C}^{m+\lambda}(\Omega)$ . If  $\lambda + (n - 1)(m + \lambda) > k$ ,  $k = \text{integer}$ ,  $k > m + \lambda$ , then  $T^n$  is a bounded map  $T^n : L_1(\Omega) \rightarrow \bar{C}^{k, k-m-\lambda}(\Omega)$ .*

*Proof.* Suppose  $\lambda + (n - 2)(m + \lambda) > 0$ . Then if  $M_n$  is the kernel of  $T^n$ ,  $M_n$  satisfies (3.8) with  $\mu_n = \lambda + (n - 1)(m + \lambda) > m + \lambda$ . Hence,  $D_x^\alpha M_n(x, y)$  is bounded for  $|\alpha| < m + \lambda$ , and so  $T^n$  is a bounded map  $T^n : L_1(\Omega) \rightarrow \bar{C}^k(\Omega)$ ,  $k < m + \lambda$ ,  $k = \text{integer}$ . Now let  $x_1, x_2 \in \Omega$  be such that  $|x_1 - x_2| < \rho(x_1)$ , let  $\alpha$  be a multi-index such that  $\sigma = m + \lambda - |\alpha| \in (0, 1)$ , and let  $M_n^\alpha(x, y) = D_x^\alpha M_n(x, y)$ . Then, writing

$$M_n^\alpha(x_2, y) - M_n^\alpha(x_1, y) = \int_0^1 \left[ \frac{d}{ds} M_n^\alpha(sx_1 + (1 - s)x_2, y) \right] ds$$

and applying Theorem 3.3, we have

$$\begin{aligned}
 |M_n^\alpha(x_1, y) - M_n^\alpha(x_2, y)| \\
 \leq C |x_1 - x_2| \int_0^1 \{ \rho(sx_1 + (1 - s)x_2)^{\sigma-1} + |sx_1 + (1 - s)x_2 - y|^{\sigma-1} \} ds.
 \end{aligned}$$

Since  $\rho(sx_1 + (1 - s)x_2) \geq s|x_1 - x_2|$  and  $|sx_1 + (1 - s)x_2 - y| \geq |s - s'| |x_1 - x_2|$  for some  $s' \in [0, 1]$ , we get

$$|M_n^\alpha(x_1, y) - M_n^\alpha(x_2, y)| \leq C |x_1 - x_2|^\sigma,$$

where  $C$  is independent of  $x_1, x_2, y$ . It follows that  $T^n$  is a bounded map  $T^n : L_1(\Omega) \rightarrow \bar{C}^{m+\lambda}(\Omega)$ .

The second part of the assertion can be proved in a similar fashion. We omit the details.  $\square$

Using the above results, the proof of Theorem 2.1 is straightforward: if  $\phi$  is a solution to (2.2), then  $\phi$  also satisfies

$$\phi(x) - (T^n \phi)(x) = \sum_{k=0}^{n-1} (T^k q)(x), \quad x \in \Omega, n \geq 1.$$

The asserted estimates then follow from Corollary 3.1, Theorem 3.1, and Theorem 3.2.  $\square$

**4. Proof of Theorem 2.2.** We assume the notation of § 2. Our first step is to carry out a coordinate transformation so as to map  $\partial\Omega$  locally onto a plane. To this end, we define the mapping  $F : R^m \rightarrow R^m$  as

$$(4.1) \quad \begin{aligned} F(x) = z : \quad z_i = x_i, \quad i = 1, \dots, m-1, \\ z_m = x_m - \psi(x_1, \dots, x_{m-1}), \end{aligned}$$

where  $\psi$  is as in Theorem 2.2. (Here and henceforth we disregard the minor notational modifications required in the case  $m = 1$ , i.e., we assume  $m \geq 2$ .) Since  $\psi \in C^\infty(R^{m-1})$ , (4.1) defines a smooth invertible transformation of  $R^m$  onto itself. Moreover,  $F(B \cap \partial\Omega)$  is a section of the plane  $x_m = 0$ . We set  $\varphi(x) = \phi(F^{-1}(x))$ ,  $f(x) = q(F^{-1}(x))$ , and  $L(x, t) = K(F^{-1}(x), F^{-1}(t))J(t)$ , where  $x, t \in \Omega_F = F(\Omega)$ ,  $K$  is the kernel of operator  $T$  in (2.2), and  $J(t)$  denotes the Jacobian of the transformation  $F^{-1}$ . In this notation, (2.2) may be rewritten in the new coordinates as

$$(4.2) \quad \varphi(x) - (\tau\varphi)(x) = f(x), \quad x \in \Omega_F,$$

where

$$(\tau\varphi)(x) = \int_{\Omega_F} L(x, t)\varphi(t) dt.$$

It may be verified from the definition of  $F$  that

$$F^{-1}(x) - F^{-1}(y) = r\Gamma(x, r, \theta), \quad r = |x - y|, \quad \theta = \frac{1}{r}(x - y),$$

where  $\Gamma$  is a smooth vector valued function and  $|\Gamma|$  is strictly positive. Hence, with  $K$  of the form (2.3), we have

$$\begin{aligned} L(x, y) &= |x - y|^\lambda \tilde{A}_1(x, r, \theta) + \tilde{A}_2(x, r, \theta), \\ r &= |x - y|, \quad \theta = \frac{1}{r}(x - y), \quad \tilde{A}_i \in C^\infty(R^m \times R^1 \times R^m). \end{aligned}$$

We introduce some notation which will be required in the proof. With  $B$  and  $B_1$  as in Theorem 2.2, let  $B', B'',$  and  $B'''$  be spheres such that  $B_1 \in B''' \in B'' \in B' \in B$ , and let  $G = F(B \cap \Omega)$ ,  $G_1 = F(B_1 \cap \Omega)$ ,  $G' = F(B' \cap \Omega)$ ,  $G'' = F(B'' \cap \Omega)$ ,  $G''' = F(B''' \cap \Omega)$ , and  $D' = F(B' \cap \partial\Omega)$ . We let  $e(x)$  be a smooth function defined on  $R^m$  such that  $e(x) = 1$  for  $x \in F(B''')$  and  $e(x) = 0$  for  $x \in R^m \setminus F(B')$ , and we define the integral operator  $\tau' \in \mathcal{T}_{\lambda, G'}$  as

$$(\tau'\varphi)(x) = \int_{G'} L(x, t) e(t)\varphi(t) dt,$$

where  $L$  is the kernel of  $\tau$ . Then, setting  $\varphi_0 = \varphi|_{G'}$ , we may rewrite (4.2) for  $x \in G'$  as

$$(4.3) \quad \varphi_0(x) - (\tau'\varphi_0)(x) = f_0(x), \quad x \in G',$$

where

$$\begin{aligned} f_0(x) &= f(x) + (\tau - \tau')\varphi(x) \\ &= f(x) + \int_{\Omega_F \setminus G''} L(x, t)(1 - e(t))\varphi(t) dt. \end{aligned}$$

Since  $f|_{G'} \in C^\infty(\bar{G})$ , it follows that  $f_0|_{G''} \in C^\infty(\bar{G}''')$ .

We state two partial integration formulae to be needed in the sequel. To this end let  $\varphi$  be defined on  $G'$  and sufficiently smooth. Then, recalling that  $D' = \partial G' \cap \partial \Omega_F$  is a section of the plane  $x_m = 0$ , and that  $e(t)$  together with all its derivatives vanishes on  $\partial G' \setminus D'$ , we have

$$(4.4) \quad \frac{\partial}{\partial x_i} (\tau' \varphi)(x) = (\tau'_i \varphi)(x) + \left( \tau' \frac{\partial \varphi}{\partial x_i} \right)(x), \quad i = 1, \dots, m-1,$$

$$(4.5) \quad \frac{\partial}{\partial x_m} (\tau' \varphi)(x) = (\tau'_m \varphi)(x) + \left( \tau' \frac{\partial \varphi}{\partial x_m} \right)(x) + \int_{D'} L(x, t) e(t) \varphi(t) dv,$$

where  $\tau'_i$  is an integral operator with the kernel

$$L_i(x, t) = \left( \frac{\partial}{\partial x_i} + \frac{\partial}{\partial t_i} \right) L(x, t) e(t).$$

Assuming the above notation, we will first prove:

LEMMA 4.1. *If  $\alpha$  is any multi-index such that  $\alpha_m = 0$ , then  $D^\alpha \varphi_0 \in C^0(\overline{G'})$  and  $(\partial/\partial x_m) D^\alpha \varphi_0 \in L_1(G')$ .*

*Proof.* Let  $\alpha$  be given such that  $\alpha_m = 0$ ,  $|\alpha| = k + 1$ , and let  $n$  be an integer such that  $\lambda + (n-1)(m+\lambda) > 1$ . Then (4.3) implies:

$$(4.6) \quad D^\alpha \varphi_0(x) = D^\alpha (\tau')^{nk} \varphi_0(x) + \sum_{j=0}^{nk-1} D^\alpha (\tau')^j f_0(x).$$

Applying repeatedly (4.4), the first term on the right side can be rewritten as

$$(4.7) \quad D^\alpha (\tau')^{nk} \varphi_0(x) = \sum_i \prod_{l=1}^n (\tau_{i,1,l}) \prod_{j=2}^{k-1} \left[ D^{\alpha^{ij}} \prod_{l=1}^n (\tau_{ijl}) \right] \varphi_0(x),$$

where  $|\alpha^{ij}| \leq 1$  and  $\tau_{ijl} \in \mathcal{T}_{\lambda, G'}$ . Now since  $\lambda + (n-1)(m+\lambda) > 1$ , it follows from Theorem 3.3 that if  $\varphi \in L_1(G')$  then  $\prod_{l=1}^n (\tau_{ijl}) \in \overline{C^0(G')} = C^0(\overline{G'})$ , and

$$\left| D^\alpha \prod_{l=1}^n (\tau_{ijl}) \varphi(x) \right| \leq C \rho(x)^{m+\lambda-1-\varepsilon} \|\varphi\|_{L_1(G')}, \quad x \in G', \quad |\alpha| = 1, \quad \varepsilon > 0,$$

with  $\rho(x) = \inf_{y \in \partial G'} |x - y|$ . Since  $G'$  has a piecewise smooth boundary, we have  $\rho(x)^{m+\lambda-1-\varepsilon} \in L_1(G')$ , provided that  $\varepsilon$  is small enough so that  $m + \lambda - 1 - \varepsilon > -1$ . Applying these arguments repeatedly in (4.7) we conclude:

$$(4.8) \quad D^\alpha (\tau')^{nk} \varphi_0 \in C^0(\overline{G'}), \quad \frac{\partial}{\partial x_m} D^\alpha (\tau')^{nk} \varphi_0 \in L_1(G').$$

To estimate the remaining terms on the right side of (4.6), let  $B_2, B_3$  be spheres such that  $B_2 \subseteq B_3 \subseteq B'''$ , and let  $G_i = F(B_i \cap \Omega)$ ,  $G_i \neq \emptyset$ . Further, let  $d(x)$  be a smooth function on  $R^m$  such that  $d(x) = 1$  for  $x \in F(B_3)$  and  $d(x) = 0$  for  $x \in R^m \setminus F(B''')$ . Recalling that  $f_0|_{G'''} \in C^\infty(\overline{G'''})$ , we have  $df_0 \in C^\infty(\overline{G'})$ . Then, applying (4.4) and Theorem 3.1, we conclude that  $D^\alpha \tau' df_0 \in C^0(\overline{G'})$ , and further, using (4.5), that  $(\partial/\partial x_m) D^\alpha \tau' df_0 \in L_1(G')$ , whenever  $\alpha_m = 0$ . Also, since  $1 - d(x) = 0$  for  $x \in G_3$ , we have  $\tau'(1-d)f_0|_{G_2} \in C^\infty(\overline{G_2})$ . So, if  $\alpha_m = 0$ , we have  $D^\alpha \tau' f_0|_{G_2} \in C^0(\overline{G_2})$  and  $(\partial/\partial x_m) D^\alpha \tau' f_0|_{G_2} \in L_1(G_2)$ .

Proceeding by induction it follows that for arbitrary  $B_2 \subseteq B'''$ ,  $G_2 = F(B_2 \cap \Omega)$ , and for any  $j, j \geq 1$ , we have

$$(4.9) \quad D^\alpha (\tau')^j f_0|_{G_2} \in C^0(\overline{G_2}), \quad \frac{\partial}{\partial x_m} D^\alpha (\tau')^j f_0|_{G_2} \in L_1(G_2), \quad \alpha_m = 0.$$



Furthermore, since  $B'$ ,  $B''$ , and  $B'''$  can be chosen arbitrarily, it follows that (4.9) holds for any  $G_2 = F(B_2 \cap \Omega)$  such that  $B_2 \Subset B$ . With this understanding, the assertion follows by combining (4.6), (4.8) and (4.9).  $\square$

Our next step is to prove the following:

LEMMA 4.2. *For any multi-index  $\alpha$  with  $\alpha_m > 0$ ,  $\varphi_0$  satisfies*

$$(4.10) \quad D^\alpha \varphi_0(x) = \Lambda(x) + \zeta(x), \quad x \in G_1,$$

where  $\zeta \in C^0(\overline{G_1})$  and  $\Lambda$  consists of a finite sum of functions of the type  $(\partial/\partial x_m)^l \prod_{i=1}^{i_0} \tau_i \xi(x)$ , where  $0 \leq l \leq \alpha_m - 1$ ,  $i_0 = i_0(\alpha_m) < \infty$ , and  $\tau_i$  and  $\xi$  are defined as

$$(\tau_i \varphi)(x) = \int_{G'} L_i(x, t) \varphi(t) dt,$$

$$\xi(x) = \int_{D'} L_{i_0+1}(x, (v, 0)) \eta(v) dv,$$

where  $\eta \in C^\infty(\overline{D'})$  and  $L_i$  admits the representation

$$L_i(x, t) = r^\lambda A_{i,1}(x, r, \theta) + A_{i,2}(x, r, \theta), \quad r = |x - t|, \quad \theta = \frac{1}{r}(x - t), \quad x, t \in \mathbb{R}^m,$$

with  $A_{ij} \in C^\infty(\mathbb{R}^m \times \mathbb{R}^1 \times \mathbb{R}^m)$ .

*Proof.* We use the formula (4.6), with  $n$  and  $k$  chosen so that  $\lambda + (n - 1)(m + \lambda) > 1$  and  $k = \alpha_m$ . First, writing  $D^\alpha = (\partial/\partial x_m)^k D^\beta$  and using (4.4), we find that

$$D^\alpha (\tau')^{nk} \varphi_0(x) = \sum_i \left( \frac{\partial}{\partial x_m} \right)^k \prod_{j=1}^{nk} (\tau_{ij}) D^{\alpha^i} \varphi_0(x),$$

where  $\alpha_m^i = 0$ ,  $|\alpha^i| \leq |\alpha|$ , and the kernels of operators  $\tau_{ij}$  are of the form

$$L_{ij}(x, y) = \left( \frac{\partial}{\partial x_1} + \frac{\partial}{\partial y_1} \right)^{\gamma_1} \cdots \left( \frac{\partial}{\partial x_{m-1}} + \frac{\partial}{\partial y_{m-1}} \right)^{\gamma_{m-1}} L(x, y) e(y),$$

where  $0 \leq \gamma_\nu \leq \alpha_\nu$ . By Lemma 4.1,  $D^{\alpha^i} \varphi_0 \in C^0(\overline{G'})$  and  $(\partial/\partial x_m) D^{\alpha^i} \varphi_0 \in L_1(G')$ . Using then (4.5), we get the further representation

$$(4.11) \quad D^\alpha (\tau')^{nk} \varphi_0(x) = \sum_{\nu=0}^{k-1} \left( \frac{\partial}{\partial x_m} \right)^\nu \sum_i \prod_{j=1}^{n_i} (\tau_{ij}) \xi_i(x)$$

$$+ \sum_i \prod_{j=1}^k \left[ \prod_{l=1}^n (\tau_{ijl}) \frac{\partial}{\partial x_m} \right] D^{\alpha^i} \varphi_0(x),$$

where the kernels of  $\tau_{ijl}$  are of the form

$$L_{ijl}(x, y) = \left( \frac{\partial}{\partial x_1} + \frac{\partial}{\partial y_1} \right)^{\gamma_1} \cdots \left( \frac{\partial}{\partial x_m} + \frac{\partial}{\partial y_m} \right)^{\gamma_m} L(x, y) e(y),$$

with  $0 \leq \gamma_i \leq \alpha_i$ , and the  $\xi_i$  are functions of the type

$$\xi_i(x) = \int_{D'} L_{ij}(x, (v, 0)) \eta(v) dv, \quad \eta(v) = \eta_0(v, 0),$$

$$(4.12) \quad \eta_0(x) = \prod_{j=1}^{n_1} (\tau_{ij}) \prod_{j=1}^{k_1} \left[ \prod_{l=1}^n (\tau_{ijl}) \frac{\partial}{\partial x_m} \right] D^{\alpha^i} \varphi_0(x),$$

$$x \in \overline{G'}, \quad n_1 \geq 0, \quad k_1 \geq 0, \quad n_1 + nk_1 = nk.$$

We claim that if  $\eta$  is as in (4.12), then  $\eta \in C^\infty(\overline{D'})$ . Indeed, by Lemma 4.1,  $D^\alpha(\partial/\partial x_m)D^{\alpha'}\varphi_0(x) = (\partial/\partial x_m)D^{\alpha'+\alpha}\varphi_0(x) \in L_1(G')$  for any  $\alpha$  such that  $\alpha_m = 0$ . So, applying (4.4) and Theorem 3.3, we have that  $D^\alpha \prod_{i=1}^n (\tau_{ij}) (\partial/\partial x_m)D^{\alpha'}\varphi_0(x) \in C^0(G')$ , and further by induction that  $D^\alpha \eta_0(x) \in C^0(\overline{G'})$  for all  $\alpha$  with  $\alpha_m = 0$ . Hence,  $\eta \in C^\infty(\overline{D'})$ .

We know from Lemma 4.1 and from Theorem 3.3 that the second term on the right side of (4.11) is in  $C^0(\overline{G'})$ . So, we have obtained the representation

$$D^\alpha(\tau')^{nk}\varphi_0(x) = \Lambda(x) + \zeta(x),$$

where  $\Lambda$  and  $\zeta$  satisfy the assertions of Lemma 4.2.

To handle the remaining terms in (4.6), let  $G_i = F(B_i \cap \Omega)$ ,  $i = 1, 2, 3$ , where  $B_i$  are spheres such that  $B_1 \subseteq B_2 \subseteq B_3 \subseteq B'$ , and write  $f_0 = f_1 + f_2$ , where  $f_1 \in C^\infty(\overline{G'})$  and  $f_2(x) = 0$  for  $x \in G_3$ . Then  $\tau'f_2|_{G_2} \in C^\infty(\overline{G_2})$  and  $D^\alpha \tau'f_1 \in C^0(\overline{G'})$  for any  $\alpha$  with  $\alpha_m = 0$ . Using (4.4) and (4.5) one then obtains the asserted splitting for  $D^\alpha \tau'f_1$ ,  $\alpha_m > 0$ , and hence for  $D^\alpha \tau'f_0$  as well. Proceeding by induction, the same is obtained for  $D^\alpha(\tau')^j f_0$ ,  $j > 1$ . The proof is complete.  $\square$

We will finally analyze the behavior of  $\Lambda(x)$  in (4.10) in more detail.

LEMMA 4.3. *Let  $\eta \in C^\infty(\overline{D'})$  and let  $L(x, t) = r^\lambda A_1(x, r, \theta) + A_2(x, r, \theta)$ ,  $x, t \in R^m$ ,  $r = |x - t|$ ,  $\theta = (1/r)(x - t)$ ,  $A_i \in C^\infty(R^m \times R^1 \times R^m)$ ,  $\lambda > -m$ . Then if  $k \geq 0$ , we have the representation*

$$\int_{D'} L(x, (v, 0))\eta(v) dv = \sum_{j=0}^{j_0} c_{jk}(x)x_m^{m+\lambda-1} (\log x_m)^j + \zeta_k(x), \quad x \in G'',$$

where  $c_{jk} \in C^\infty(\overline{G''})$ ,  $\zeta_k \in C^k(\overline{G''})$ ,  $j_0 = 1$  if  $\lambda = \text{integer}$  and  $j_0 = 0$  otherwise.

*Proof.* Upon expanding  $L(x, t)$  as

$$\begin{aligned} L(x, t) &= r^\lambda A_1(x, r, \theta) + A_2(x, r, \theta) \\ &= \sum_{i=0}^n \frac{1}{i!} \left[ \left( \frac{\partial}{\partial r} \right)^i A_1(x, r, \theta) \Big|_{r=0} \right] r^{\lambda+i} + L_n(x, t), \end{aligned}$$

we have  $L_n \in \mathcal{H}_{\lambda+n, G'}$ , and so  $\zeta_n(x) = \int_{D'} L_n(x, (v, 0))\eta(v) dv \in C^k(\overline{G'})$ , with  $k \rightarrow \infty$  as  $n \rightarrow \infty$ . Therefore, it suffices to consider the behavior of functions of the type

$$\xi_i(x) = \int_{D'} r^{\lambda+i} L_i(x, \theta)\eta(v) dv, \quad r = r(x, v), \quad \theta = \theta(x, v), \quad L_i \in C^\infty(R^m \times R^m).$$

Let  $x = (u, x_m)$ ,  $u \in R^{m-1}$ , and choose  $\rho_0$  so that the  $\rho_0$ -neighborhood  $D_{u, \rho_0}$  of  $u$ ,  $D_{u, \rho_0} = \{v \in R^{m-1}; |v - u| < \rho_0\}$ , is contained in  $D'$  for all  $u$  such that  $(u, 0) \in \overline{G''} \cap \partial\Omega_F$ . Then if

$$\xi_{i,0}(x) = \int_{D_{u, \rho_0}} r^{\lambda+i} L_i(x, \theta)\eta(v) dv,$$

we have  $\xi_i - \xi_{i,0} \in C^\infty(\overline{G''})$ , so it suffices to study the functions  $\xi_{i,0}(x)$ . Setting  $v = u + \rho\omega$ ,  $\rho = |v - u|$ ,  $|\omega| = 1$ , and expanding  $\eta(v)$  into a Taylor series at  $v = u$ , we get

$$\xi_{i,0}(x) = \int_0^{\rho_0} \int_{D_0} r^{\lambda+i} L_i(x, \theta) \left[ \sum_{j=0}^n \rho^j c_j(u, \omega) + \zeta_n(u, \rho, \omega) \right] \rho^{m-2} d\rho d\omega,$$

where  $D_0$  is the unit sphere in  $R^{m-1}$ ,  $c_j \in C^\infty(R^{m-1} \times R^{m-1})$ ,  $r = (x_m^2 + \rho^2)^{1/2}$ ,  $\theta_i = (1 - \theta_m^2)^{1/2} \omega_i$ ,  $i = 1, \dots, m-1$ ,  $\theta_m = x_m(x_m^2 + \rho^2)^{-1/2}$ , and  $\zeta_n$  satisfies

$$(4.13) \quad |D_u^\alpha \zeta_n(u, \rho, \omega)| \leq C\rho^{n-|\alpha|}, \quad 0 \leq |\alpha| \leq n.$$

Since  $\rho \leq r$ , it follows from (4.13) that if

$$\zeta_n(x) = \int_0^{\rho_0} \int_{D_0} r^{\lambda+i} L_i(x, \theta) \zeta_n(u, \rho, \omega) \rho^{m-2} d\rho d\omega,$$

then  $\zeta_n \in C^k(\overline{G_1})$  with  $k \rightarrow \infty$  as  $n \rightarrow \infty$ . So, we have reduced the problem to the study of functions of the form

$$(4.14) \quad \xi_{ij}(x) = \int_0^{\rho_0} \int_{D_0} r^{\lambda+i} \rho^{m-2+j} A(x, \theta, \omega) d\rho d\omega, \quad i, j \geq 0,$$

where  $A \in C^\infty(\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^{m-1})$ .

Upon performing the integration over  $\omega$ , (4.14) can be rewritten as

$$\xi_{ij}(x) = \int_0^{\rho_0} r^{\lambda+i} \rho^{m-2+j} A(x, \theta_m, (1-\theta_m^2)^{1/2}) d\rho, \quad A \in C^\infty(\mathbb{R}^m \times \mathbb{R}^1 \times \mathbb{R}^1).$$

Further, using the relations  $r = x_m/\theta_m$ ,  $\rho = (1-\theta_m^2)^{1/2} x_m/\theta_m$ , and changing the integration variable, we obtain

$$\begin{aligned} \xi_{ij}(x) &= x_m^{m+\lambda+i+j-1} \int_{a(x_m)}^1 \theta_m^{-m-\lambda-i-j} (1-\theta_m^2)^{(m+j-3)/2} A(x, \theta_m, (1-\theta_m^2)^{1/2}) d\theta \\ &= x_m^{m+\lambda+i+j-1} \int_{a(x_m)}^{1/2} \theta_m^{-m-\lambda-i-j} A_1(x, \theta_m) d\theta_m + x_m^{m+\lambda+i+j-1} \zeta(x), \end{aligned}$$

where  $a(x_m) = x_m(\rho_0^2 + x_m^2)^{-1/2}$ ,  $A_1 \in C^\infty(\mathbb{R}^m \times \mathbb{R}^1)$  and  $\zeta \in C^\infty(\overline{G''})$ . Now choose  $n$  so that  $-m-\lambda-i-j+n > -1$ , and expand  $A_1(x, \theta_m)$  into a Taylor series in  $\theta_m$  to obtain

$$(4.15) \quad \begin{aligned} \xi_{ij}(x) &= x_m^{m+\lambda+i+j-1} \left\{ \zeta(x) + \int_0^{1/2} \theta_m^{-m-\lambda-i-j} \Delta_n(x, \theta_m) d\theta_m \right. \\ &\quad \left. + \sum_{l=0}^{n-1} \frac{1}{l!} \left( \frac{\partial}{\partial \theta_m} \right)^l A_1(x, \theta_m) \Big|_{\theta_m=0} \int_{a(x_m)}^{1/2} \theta_m^{-m-\lambda-i-j+l} d\theta_m \right\} \\ &\quad - x_m^{m+\lambda+i+j-1} \int_0^{a(x_m)} \theta_m^{-m-\lambda-i-j} \Delta_n(x, \theta_m) d\theta_m, \end{aligned}$$

where  $\Delta_n(x, \theta_m)$  satisfies

$$(4.16) \quad \left| D_x^\alpha \left( \frac{\partial}{\partial \theta_m} \right)^l \Delta_n(x, \theta_m) \right| \leq C |\theta_m|^{n-l}, \quad C = C(\alpha, n, l) < \infty.$$

We can further simplify (4.15) as

$$(4.17) \quad \begin{aligned} \xi_{ij}(x) &= - \int_0^{b(x_m)} s^{-m-\lambda-i-j} \Delta_n(x, sx_m) ds \\ &\quad + \begin{cases} x_m^{m+\lambda+i+j-1} \log(x_m) \psi_1(x) + \psi_2(x), & \lambda = \text{integer,} \\ x_m^{m+\lambda+i+j-1} \psi_1(x) + \psi_2(x), & \lambda \neq \text{integer,} \end{cases} \end{aligned}$$

where  $b(x_m) = (\rho_0^2 + x_m^2)^{-1/2}$  and  $\psi_1, \psi_2 \in C^\infty(\overline{G''})$ .

In view of (4.16), the first term in (4.17) can be embedded in  $\psi_2(x)$ . So, recalling that

$$\int_{D'} L(x, (v, 0)) \eta(v) dv = \sum_{i=0}^n \sum_{j=0}^n \xi_{ij}(x) + \zeta_n(x),$$

where  $\zeta_n \in C^k(\overline{G''})$  with  $k \rightarrow \infty$  as  $n \rightarrow \infty$ , the proof is complete.  $\square$

LEMMA 4.4. Let  $s > -1$ ,  $\nu = \text{integer}$ ,  $\nu \geq 0$ , and let  $\varphi \in C^k(\overline{G''})$ . Then there exists an integer  $k_0 = k_0(k)$ , with  $k_0 \rightarrow \infty$  as  $k \rightarrow \infty$ , such that if  $0 \leq l < k_0$  and if  $L(x, t)$  satisfies the assumptions of Lemma 4.3, then

$$\int_{G''} L(x, t) t_m^s (\log(t_m))^\nu \varphi(t) dt = \sum_{j=0}^{i_0} c_{jl}(x) x_m^{m+\lambda+s} (\log(x_m))^j + \zeta_l(x), \quad x \in G''$$

where  $C_{jl}, \zeta_l \in C^l(\overline{G''})$  and

$$j_0 = \begin{cases} \nu + 1, & \text{if } \lambda = \text{integer or } m + \lambda + s = \text{integer,} \\ \nu, & \text{otherwise.} \end{cases}$$

*Proof.* We have

$$\int_{G''} L(x, t) t_m^s (\log(t_m))^\nu \varphi(t) dt = \int_0^a N(x, t_m) t_m^s (\log(t_m))^\nu dt_m,$$

where  $a = \sup \{t_m; t \in G''\}$  and

$$(4.18) \quad N(x, t_m) = \int_{D_t} L(x, (v, t_m)) \varphi((v, t_m)) dv,$$

with  $D_t = \{v \in R^{m-1}; (v, t_m) \in G''\}$ . Now since  $\varphi((v, t_m)) \in C^k(\overline{D_t})$  for each  $t \in G''$  in (4.18), we find, essentially by repeating the arguments in the proof of Lemma 4.3, that if  $l < k_0$ ,  $k_0 = k_0(k) \rightarrow \infty$  as  $k \rightarrow \infty$ , then

$$N(x, t_m) = \begin{cases} |x_m - t_m|^{m+\lambda-1} \log|x_m - t_m| \psi_1(x) + \psi_2(x), & \lambda = \text{integer,} \\ |x_m - t_m|^{m+\lambda-1} \psi_1(x) + \psi_2(x), & \lambda \neq \text{integer,} \end{cases}$$

with  $\psi_1, \psi_2 \in C^l(\overline{G''})$ . The remaining part of the proof is an exercise of integral calculus; the details can be found in [7], [8].  $\square$

We are now ready to prove Theorem 2.2. We apply first Lemma 4.3 and Lemma 4.4 together with an induction argument to obtain for the function  $\Lambda(x)$  in (4.10) the representation

$$\Lambda(x) = \sum_{l=0}^{\alpha_m-1} \sum_{i=1}^{i_0} \sum_{j=0}^{j_0(i)} \left( \frac{\partial}{\partial x_m} \right)^l x_m^{(m+\lambda)i-1} (\log(x_m))^j c_{ijk}(x) + \eta_k(x), \quad x \in G_1, \quad k \geq 0,$$

where  $i_0$  and  $j_0(i)$  are as in Theorem 2.2 and  $c_{ijk}, \eta_k \in C^k(\overline{G_1})$ . From this and (4.10) it is easily verified by integration that for some  $d_{ijk} \in C^k(\overline{G_1})$ ,

$$\varphi_0(x) - \sum_{i=1}^{i_0} \sum_{j=0}^{j_0(i)} x_m^{(m+\lambda)i} (\log(x_m))^j d_{ijk}(x) \in C^k(\overline{G_1}).$$

Upon changing the variables back to the original ones, the proof of Theorem 2.2 is complete.  $\square$

#### REFERENCES

- [1] K. ATKINSON, *The numerical solution of Fredholm integral equations of the second kind with singular kernels*, Numer. Math., 19 (1972), pp. 248–259.
- [2] G. I. BELL AND S. GLASSTONE, *Nuclear Reactor Theory*, Van Nostrand-Reinhold, New York, 1971.
- [3] R. GONZÁLEZ AND R. KRESS, *On the treatment of a Dirichlet-Neumann mixed boundary value problem by an integral equation method*, this Journal, 8 (1977), pp. 504–517.
- [4] H. G. KAPER AND R. B. KELLOGG, *Asymptotic behavior of the solution of the integral transport equation in slab geometry*, SIAM J. Appl. Math., 32 (1977), pp. 191–200.
- [5] S. G. MIKHLIN, *Integral Equations*, Pergamon Press, London, 1964.

- [6] J. PITKÄRANTA, *On the differential properties of solutions to Fredholm equations with weakly singular kernels*, J. Inst. Math. Appl., 24 (1979), pp. 109–119.
- [7] ———, *Asymptotic behavior of the solution to the integral transport equation in the vicinity of a curved material interface*, SIAM J. Appl. Math., 36 (1979), pp. 200–218.
- [8] G. R. RICHTER, *On weakly singular Fredholm integral equations with displacement kernels*, J. Math. Anal. Appl., 55 (1976), pp. 32–42.
- [9] C. SCHNEIDER, *Regularity of the solution to a class of weakly singular Fredholm integral equations of the second kind*, Integral Equations Operator Theory, 2 (1979), pp. 62–68.

## ON FREQUENCY DOMAIN STABILITY FOR EVOLUTION EQUATIONS IN HILBERT SPACES VIA THE ALGEBRAIC RICCATI EQUATION\*

D. WEXLER†

**Abstract.** We establish Lyapunov type stability for an evolution equation in a Hilbert space by using some energy functions which are obtained via the algebraic Riccati equation. We discuss also briefly the advantages and limitations of this infinite-dimensional extension of a well-known method. Our abstract setting is motivated by some special systems arising in reactor dynamics and retarded differential difference equations.

**1. Introduction.** We discuss Lyapunov type stability for the zero solution of the differential system

$$(1.1) \quad \frac{dx}{dt} = Ax + \phi(\sigma)b, \quad \frac{d\sigma}{dt} = \langle c, x \rangle - \phi(\sigma)\rho,$$

where the linear operator  $A$  generates a  $C_0$ - (i.e., quasi-bounded) semigroup  $S$  on the real Hilbert space  $X$  with inner product  $\langle \cdot, \cdot \rangle$  and norm  $|\cdot|$ ,  $b, c \in X$ ,  $\rho \in \mathbb{R}$ , and  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  is a nonlinear, locally Lipschitz function with  $r\phi(r) > 0$  for all  $r \neq 0$  (so that  $\phi(0) = 0$ ). In addition to this, we assume  $S$  to be exponentially stable, which means that there exist  $M \geq 1$  and  $\alpha > 0$  such that

$$(1.2) \quad |S(t)|_{\mathcal{L}(X)} \leq M e^{-\alpha t}, \quad \text{for all } t \geq 0,$$

where  $\mathcal{L}(X)$  denotes the Banach space of bounded linear operators from  $X$  to  $X$ ; for the theory of linear  $C_0$ -semigroups, we refer the reader to [11, Chap. IX]. The above system will be viewed in the Hilbert space  $\mathcal{X} = X \times \mathbb{R}$  with inner product

$$\langle (x_1, r_1), (x_2, r_2) \rangle_{\mathcal{X}} = \langle x_1, x_2 \rangle + r_1 r_2.$$

System (1.1) is an abstract version for some significant special control systems, among which we would like to mention first the integrodifferential system

$$(1.3) \quad \begin{aligned} \frac{\partial}{\partial t} T(t, \xi) &= \frac{\partial}{\partial \xi} \left[ p_1(\xi) \frac{\partial}{\partial \xi} T(t, \xi) \right] + p_2(\xi) T(t, \xi) + \phi(\sigma(t))b(\xi), \\ \frac{d}{dt} \sigma(t) &= \int_{\gamma_1}^{\gamma_2} c(\xi) T(t, \xi) d\xi, \quad \text{for all } t > 0 \text{ and almost all } \xi \in ]\gamma_1, \gamma_2[, \end{aligned}$$

subject to initial conditions and to homogeneous boundary conditions of Dirichlet-Neumann type. Systems of form (1.3) arise as dynamic models of one-dimensional continuous medium nuclear reactors. They have been studied extensively by Levin and Nohel [12], Miller [14], Bronikowski, Hall and Nohel [1] and others (see [10], [19] for more extensive bibliography) by reducing them to certain nonlinear Volterra equations, which have been discussed by means of some energy functions and/or transform methods. Infante and Walker [10] discussed (1.3) in its abstract form (1.1) with  $A$  selfadjoint, by using the theory of nonlinear  $C_0$ -semigroups combined with some estimates obtained on the basis of a Lyapunov function, which is very similar to that used previously in the theory of absolute stability of differential equations in finite-dimensional spaces. Indeed, if we define  $A$  in  $X = L^2(\gamma_1, \gamma_2)$  as the Sturm-Liouville

\* Received by the editors October 23, 1979, and in revised form February 27, 1980.

† Department of Mathematics, Facultés Universitaires N.-D. de la Paix, 61, rue de Bruxelles, B-5000 Namur, Belgium.

operator

$$Au(\xi) = \frac{d}{d\xi} \left[ p_1(\xi) \frac{d}{d\xi} u(\xi) \right] + p_2(\xi)u(\xi),$$

and require appropriate conditions on  $A$ , then (1.1) with  $\rho = 0$  is an abstract version for (1.3) (see [10], [19] for details). Clearly much more general integrodifferential systems with elliptic operator  $A$  may be written in the form (1.1).

As a second example motivating our general setting, consider the retarded differential difference system

$$(1.4) \quad \begin{aligned} \dot{y}(t) &= A_1y(t) + A_2y(t-r) + \phi(\sigma(t))\tilde{b}, \\ \dot{\sigma}(t) &= \langle \tilde{c}, y(t) \rangle - \phi(\sigma(t))\rho, \end{aligned}$$

where  $A_1, A_2$  are  $n \times n$  matrices,  $r > 0$  and  $\tilde{b}, \tilde{c} \in \mathbb{R}^n$ . Similar systems have been discussed in control theory for retarded equations by Halanay [7, Chap. 4], Corduneanu [2] and others by applying Popov type frequency domain methods to the corresponding Volterra equation for  $\sigma$ . For our purpose, the following initial conditions are suitable:

$$\begin{aligned} \sigma(0) &= \sigma_0 \in \mathbb{R}, & y(0) &= \eta \in \mathbb{R}^n, \\ y &= \psi \text{ a.e. in } [-r, 0], & \text{where } \psi &\in L^2(-r, 0; \mathbb{R}^n). \end{aligned}$$

By using the semigroup associated to linear retarded equations [8], we can easily see that system (1.4) may be written in form (1.1) with  $X = L^2(-r, 0; \mathbb{R}^n) \times \mathbb{R}^n$ ,  $b = (0, \tilde{b})$ ,  $c = (0, \tilde{c})$  and  $A$  defined as

$$A(\psi, \eta) = (\dot{\psi}, A_1\eta + A_2\psi(-r)),$$

with domain

$$D(A) = \{(\psi, \eta) \in X : \psi \in AC([-r, 0], \mathbb{R}^n), \dot{\psi} \in L^2(-r, 0, \mathbb{R}^n), \psi(0) = \eta\}.$$

Recall now that in the finite-dimensional case, powerful methods to discuss stability of system (1.1) are available. Let us mention here the following ones: 1) the application of Popov type frequency domain methods to the Volterra equation for  $\sigma$  associated to (1.1) [3, Chap. 3]; 2) the application of the Kalman-Yakubovich Lemma (a result concerning a special case of the Algebraic Riccati Equation in optimal control theory) to construct Lyapunov functions for (1.1) [15]; and 3) the application of Popov hyperstability [17], [6]. The above methods lead to so-called frequency domain stability criteria, which are expressed in terms of  $b, c$  and the resolvent of the complexification  $A^c$  of  $A$ . In finite dimension, such criteria are known to be most general and easy-to-check.

We are concerned here with some infinite-dimensional extensions of the above criteria. Although these extensions are still significant, to check them is not always an easy matter, for in the infinite-dimensional case, it may be difficult to handle the resolvent of  $A^c$ ; see [19] for more details. In addition to this, it seems that in infinite dimension, the above three methods have no more the same area of application.

In [19] we have applied the first method to our infinite-dimensional setting under the additional assumption that the semigroup  $S$  is differentiable and  $\rho = 0$  (note that this extension holds also for  $\rho \geq 0$ ). Although most significant  $C_0$ -semigroups are differentiable, this condition does not seem quite natural to the problem, but only related to some technical points in the proofs. Notice also that the semigroups associated to retarded differential equations are not differentiable. For these reasons, we would like to discuss here the application of the Algebraic (Operatorial) Riccati Equation (for short ORE). We use the Hilbert space version of the ORE by Yakubovich [21].

In this way we can drop the differentiability condition on  $S$ , but we have to introduce some other restrictions. As well known in finite dimension, the exact controllability in finite time of  $(A, b)$  is a very useful and natural condition for the ORE associated to the stability problem under consideration. This condition is meaningless in our setting, for it follows by a result of Triggiani [18], that in infinite dimension  $(A, b)$  is never exactly controllable in finite time. Moreover, weaker controllability concepts seem to be of little interest here. However, under the assumption that the semigroup  $S$  is exponentially stable, we may use Yakubovich results for the associated ORE in the so-called nondegenerate case, which does not require exact controllability of  $(A, b)$ . The above narrows our stability results to the case  $\rho > 0$ ; see Remark 1. The case  $\rho = 0$  (which is of interest too) is related to the ORE in the degenerate case. In this latter case, under the lack of controllability, it seems difficult to characterize the existence of solutions for the ORE in frequency domain terms. We recall that, in the case  $\rho = 0$ , frequency domain stability may be obtained by other means under the differentiability condition on  $S$  [19].

Notice also that to the opposite of the finite-dimensional case, the energy functions we obtain by using the ORE are no more “truly” Lyapunov functions, for they are not necessarily coercive, see § 3. This is related to some specific features of the Lyapunov Equation (for short LE) in infinite dimension, see § 2. However, we may still use these energy functions to prove stability in our setting.

**2. The Lyapunov and Riccati equations.** It is useful to consider also the complexification of the spaces and operators under consideration. The elements of the complexification  $X^c$  of  $X$  will be written as  $x + iy$ ,  $x, y \in X$ , and the inner product of  $X^c$  will be denoted by  $\langle \cdot, \cdot \rangle_{X^c}$ . For any operator  $E$  in  $X$  we denote by  $E^c$  the linear operator in  $X^c$  defined by

$$E^c(x + iy) = Ex + iEy, \quad \text{with domain } D(E^c) = D(E) + iD(E).$$

$I$  denotes the identity operator on  $X$ , so that  $I^c$  is the identity on  $X^c$ . The complexifications of the other spaces and operators which will appear below are defined in a similar way.

In this section, we assume that  $A$  generates a  $C_0$ -semigroup  $S$  on  $X$  which satisfies (1.2). Recall that, by the Hille-Yosida Theorem, these assumptions are equivalent to the following:  $A$  is densely defined, the resolvent set of  $A^c$  contains the halfplane  $\text{Re } \lambda > -\alpha$  and, for each  $n = 1, 2, 3, \dots$ , we have

$$(2.1) \quad \|(\lambda I^c - A^c)^{-n}\|_{\mathcal{L}(X^c)} \leq M(\text{Re } \lambda + \alpha)^{-n}, \quad \text{for all } \lambda \in \mathbb{C}, \text{Re } \lambda > -\alpha.$$

We see in particular that  $A^{-1} \in \mathcal{L}(X)$ .

The LE in infinite-dimensional spaces has been considered by Datko [4], [5] and Pazy [16]. As they have not discussed the uniqueness property and we make use of it in § 3, we state the following simple lemma.

LEMMA 1. For each selfadjoint operator  $P \in \mathcal{L}(X)$ , the operator  $K \in \mathcal{L}(X)$  defined by

$$Kx = \int_0^{+\infty} S^*(s)PS(s) ds \quad (\text{the symbol } * \text{ stands for the adjoint operator}),$$

is the unique selfadjoint operator in  $\mathcal{L}(X)$  satisfying the LE

$$2\langle KAx, x \rangle = \langle -Px, x \rangle, \quad \text{for all } x \in D(A).$$

Moreover, if  $P$  is positive, so is  $K$ .



*Proof.* As in the finite-dimensional case, (1.2) implies the convergence of the above integral and the fact that the linear operator  $K$  is bounded. Selfadjointness is obvious. To see that  $K$  satisfies the LE, note that

$$Kx = \int_t^{+\infty} S^*(s-t)PS(s-t)x \, ds,$$

differentiate  $\langle Kx, x \rangle$  with respect to  $t$  and take into account that  $d[S(t)x]/dt = AS(t)x = S(t)Ax$ , for all  $x \in D(A)$  and  $t \geq 0$ . To prove uniqueness, assume that  $\hat{K} \in \mathcal{L}(X)$  is another selfadjoint operator which satisfies the LE. Fix  $x \in D(A)$  and put

$$\psi(t) = \langle KS(t)x, S(t)x \rangle, \quad \hat{\psi}(t) = \langle \hat{K}S(t)x, S(t)x \rangle.$$

Since  $K$  and  $\hat{K}$  satisfy the same LE, we have  $d\psi/dt = d\hat{\psi}/dt$ ; hence by  $\lim_{t \rightarrow +\infty} \psi(t) = \lim_{t \rightarrow +\infty} \hat{\psi}(t) = 0$ , it follows that  $\psi(t) = \hat{\psi}(t)$  for all  $t \geq 0$ . We see so that  $\langle Kx, x \rangle = \langle \hat{K}x, x \rangle$ , for all  $x \in D(A)$ , and hence for all  $x \in X$  (use continuity of  $K$  and  $\hat{K}$  and the fact that  $D(A)$  is dense in  $X$ ). Since  $K$  and  $\hat{K}$  are selfadjoint, it follows then that  $\hat{K} = K$ .

Note that, according to [16], in the infinite-dimensional case,  $K$  is not necessarily coercive, even if  $P$  is so, and this narrows the field of applications of the Lyapunov Theorem in infinite dimension.

To state Yakubovich' results [21] in the form we use, let us consider another real Hilbert space  $U$ , an operator  $B \in \mathcal{L}(U, X)$  and a continuous, quadratic form  $F$  defined on the Hilbert space  $X \times U$  by

$$F(x, u) = \langle F_1x, x \rangle_X + 2\langle F_2x, u \rangle_U + \langle F_3u, u \rangle_U, \quad x \in X, \quad u \in U,$$

where  $F_1 \in \mathcal{L}(X)$ ,  $F_3 \in \mathcal{L}(U)$  are selfadjoint and  $F_2 \in \mathcal{L}(X, U)$ . The complexification  $F^c$  of  $F$  is the Hermitian form defined on  $X^c \times U^c$  by

$$F^c(z, w) = \langle F_1^c z, z \rangle_{X^c} + 2 \operatorname{Re} \langle F_2^c z, w \rangle_{U^c} + \langle F_3^c w, w \rangle_{U^c}, \quad z \in X^c, \quad w \in U^c.$$

THEOREM 1 [21]. *If for some  $\delta > 0$ ,*

$$(2.2) \quad F^c((i\omega I^c - A^c)^{-1}B^c w, w) \geq \delta |w|_{U^c}^2, \quad \text{for all } w \in U^c, \omega \in \mathbb{R},$$

*then there exists a selfadjoint operator  $H \in \mathcal{L}(X)$  such that  $H$  and  $h = -F_3^{-1} (B^*H + F_2)$  satisfy*

$$(2.3) \quad 2\langle Ax + Bu, Hx \rangle + F(x, u) = |F_3^{1/2} (u - hx)|_{U^c}^2, \quad \text{for all } (x, u) \in D(A) \times U.$$

Note that, since according to [11, Chap. IX],

$$(i\omega I^c - A^c)^{-1}z = \int_0^{+\infty} e^{-i\omega t} S^c(t)z \, dt, \quad \text{for all } z \in X^c, \omega \in \mathbb{R},$$

we have, by the Riemann-Lebesgue Theorem,

$$(2.4) \quad (i\omega I^c - A^c)^{-1}z \rightarrow 0 \quad \text{as } \omega \rightarrow \infty, \quad \text{for all } z \in X^c,$$

so that the frequency domain condition (2.2) implies that  $F^c(0, w) \geq \delta |w|_{U^c}^2$ , hence  $F_3$  is coercive.

Relation (2.3) is nothing else but the ORE written in terms of forms (see Willems [20] for the finite-dimensional case).

Yakubovich' proofs rely on optimal control methods for the continuous-time infinite-dimensional regulator problem. His results in [21] are given for bounded Hilbert space operators  $A$ , but it is easy to see that his proofs may be adapted to unbounded operators  $A$  which generate  $C_0$ -semigroups. For the convenience of the

reader and following the referee’s suggestion, we sketch this adaptation in an appendix (§ 4). Theorem 1 is a consequence of the real form of Theorem 3 in § 4. It follows easily when the complex spaces under consideration in Theorem 3 are replaced by the complexifications of the above real spaces  $X$  and  $U$ . See also [22] for related problems and [13] for generalizations to some unbounded operators  $A$  in a setting which is however different from our one. Note also the discrete-time infinite-dimensional version of the ORE by Helton, based on a spectral factorization approach [9].

**3. Stability.** Throughout this section, we assume that  $A$  generates a  $C_0$ -semigroup  $S$  which satisfies (1.2).

The function  $(x, \sigma)$  from the interval  $J$  with origin 0 to  $\mathcal{X}$  is said to be a *solution of (1.1) on  $J$  with initial data  $(x_0, \sigma_0) \in \mathcal{X}$*  if

$$(3.1) \quad x(t) = S(t)x_0 + \int_0^t \phi(\sigma(s))S(t-s)b \, ds, \quad \text{for all } t \in J,$$

and  $\sigma \in C^1(J)$  satisfies

$$(3.2) \quad \sigma(0) = \sigma_0, \quad \frac{d\sigma}{dt}(t) = \langle c, x(t) \rangle - \phi(\sigma(t))\rho, \quad \text{for all } t \in J.$$

Proposition 1 below reduces the initial data problem for (1.1) to the scalar nonlinear Volterra equation (3.3).

**PROPOSITION 1.** *For each  $(x_0, \sigma_0) \in \mathcal{X}$  and each nonvoid interval  $J$  with origin 0, there exists at most one solution of (1.1) on  $J$  with initial data  $(x_0, \sigma_0)$ . This solution is  $(x, \sigma)$  if and only if  $\sigma$  is a solution of the integral equation*

$$(3.3) \quad \begin{aligned} \sigma(t) = & \sigma_0 + \langle c, S(t)A^{-1}x_0 - A^{-1}x_0 \rangle \\ & + \int_0^t \phi(\sigma(s))[\langle c, S(t-s)A^{-1}b \rangle - \langle c, A^{-1}b \rangle - \rho] \, ds, \end{aligned}$$

*continuous on  $J$  and  $x$  satisfies (3.1). Moreover, if  $(x, \sigma)$  is a solution on  $J$  with initial data  $(x_0, \sigma_0)$  and if  $x_0 \in D(A)$ , then the component  $x$  is also of class  $C^1$  and satisfies*

$$x(t) \in D(A), \quad \frac{dx}{dt}(t) = Ax(t) + \phi(\sigma(t))b, \quad \text{for all } t \in J.$$

The proof is similar to that for Proposition 1 in [19].

We shall make use of the following “local semigroup property”.

**PROPOSITION 2.** *If  $(x_1, \sigma_1)$  is a solution of (1.1) on  $[0, \theta_1]$  with initial data  $(x_0, \sigma_0)$  and if  $(x_2, \sigma_2)$  is a solution on  $[0, \theta_2]$  with initial data  $(x_1(\theta_1), \sigma_1(\theta_1))$ , then the function  $(x, \sigma)$  defined on  $[0, \theta_1 + \theta_2]$  by*

$$\begin{aligned} x(t) = x_1(t) & \quad \text{if } t \in [0, \theta_1], & \quad x(t) = x_2(t - \theta_1) & \quad \text{if } t \in [\theta_1, \theta_1 + \theta_2], \\ \sigma(t) = \sigma_1(t) & \quad \text{if } t \in [0, \theta_1], & \quad \sigma(t) = \sigma_2(t - \theta_1) & \quad \text{if } t \in [\theta_1, \theta_1 + \theta_2], \end{aligned}$$

*is the solution of (1.1) on  $[0, \theta_1 + \theta_2]$  with initial data  $(x_0, \sigma_0)$ .*

The proof involves rather long but straightforward calculations based on Proposition 1 and the semigroup property of  $S$ , so we omit it.

The zero solution of (1.1) is said to be *stable in the large* if: (i) for each  $(x_0, \sigma_0) \in \mathcal{X}$ , there exists a solution of (1.1) on  $\mathbb{R}^+$  with initial data  $(x_0, \sigma_0)$  (uniqueness is insured by Proposition 1); and (ii) there exists a continuous, strictly increasing function  $\Pi: \mathbb{R}^+ \rightarrow \mathbb{R}^+$

with  $\Pi(0) = 0$ , such that, for any solution  $(x, \sigma)$  with initial data  $(x_0, \sigma_0)$  and any  $r > 0$ ,

$$|(x_0, \sigma_0)|_{\mathcal{X}} \leq r \text{ implies } |(x(t), \sigma(t))|_{\mathcal{X}} \leq \Pi(r) \quad \forall t \geq 0.$$

The zero solution is said to be *uniformly asymptotically stable in the large* if it is stable in the large and if, for any bounded set  $\mathcal{B}$  in  $\mathcal{X}$ , the solution  $(x, \sigma)$  of (1.1) with initial data  $(x_0, \sigma_0)$  tends to zero as  $t \rightarrow +\infty$ , uniformly with respect to  $(x_0, \sigma_0) \in \mathcal{B}$ .

We now state our main result.

**THEOREM 2.** *Assume that  $A$  generates a  $C_0$ -semigroup  $S$  which satisfies (1.2). If, in addition to this,*

- (i)  $\lim_{|s| \rightarrow \infty} \int_0^s \phi(r) dr = +\infty$ , and
- (ii) *there exists  $\delta > 0$  such that*

$$\rho - \operatorname{Re} \langle c, (i\omega I^c - A^c)^{-1} b \rangle_{X^c} \geq \delta, \text{ for all } \omega \in \mathbb{R},$$

*then the zero solution of (1.1) is uniformly asymptotically stable in the large.*

Our proof is in three steps.

*Step I, local existence and continuous dependence on initial data.* For each  $r > 0$ , we may find  $\theta > 0$  such that, for any  $(x_0, \sigma_0) \in \mathcal{X}$  with  $|(x_0, \sigma_0)|_{\mathcal{X}} \leq r$ , there exists one and only one solution of (1.1) on  $[0, \theta]$  with initial data  $(x_0, \sigma_0)$ . It is important to the sequel to note that  $\theta$  does not depend on the initial data.

The above is an easy consequence of the Contraction Mapping Principle and Proposition 1. Indeed, since  $\phi$  is locally Lipschitz, there exists a continuous strictly increasing function  $l: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  such that, for each  $r > 0$ ,

$$|\phi(r_1) - \phi(r_2)| \leq l(r)|r_1 - r_2|, \text{ for all } r_1, r_2 \in [-r, r].$$

Fix now  $r > 0$ , put  $\beta = 2r[1 + (M + 1)|c| \|A^{-1}|_{\mathcal{L}(X)}]$  and denote by  $\theta$  the unique solution of

$$\theta + \int_0^\theta |\langle c, S(s)A^{-1}b \rangle - \langle c, A^{-1}b \rangle - \rho| ds = (2l(\beta))^{-1}.$$

Choose an arbitrary  $(x_0, \sigma_0) \in \mathcal{X}$  with  $|(x_0, \sigma_0)|_{\mathcal{X}} \leq r$  and let  $\mathcal{C}$  be the subspace of functions  $\sigma \in C([0, \theta])$  satisfying  $\sigma(0) = \sigma_0$  and  $|\sigma| \leq \beta$  on  $[0, \theta]$ . Clearly  $\mathcal{C}$  is a complete metric space. For each  $\sigma \in \mathcal{C}$ , define the function  $\mathcal{U}\sigma$  on  $[0, \theta]$  by

$$\begin{aligned} \mathcal{U}\sigma(t) &= \sigma_0 + \langle c, S(t)A^{-1}x_0 - A^{-1}x_0 \rangle \\ &\quad + \int_0^t \phi(\sigma(s)) [\langle c, S(t-s)A^{-1}b \rangle - \langle c, A^{-1}b \rangle - \rho] ds. \end{aligned}$$

By using (1.2), the Lipschitz property of  $\phi$ ,  $\phi(0) = 0$  and the choice of  $\beta$  and  $\theta$ , we see easily that  $\mathcal{U}$  is a strict contraction in  $\mathcal{C}$ ; hence the Contraction Mapping Principle implies that equation (3.3) possesses one and only one solution on  $[0, \theta]$ . Our local existence claim follows now by Proposition 1.

Apply then Propositions 1 and 2 and the Lipschitz property of  $\phi$  to obtain continuous dependence on initial data: if  $(x_n, \sigma_n)$  is a uniformly bounded sequence of solutions of (1.1) on  $[0, T]$ ,  $T > 0$ , with initial data  $(x_{0n}, \sigma_{0n})$  and if  $(x_{0n}, \sigma_{0n}) \rightarrow (x_0, \sigma_0)$ , then there exists one and only one solution  $(x, \sigma)$  on  $[0, T]$  with initial data  $(x_0, \sigma_0)$  and  $(x_n, \sigma_n) \rightarrow (x, \sigma)$ , uniformly on  $[0, T]$ .

*Step II, stability in the large.* Note first that, since  $r\phi(r) > 0$  for all  $r \neq 0$ , we have that  $\int_0^s \phi(r) dr > 0$  for all  $s \neq 0$ . The functions  $s \mapsto \int_0^s \phi(r) dr$  from  $\mathbb{R}^-$  to  $\mathbb{R}^+$  and  $s \mapsto \int_0^s \phi(r) dr$  from  $\mathbb{R}^+$  to  $\mathbb{R}^+$  are continuous, strictly decreasing (respectively increasing) and equal zero at zero. In addition to this, condition (i) of Theorem 2 implies that these functions

tend to  $+\infty$  as  $s \rightarrow -\infty$ , (respectively as  $s \rightarrow +\infty$ ). Denote by  $\gamma^-: \mathbb{R}^+ \rightarrow \mathbb{R}^-$  and by  $\gamma^+: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  respectively the reciprocals of the above functions and put  $\gamma = \max(-\gamma^-, \gamma^+)$ . It follows that  $\gamma: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is continuous, strictly increasing,  $\gamma(0) = 0$ ,  $\gamma(s) \rightarrow +\infty$  as  $s \rightarrow +\infty$  and

$$(3.4) \quad \int_0^s \phi(r) \, dr \leq a \quad \text{implies} \quad |s| \leq \gamma(a), \quad \text{for all } s \in \mathbb{R} \text{ and } a \geq 0.$$

By (2.1) with  $n = 1$  and condition (ii) of Theorem 2, we may choose  $\delta' > 0$  such that

$$(3.5) \quad \rho - \delta' - \operatorname{Re} \langle c, (i\omega I^c - A^c)^{-1} b \rangle_{X^c} - \delta' |(i\omega I^c - A^c)^{-1} b|_{X^c}^2 \geq 2^{-1} \delta \quad \forall \omega \in \mathbb{R}.$$

Define the quadratic form  $F$  on  $X \times \mathbb{R}$  by

$$F(x, u) = -\delta' |x|^2 - \langle x, c \rangle u + (\rho - \delta') u^2.$$

Since, by (3.5),

$$F^c((i\omega I^c - A^c)^{-1} b w, w) \geq 2^{-1} \delta |w|^2, \quad \text{for all } w \in \mathbb{C} \text{ and } \omega \in \mathbb{R},$$

we may apply Theorem 1 with  $U = \mathbb{R}$  and  $B: \mathbb{R} \rightarrow X$  defined as  $Bu = ub$ , to see that there exists a selfadjoint  $H = \mathcal{L}(X)$  and  $h \in \mathcal{L}(X, \mathbb{R})$  such that

$$(3.6) \quad \begin{aligned} 2\langle Ax + ub, Hx \rangle - \langle x, c \rangle u + \rho u^2 &= \delta' (|x|^2 + u^2) + \rho(u - hx)^2, \\ &\text{for all } (x, u) \in D(A) \times \mathbb{R}. \end{aligned}$$

For  $u = 0$ , we get the LE

$$2\langle -H Ax, x \rangle = \langle -(\delta' I + \rho h^* h)x, x \rangle, \quad \text{for all } x \in D(A),$$

and then Lemma 1 implies that  $\langle -Hx, x \rangle \geq 0$  for all  $x \in X$ ; hence the function

$$W: X \times \mathbb{R} \rightarrow \mathbb{R}, \quad W(x, \sigma) = \langle -Hx, x \rangle + \int_0^\sigma \phi(s) \, ds$$

is also positive. As seen in § 2,  $-H$  is not necessarily coercive, so that  $W$  is not a “truly” Lyapunov function. However, its properties allow us to prove stability.

Fix an arbitrary  $r > 0$  and choose any  $(x_0, \sigma_0) \in \mathcal{X}$  with  $(x_0, \sigma_0)|_{\mathcal{X}} \leq r$ . By using Step I and Proposition 2, we may see that there exists one and only one maximal solution  $(x, \sigma)$  with initial data  $(x_0, \sigma_0)$  and that its existence interval  $[0, \tau[$ ,  $0 < \tau \leq +\infty$ , is half-open (as usual, a solution through  $(x_0, \sigma_0)$  is called maximal if it has no proper extensions).

Assume first  $x_0 \in D(A)$ , so that, according to Proposition 1, the function  $W$  is differentiable along this solution. By using (3.6), it follows that

$$(3.7) \quad \dot{W}(x(t), \sigma(t)) = -\delta' [|x(t)|^2 + \phi(\sigma(t))^2] - \rho [\phi(\sigma(t)) - hx(t)]^2, \quad \text{for all } t \in [0, \tau[,$$

hence  $W(x(\cdot), \sigma(\cdot))$  is decreasing. Then, by using the positivity of  $-H$ ,  $s\phi(s) \geq 0$ , the Lipschitz property of  $\phi$  and  $\phi(0) = 0$ , we see that, for all  $t \geq 0$ ,

$$(3.8) \quad \begin{aligned} \int_0^{\sigma(t)} \phi(s) \, ds &\leq W(x(t), \sigma(t)) \leq W(x_0, \sigma_0) \\ &\leq r^2 |H|_{\mathcal{L}(X)} + \max \left( \int_0^r \phi(s) \, ds, \int_0^{-r} \phi(s) \, ds \right) \leq r^2 (|H|_{\mathcal{L}(X)} + 2^{-1} l(r)); \end{aligned}$$

hence, by (3.4), we have  $|\sigma(t)| \leq \mu(r)$  for all  $t \in [0, \tau[$ , where

$$\mu: \mathbb{R}^+ \rightarrow \mathbb{R}^+, \quad \mu(s) = \gamma(s^2 |H|_{\mathcal{L}(X)} + 2^{-1} l(r)).$$

Use then (3.1), (1.2), the Lipschitz property of  $\phi$  and  $\phi(0) = 0$  to see that  $|x(t)| \leq \nu(r)$  for

all  $t \in [0, \tau[$ , where

$$\nu : \mathbb{R}^+ \rightarrow \mathbb{R}^+, \quad \nu(s) = rM(1 + \alpha^{-1}\mu(r)l(\mu(r))).$$

It follows that the function  $\Pi = (\mu^2 + \nu^2)^{1/2}$  is continuous, strictly increasing,  $\Pi(0) = 0$  and  $|(x(t), \sigma(t))|_{\mathcal{X}} \leq \Pi(r)$ , for all  $t \in [0, \tau[$ . Use now again Step I to find a  $\theta > 0$  such that all of the solutions with initial data in the closed ball with center 0 and radius  $\Pi(r)$  are defined on the same interval  $[0, \theta]$  and then apply Proposition 2 to see that  $\tau = +\infty$ . It follows that

$$(3.9) \quad |(x(t), \sigma(t))|_{\mathcal{X}} \leq \Pi(r), \quad \text{for all } t \in \mathbb{R}^+.$$

Finally, consider an arbitrary  $x_0 \in X$ . Since  $D(A)$  is dense in  $X$ , we may choose a sequence  $(x_{0n}, \sigma_{0n})$  in  $\mathcal{X}$  such that  $x_{0n} \in D(A)$ ,  $|(x_{0n}, \sigma_{0n})|_{\mathcal{X}} \leq r$  and  $(x_{0n}, \sigma_{0n}) \rightarrow (x_0, \sigma_0)$ . Denote by  $(x_n, \sigma_n)$  the solution with initial data  $(x_{0n}, \sigma_{0n})$ , so that, as seen above,  $|(x_n(t), \sigma_n(t))|_{\mathcal{X}} \leq \Pi(r)$ , for all  $n \in \mathbb{N}$  and  $t \in \mathbb{R}^+$ . Apply then continuous dependence on initial data to infer that  $\tau = +\infty$  and (3.9) holds whatever will be  $x_0 \in X$ .

*Step III, uniformly asymptotic stability in the large.* Since we have stability in the large and the semigroup property, it suffices to establish the following: for each bounded set  $\mathcal{B}$  in  $\mathcal{X}$  and each  $\varepsilon > 0$ , there exists  $T \geq 0$  such that, for any solution  $(x, \sigma)$  with initial data  $(x_0, \sigma_0) \in \mathcal{B}$ , we may find  $\bar{t} \in [0, T]$  such that  $|(x(\bar{t}), \sigma(\bar{t}))|_{\mathcal{X}} \leq \Pi^{-1}(\varepsilon)$ , where  $\Pi^{-1}$  is the function reciprocal to  $\Pi$ .

Suppose the above claim does not hold. There exists then a bounded set  $\mathcal{B}$  in  $\mathcal{X}$ ,  $\varepsilon > 0$ , a sequence  $(T_n)$  of positive numbers with  $T_n \rightarrow \infty$  and a sequence  $((x_{0n}, \sigma_{0n}))$  in  $\mathcal{B}$  such that the solution  $(x_n, \sigma_n)$  through  $(x_{0n}, \sigma_{0n})$  satisfies

$$(3.10) \quad |(x_n(t), \sigma_n(t))|_{\mathcal{X}} > \Pi^{-1}(\varepsilon), \quad \text{for all } t \in [0, T_n].$$

By integrating (3.7) on  $[0, t]$ , we see that, if  $x_0 \in D(A)$ , the solution  $(x, \sigma)$  with initial data  $(x_0, \sigma_0)$  satisfies

$$(3.11) \quad \delta' \int_0^t [|x(s)|^2 + \phi(\sigma(s))^2] ds \leq W(x_0, \sigma_0), \quad \text{for all } t \geq 0.$$

Use then the fact that  $D(A)$  is dense in  $X$ , stability in the large and the continuous dependence on initial data to see that (3.11) holds for all  $x_0 \in X$ . Replace  $(x_0, \sigma_0)$  by  $(x_{0n}, \sigma_{0n})$ ,  $t$  by  $T_n$  and estimate  $W(x_0, \sigma_0)$  as in (3.8) to infer that

$$(3.12) \quad \delta' \int_0^{T_n} [|x_n(s)|^2 + \phi(\sigma_n(s))^2] ds \leq r^2(|H|_{\mathcal{X}(X)} + 2^{-1}l(r)) \quad \forall n \in \mathbb{N},$$

where  $r = \sup \{|(x_0, \sigma_0)|_{\mathcal{X}} : (x_0, \sigma_0) \in \mathcal{B}\}$ . By stability, we have  $|(x_n(t), \sigma_n(t))|_{\mathcal{X}} \leq \Pi(r)$ , for all  $n$  and  $t$ ; hence, with (3.10), we obtain  $\Pi^{-1}(\varepsilon) < \Pi(r)$ . Put then

$$\eta = \inf \{\phi(s)^2 : 2^{-1/2}\Pi^{-1}(\varepsilon) \leq |s| \leq \Pi(r)\},$$

so that  $\eta > 0$ . By using (3.12), we see that, for sufficiently large  $n \in \mathbb{N}$ , there exists  $t_n \in [0, T_n]$  with

$$|x_n(t_n)| \leq 2^{-1/2}\Pi^{-1}(\varepsilon) \quad \text{and} \quad \phi(\sigma_n(t_n))^2 < \eta.$$

Since, by stability in the large, we have  $|\sigma_n(t_n)| \leq \Pi(r)$ , it follows then by the definition of  $\eta$  that  $|\sigma_n(t_n)| < 2^{-1/2}\Pi^{-1}(\varepsilon)$ , hence

$$|(x_n(t_n), \sigma_n(t_n))|_{\mathcal{X}} < \Pi^{-1}(\varepsilon), \quad \text{for sufficiently large } n \in \mathbb{N},$$

which contradicts (3.10). The proof of Theorem 2 is complete.

*Remark 1.* By using (2.4), we see that condition (ii) of Theorem 2 implies  $\rho > 0$ .

*Remark 2.* As in the finite-dimensional case, we may replace condition (ii) of Theorem 2 by the following, more general condition:  $\rho + \langle c, A^{-1}b \rangle > 0$ , and there exists  $\delta > 0$  and  $\rho \geq 0$  such that

$$\rho - \operatorname{Re} \langle c, (i\omega I^c - A^c)^{-1}(I + pA^{-1})b \rangle_{X^c} \geq \delta, \quad \text{for all } \omega \in \mathbb{R}.$$

We have then to replace in the proof  $F$  and  $W$  by  $F_1$ , respectively  $W_1$ , which are defined as

$$\begin{aligned} F_1(x, u) &= F(x, u) - p \langle c, A^{-1}x \rangle u, \\ W_1(x, u) &= W(x, u) + 2p(\langle c, A^{-1}x \rangle - \sigma)^2(\rho + \langle c, A^{-1}b \rangle)^{-1}. \end{aligned}$$

**4. Appendix.** Throughout this section,  $X$  and  $U$  are complex Hilbert spaces,  $A$  is the generator of a  $C_0$ -semigroup  $S$  on  $X$  which satisfies (1.2),  $B \in \mathcal{L}(U, X)$  and  $F : X \times U \rightarrow \mathbb{R}$  is a continuous Hermitian form,

$$F(x, u) = \langle F_1x, x \rangle_X + 2 \operatorname{Re} \langle F_2x, u \rangle_U + \langle F_3u, u \rangle_U,$$

where  $F_1 \in \mathcal{L}(X)$  and  $F_3 \in \mathcal{L}(U)$  are selfadjoint and  $F_2 \in \mathcal{L}(X, U)$ . Let us state Yakubovich' result [21] for the nondegenerate case in this setting.

**THEOREM 3 [21].** *For the existence of a selfadjoint operator  $H \in \mathcal{L}(X)$  such that the form*

$$(x, u) \mapsto \operatorname{Re} \langle Ax + Bu, Hx \rangle_X + F(x, u)$$

*is coercive on  $D(A) \times U$ , it is necessary and sufficient that for some  $\delta > 0$ ,*

$$(4.1) \quad F((i\omega I_X - A)^{-1}Bu, u) \geq \delta |u|_U^2 \quad \forall u \in U \text{ and } \omega \in \mathbb{R}.$$

*If condition (4.1) holds, then there exists a unique selfadjoint  $H^+ \in \mathcal{L}(X)$  such that  $H^+$  and  $h^+ = -F_3^{-1}(B^*H^+ + F_2)$  satisfy the following two conditions:*

- (i)  $2 \operatorname{Re} \langle Ax + Bu, H^+x \rangle_X + F(x, u) = |F_3^{1/2}(u - h^+x)|_U^2 \quad \forall (x, u) \in D(A) \times U;$
- (ii) *the  $C_0$ -semigroup  $S^+$  generated by the operator  $A^+ = A + Bh^+$  satisfies  $S^+(\cdot)a \in L^2(\mathbb{R}^+, X)$  for each  $a \in X$ .*

We sketch here the technical changes to be made in Yakubovich' proofs [21] in order to adapt them to our setting.

Consider the Cauchy problem on  $\mathbb{R}^+$

$$(4.2) \quad \frac{dx}{dt}(\cdot) = Ax(\cdot) + Bu(\cdot), \quad x(0) = a \in X;$$

throughout this section, we consider only controls  $u(\cdot) \in L^2(\mathbb{R}^+, U)$ . The function

$$x(t) = S(t)a + \int_0^t S(t-s)Bu(s) ds, \quad t \geq 0,$$

is called the *solution* of (4.2). Clearly  $x(\cdot) \in L^2(\mathbb{R}^+, X)$ . For each  $a \in X$  define the Hermitian functional  $J_a$  on  $L^2(\mathbb{R}^+, U)$  by

$$J_a(u(\cdot)) = \int_0^{+\infty} F(x(t), u(t)) dt, \quad \text{with } x(\cdot) \text{ the solution of (4.2).}$$

We denote by  $\tilde{f}$  the Fourier transform of a function  $f \in L^2(\mathbb{R}, Z)$ , where  $Z$  is a complex Hilbert space. So  $\tilde{f}$  is defined for almost all  $\omega \in \mathbb{R}$  as the limit in  $L^2(\mathbb{R}, Z)$  of the

function

$$\beta \mapsto (2\pi)^{-1/2} \int_{-\beta}^{\beta} e^{-i\omega t} f(t) dt, \quad \text{as } \beta \rightarrow +\infty.$$

When  $f$  is defined only on  $\mathbb{R}^+$ , we extend it by 0 on  $\mathbb{R}^-$  and still denote by  $\tilde{f}$  the Fourier transform of this extension.

LEMMA 2. *The Fourier transform of the solution  $x(\cdot)$  of (4.2) satisfies*

$$\tilde{x}(\omega) = (i\omega I_X - A)^{-1} (B\tilde{u}(\omega) + (2\pi)^{-1/2} a), \quad \text{a.e. in } \mathbb{R}.$$

*Proof.* This is quite trivial when  $A$  is bounded. For unbounded  $A$ , we proceed as follows. Put

$$y(t) = S(t)a$$

and

$$z(t) = \int_0^t S(t-s)Bu(s) ds, \quad t \geq 0,$$

so that  $\tilde{x}(\cdot) = \tilde{y}(\cdot) + \tilde{z}(\cdot)$ . Clearly

$$\tilde{y}(\omega) = (2\pi)^{-1/2} \int_0^{\infty} e^{-i\omega t} S(t)a dt = (2\pi)^{-1/2} (i\omega I - A)^{-1} a \quad \forall \omega \in \mathbb{R}.$$

To prove that

$$(4.3) \quad \tilde{z}(\omega) = (i\omega I - A)^{-1} B\tilde{u}(\omega), \quad \text{a.e. in } \mathbb{R},$$

assume first that  $u(\cdot)$  is continuously differentiable and  $u(\cdot), du(\cdot)/dt \in L^2(\mathbb{R}^+, U)$ . Then, according to [11, Chap. IX],  $z(\cdot)$  is continuously differentiable,

$$(4.4) \quad z(t) \in D(A) \quad \text{and} \quad \frac{dz}{dt}(t) = Az(t) + Bu(t), \quad \text{for all } t \geq 0.$$

On the other hand, we may then differentiate the integral defining  $z(\cdot)$  to see that

$$\frac{dz}{dt}(t) = S(t)Bu(0) + \int_0^t S(s)B \frac{du}{dt}(t-s) ds \quad \forall t \geq 0,$$

hence  $dz/dt$  belongs to  $L^2(\mathbb{R}^+, X)$ , which combined with  $z(0) = 0$  implies

$$\frac{\tilde{dz}}{dt}(\omega) = i\omega \tilde{z}(\omega), \quad \text{a.e. in } \mathbb{R}.$$

By (4.4), we have

$$A^{-1} \left( \frac{dz}{dt}(t) - Bu(t) \right) = z(t) \quad \forall t \geq 0,$$

and then, since  $A^{-1} \in \mathcal{L}(X)$ , the application of Fourier transform yields

$$A^{-1} (i\omega \tilde{z}(\omega) - B\tilde{u}(\omega)) = \tilde{z}(\omega), \quad \text{a.e. in } \mathbb{R},$$

hence (4.3) follows. To see that (4.3) holds for an arbitrary element  $u(\cdot)$  in  $L^2(\mathbb{R}^+, U)$ , it suffices then to note that the operators involved in (4.3) are continuous and the space of continuously differentiable functions  $u(\cdot)$  with  $u(\cdot)$  and  $du(\cdot)/dt$  in  $L^2(\mathbb{R}^+, U)$  is dense in  $L^2(\mathbb{R}^+, U)$ .

By using the above lemma, we see that the Lemmas 2 and 3 in [21] hold also in the above setting. We infer so the existence and uniqueness of an optimal control  $u^0(\cdot, a)$  minimizing  $J_a$ . Denote by  $x^0(\cdot, a)$  the corresponding optimal state and by  $V(a)$  the minimum of  $J_a$ . Lemma 5 in [21] clearly applies, so that  $a \mapsto V(a)$  is a continuous Hermitian form on  $X$ . We denote by  $H^+$  the selfadjoint operator associated to  $V$ , so that  $H^+ \in \mathcal{L}(X)$ .

LEMMA 3. *If the solution  $x(\cdot)$  of (4.2) is continuously differentiable and if  $dx(\cdot)/dt \in L^2(\mathbb{R}^+, X)$ , then*

$$x(t) \in D(A) \quad \text{and} \quad \frac{dx}{dt}(t) = Ax(t) + Bu(t), \quad \text{a.e. in } \mathbb{R}^+.$$

*Proof.* Clearly

$$(4.5) \quad \frac{\tilde{dx}}{dt}(\omega) = i\omega\tilde{x}(\omega) - (2\pi)^{-1/2}a, \quad \text{a.e. in } \mathbb{R}.$$

By Lemma 2,

$$\tilde{x}(\omega) \in D(A) \quad \text{and} \quad A\tilde{x}(\omega) = i\omega\tilde{x}(\omega) - (2\pi)^{-1/2}a - B\tilde{u}(\omega), \quad \text{a.e. in } \mathbb{R};$$

hence

$$A^{-1}\left(\frac{\tilde{dx}}{dt}(\omega) - B\tilde{u}(\omega)\right) = \tilde{x}(\omega), \quad \text{a.e. in } \mathbb{R}.$$

Apply then the reciprocal Fourier transform to see that

$$A^{-1}\left(\frac{dx}{dt}(t) - Bu(t)\right) = x(t), \quad \text{a.e. in } \mathbb{R}^+,$$

and this proves Lemma 3.

We adapt Lemma 6 in [21] as follows:

LEMMA 4. *Assume the frequency domain condition (4.1) holds and put, for each  $t \geq 0$  and  $a \in X$ ,  $S'(t)a = x^0(t, a)$ . Then:*

- (i)  *$S'$  is a  $C_0$ -semigroup on  $X$  and  $S'(\cdot)a \in L^2(\mathbb{R}^+, X)$  for each  $a \in X$ ;*
- (ii) *the generator  $A'$  of  $S'$  satisfies  $D(A') \subset D(A)$  (so that  $a \in D(A')$  implies  $x^0(t, a) \in D(A)$  for all  $t \geq 0$  and  $x^0(\cdot, a) \in C^1(\mathbb{R}^+, X)$ ); moreover, for each  $a \in D(A')$ , we have*

$$(4.6) \quad \frac{dx^0}{dt}(t, a) = Ax^0(t, a) + Bu^0(t, a), \quad \text{a.e. in } \mathbb{R}^+;$$

- (iii) *for each  $a \in X$  and each  $s \geq 0$ , we have*

$$u^0(s+t, a) = u^0(t, x^0(s, a)), \quad \text{a.e. in } \mathbb{R}^+.$$

*Proof.* The fact that  $S'(t+s) = S'(t)S'(s)$ , for all  $s, t \geq 0$  and (iii) follow as in [21];  $S'(0) = I$  and continuity of  $S'(\cdot)a$  at 0 follow from the properties of  $x^0(\cdot, a)$ . As in the proof of Lemma 5 of [21], we may see that the operator  $a \mapsto u^0(\cdot, a)$  from  $X$  to  $L^2(\mathbb{R}^+, U)$  is linear and bounded, and this allows us to see easily that  $x^0(t, \cdot) \in \mathcal{L}(X)$  for each  $t \geq 0$ , which achieves the proof of (i).

To prove (ii), we fix  $a \in D(A')$ . Then,  $S'(\cdot)a \in C^1(\mathbb{R}^+, X)$  and

$$\frac{d}{dt}S'(t)a = A'S'(t)a = S'(t)A'a, \quad \text{for each } t \geq 0.$$



We see then that  $x^0(\cdot, a) \in C^1(\mathbb{R}^+, X)$  and  $dx^0(\cdot, a)/dt = x^0(\cdot, A'a)$ , so that by (i) we have  $dx^0(\cdot, a)/dt \in L^2(\mathbb{R}^+, X)$ ; hence (4.6) follows by Lemma 3 with  $x^0(\cdot, a)$  instead of  $x(\cdot)$ . Moreover, it follows that

$$\overbrace{\frac{dx^0(\cdot, a)}{dt}}(\omega) = \overbrace{x^0(\cdot, A'a)}(\omega), \quad \text{a.e.};$$

hence by (4.5) with  $x^0(\cdot, a)$  instead of  $x(\cdot)$ ,

$$(2\pi)^{-1/2}a = i\omega \overbrace{x^0(\cdot, a)}(\omega) - \overbrace{x^0(\cdot, A'a)}(\omega), \quad \text{a.e. in } \mathbb{R}.$$

Since Lemma 2 implies that the right-hand side belongs a.e. to  $D(A)$ , it follows that  $a \in D(A)$ .

LEMMA 5. *Assume the frequency domain condition (4.1) holds. Then each solution  $x(\cdot)$  of (4.2) satisfies the dissipation inequality*

$$(4.7) \quad V(x(s)) - V(x(\sigma)) + \int_{\sigma}^s F(x(t), u(t)) dt \geq 0 \quad \forall \sigma, s, 0 \leq \sigma \leq s.$$

Equality holds in (4.7) if and only if, for some  $a \in X$ ,  $u(\cdot) = u^0(\cdot, a)$ , so that  $x(\cdot) = x^0(\cdot, a)$ .

*Proof.* As usual, define  $\hat{u}(\cdot)$  a.e. in  $\mathbb{R}^+$  by

$$\begin{aligned} \hat{u}(t) &= u(t + \sigma), \quad \text{a.e. in } [0, s - \sigma], \\ \hat{u}(t) &= u^0(t - s + \sigma, x(s)), \quad \text{a.e. in } ]s - \sigma, +\infty[. \end{aligned}$$

Clearly  $\hat{u}(\cdot) \in L^2(\mathbb{R}^+, U)$  and the solution of

$$\frac{d\hat{x}}{dt}(\cdot) = A\hat{x}(\cdot) + B\hat{u}(\cdot), \quad \hat{x}(0) = x(\sigma),$$

is

$$\begin{aligned} \hat{x}(t) &= x(t + \sigma) \quad \text{on } [0, s - \sigma], \\ \hat{x}(t) &= x^0(t - s + \sigma, x(s)) \quad \text{on } ]s - \sigma, +\infty[. \end{aligned}$$

Then it is easy to see that

$$(4.8) \quad J_{x(\sigma)}(\hat{u}(\cdot)) = \int_{\sigma}^s F(x(t), u(t)) dt + V(x(s)),$$

and this, combined with  $J_{x(\sigma)}(\hat{u}(\cdot)) \geq V(x(\sigma))$  implies (4.7). If  $u(\cdot) = u^0(\cdot, a)$ , then by using Lemma 4, we may see that  $\hat{u}(\cdot) = u^0(\cdot, x^0(\sigma, a))$ , so that (4.8) implies (4.7) with equality. Conversely, equality in (4.7) combined with (4.8) implies, for  $\sigma = 0$ , that  $J_a(u(\cdot)) = V(a)$ ; hence  $u(\cdot) = u^0(\cdot, a)$  by the uniqueness of the optimal control.

We adapt Lemma 7 of [21] as follows:

LEMMA 6. *Assume the frequency domain condition (4.1) holds and define the functional*

$$G : D(A) \times U \rightarrow \mathbb{R} \quad \text{by } G(x, u) = 2 \operatorname{Re}\langle Ax + Bu, H^+x \rangle + F(x, u).$$

Then:

- (i)  $G(a, v) \geq 0$ , for all  $a \in D(A)$  and  $v \in U$ ;
- (ii) for each  $a \in D(A')$ , there exists a set  $N \subset \mathbb{R}^+$  of Lebesgue measure zero such that

$$G(x^0(t, a), u^0(t, a)) = 0 \quad \forall t \in \mathbb{R}^+ \setminus N.$$

*Proof.* Choose  $u_0(\cdot) \in C^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)$  with  $u_0(0) = 1$  and put  $u(\cdot) = u_0(\cdot)v$ , so that  $u(\cdot) \in C^1(\mathbb{R}^+, U) \cap L^2(\mathbb{R}^+, U)$  and  $u(0) = v$ . Then, according to [11, Chap. IX], the solution  $x(\cdot)$  of (4.2) is continuously differentiable and

$$x(t) \in D(A)$$

and

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t), \quad \text{for all } t \geq 0.$$

It follows that  $V(x(\cdot)) \in C^1(\mathbb{R}^+)$ . We may then use (4.7) to infer that

$$\left. \frac{dV(x(\cdot))}{dt} \right|_{t=0} + F(x(0), u(0)) \geq 0;$$

hence (i) follows.

To prove (ii), we fix  $a \in D(A')$ . By Lemma 4, there exists  $N_1 \subset \mathbb{R}^+$  of Lebesgue measure zero such that

$$(4.9) \quad \frac{dx^0(t, a)}{dt} = Ax^0(t, a) + Bu^0(t, a) \quad \forall t \in \mathbb{R}^+ \setminus N_1.$$

By Lemma 5 we have

$$V(x^0(t, a)) = V(a) - \int_0^t F(x(s, a), u^0(s, a)) ds, \quad \text{for all } t \geq 0.$$

It follows that  $V(x^0(\cdot, a))$  is locally absolutely continuous and there exists  $N_2 \subset \mathbb{R}^+$  of measure zero such that

$$\frac{d}{dt} V(x^0(t, a)) + F(x^0(t, a), u^0(t, a)) = 0 \quad \forall t \in \mathbb{R}^+ \setminus N_2,$$

which, combined with (4.9), implies (ii) with  $N = N_1 \cup N_2$ .

*Proof of Theorem 3.* Condition (4.1) implies that the operators  $H^+$  and  $h^+$  satisfy conditions (i) and (ii) of Theorem 3. Indeed, for each  $a \in D(A')$ , there exists a set  $N \subset \mathbb{R}^+$  of Lebesgue measure zero such that

$$(4.10) \quad u^0(t, a) = h^+ x^0(t, a) \quad \forall t \in \mathbb{R}^+ \setminus N,$$

and

$$(4.11) \quad G(x^0(t, a), u) = |F_3^{1/2}(u - h^+ x^0(t, a))|_U^2 \quad \forall t \in \mathbb{R}^+ \setminus N \text{ and } u \in U.$$

This follows easily by applying our Lemma 6 and Lemma 8 of [21] with

$$\Gamma = F_3, \quad g = (B^*H^+ + F_2)x^0(t, a)$$

and

$$\gamma = \text{Re}\langle [2H^+A + F_1]x^0(t, a), x^0(t, a) \rangle.$$

Clearly we may then assume that (4.10) holds on all of  $\mathbb{R}^+$  (modify, if necessary,  $u^0(\cdot, a)$  on  $N$ ).

We have

$$(4.12) \quad \frac{d}{dt} S'(t)a = (A + Bh^+)S'(t)a \quad \forall t \geq 0 \text{ and } a \in D(A').$$

Indeed,  $a \in D(A')$  implies, by Lemma 4, that  $a \in D(A)$  and  $x^0(\cdot, a) \in C^1(\mathbb{R}^+, X)$ ; hence, by (4.10),  $u^0(\cdot, a) \in C^1(\mathbb{R}, U)$ , so that, according to [11, Chap. IX],

$$\frac{d}{dt} x^0(t, a) = Ax^0(t, a) + Bu^0(t, a) \quad \forall t \geq 0,$$

and (4.12) follows by using again (4.10).

By (4.12), it follows that, if  $a \in D(A')$ , the function  $t \mapsto Ax^0(t, a)$  is continuous on  $\mathbb{R}^+$ , so that, by letting  $t \rightarrow 0$  in (4.11), we obtain the equality of (i) of Theorem 3 for  $(x, u) \in D(A') \times U$ .

To see that  $D(A') = D(A)$  and  $A' = A + Bh^+$ , we observe that, according to [11, Chap. IX], the operator  $A + Bh^+$  with domain  $D(A)$  generates a  $C_0$ -semigroup  $S^+$ . This, combined with (4.12), implies that  $S'(t)a = S^+(t)a$ , for all  $t \geq 0$  and  $a \in D(A')$ . Then, since  $D(A')$  is dense in  $X$ , we have  $S' = S^+$ ; hence  $D(A') = D(A)$  and  $A' = A + Bh^+$ .

*Uniqueness of  $H^+$ .* Assume that the selfadjoint operator  $\hat{H} \in \mathcal{L}(X)$  is such that  $\hat{H}$  and  $\hat{h} = -F_3^{-1}(B^*\hat{H} + F_2)$  satisfy also conditions (i) and (ii) of Theorem 3. Put  $\hat{V}(a) = \langle \hat{H}a, a \rangle$ . Then, as in [21, § 1], we find that for each solution  $x(\cdot)$  of (4.2) with  $a \in D(A)$ ,

$$\int_0^{+\infty} F(x(t), u(t)) dt = \hat{V}(a) + \int_0^{+\infty} |F_3^{1/2}[u(t) - \hat{h}x(t)]|^2 dt,$$

(consider first the case  $u(\cdot) \in L^2(\mathbb{R}^+, U) \cap C^1(\mathbb{R}^+, U)$ , for which  $x(\cdot) \in C^1(\mathbb{R}^+, U)$  and then use the density of  $L^2(\mathbb{R}^+, U) \cap C^1(\mathbb{R}^+, U)$  in  $L^2(\mathbb{R}^+, U)$ ). By taking the infimum for  $u(\cdot) \in L^2(\mathbb{R}^+, U)$ , we see then that  $V(a) = \hat{V}(a)$  for all  $a \in D(A)$ ; hence (by continuity of  $V$  and  $\hat{V}$ ) for all  $a \in X$ . This implies that  $H^+ = \hat{H}$ .

The other claims of Theorem 3 follow now as in [21].

Note that Theorem 3 admits also a real variant which applies when  $X$  and  $U$  are real Hilbert spaces. It is stated in the same way with only two modifications: we have to delete the "Re" and to replace, in (4.1),  $U, B$  and  $A$  by their complexifications. This variant follows easily by applying Theorem 3 to the corresponding complexifications and by observing that then the Hermitian form  $V$  on  $X^c$  is the complexification of a quadratic form  $V'$  on  $X$ .

#### REFERENCES

- [1] T. A. BRONIKOWSKI, J. E. HALL AND J. A. NOHEL, *Quantitative estimates for a nonlinear system of integrodifferential equations arising in reactor dynamics*, this Journal, 3 (1972), pp. 567-588.
- [2] C. CORDUNEANU, *Absolute stability of some integro-differential systems*, in Ordinary Differential Equations, 1971 NRL-MRC Conference, Academic Press, 1972.
- [3] ———, *Integral Equations and Stability of Feedback Systems*, Academic Press, New York, 1973.
- [4] R. DATKO, *An extension of a theorem of A. M. Liapunov to semi-groups of operators*, J. Math. Anal. Appl., 24 (1968), pp. 290-295.
- [5] ———, *Extending a theorem of A. M. Liapunov to Hilbert spaces*, J. Math. Anal. Appl., 32 (1970), pp. 610-616.
- [6] P. FAURE, M. CLERGET, AND F. GERMAIN, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.
- [7] A. HALANAY, *Differential Equations*, Academic Press, New York, 1966.
- [8] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
- [9] J. W. HELTON, *A spectral factorization approach to the distributed stable regulator problem; the algebraic Riccati equation*, SIAM J. Control and Optimization, 14 (1976) 6, pp. 639-661.
- [10] E. F. INFANTE AND J. A. WALKER, *On the stability properties of an equation arising in reactor dynamics*, J. Math. Anal. Appl., 55 (1976), pp. 112-124.
- [11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

- [12] J. J. LEVIN AND J. A. NOHEL, *On a system of integro-differential equations occurring in reactor dynamics*, J. Math. Mech., 9 (1960), pp. 347–368.
- [13] A. L. LIKHTARNIKOV AND V. A. YAKUBOVICH, *The frequency theorem for equations of evolutionary type*, Siberian Math. J., 17 (1976), pp. 790–803.
- [14] R. K. MILLER, *An unstable nonlinear integrodifferential system*, Proceedings U.S.-Japan Seminar on Differential and Functional Equations, Benjamin, New York, 1967, pp. 479–489.
- [15] K. S. NARENDRA AND J. H. TAYLOR, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.
- [16] A. PAZY, *On the application of Lyapunov's theorem in Hilbert spaces*, this Journal, 3 (1972), pp. 291–294.
- [17] V.-M. POPOV, *Hyperstability of Control Systems*, Springer-Verlag, New York, 1973.
- [18] R. TRIGGIANI, *On the lack of exact controllability for mild solutions in Banach spaces*, J. Math. Anal. Appl., 50 (1975), pp. 438–446.
- [19] D. WEXLER, *Frequency domain stability for a class of equations arising in reactor dynamics*, this Journal, 10 (1979), pp. 118–138.
- [20] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.
- [21] V. A. YAKUBOVICH, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces, with an application to some problems in the synthesis of optimal controls I*, Siberian Math. J., 15 (1974), pp. 457–476.
- [22] ———, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces with an application to some problems in the synthesis of optimal controls II*, Siberian Math. J., 16 (1975), pp. 828–845.

## GENERALIZED CONVEXITY\*

S. GUDDER† AND F. SCHROECK‡

**Abstract.** Classical convexity theory is not broad enough to treat certain blending situations arising in nonlinear quantum mechanics, color vision, threshold phenomena, and chemistry. By relaxing some of the axioms of the classical theory, we obtain a generalized convexity theory which can be applied to nonlinear blending investigations. The axiomatic structure is derived, representation theorems are proved, threshold phenomena and convex topologies are considered, and a generalized probability theory is developed. Examples from color vision, petroleum engineering, and quantum mechanics are presented.

**1. Introduction.** In the early 1940's, J. von Neumann and O. Morgenstern [14] employed abstract convex structures in their theory of games and economic behavior. Important contributions were made a few years later by M. Stone [12] (he called such a structure a barycentric calculus). Since then, convex structures have been applied to studies in color vision [4], utility theory [6], [13], quantum mechanics [1], [2], [3], [5], [8], [9], and petroleum engineering [10], [11].

In order to apply this theory to a larger class of practical situations, we argue that some of the axioms of a convex structure must be weakened. We then study the consequences of such relaxations. Representation theorems, threshold phenomena, and convex topologies are considered. Finally, we develop a generalized probability theory based on these structures.

We feel that the resulting generalized convexity systems can be used as a vehicle for investigating nonlinear blending. To illustrate this, possible applications to color vision, petroleum engineering and quantum mechanics are presented.

**2. Convex prestructures.** A convex prestructure is a nonempty set  $S$  together with a map from  $[0, 1] \times S \times S$  to  $S$  denoted by  $(\lambda, x, y) \rightarrow \langle \lambda, x, y \rangle$ . We think of  $S$  as a set of elements that can be blended or mixed, and  $\langle \lambda, x, y \rangle$  denotes a blend of  $x$  and  $y$  in which the concentration (or proportion) of  $x$  is  $\lambda$  and the concentration of  $y$  is  $(1 - \lambda)$ . A *convex structure* is a convex prestructure  $(S, \langle \cdot, \cdot, \cdot \rangle)$  satisfying the following five postulates:

- (P1)  $\langle \lambda, x, y \rangle = \langle 1 - \lambda, y, x \rangle$  for all  $\lambda \in [0, 1]$ ,  $x, y \in S$ ;
- (P2)  $\langle \lambda, x, \langle \mu, y, z \rangle \rangle = \langle \lambda + (1 - \lambda)\mu, \langle \lambda[\lambda + (1 - \lambda)\mu]^{-1}, x, y \rangle, z \rangle$  for all  $\lambda, \mu \in [0, 1]$  with  $\lambda + (1 - \lambda)\mu \neq 0$  and  $x, y, z \in S$ ;
- (P3)  $\langle \lambda, x, x \rangle = x$  for all  $\lambda \in [0, 1]$ ,  $x \in S$ ;
- (P4) if  $\langle \lambda, x, y \rangle = \langle \lambda, x, z \rangle$  for some  $\lambda \neq 1$  and some  $x \in S$ , then  $y = z$ ;
- (P5)  $\langle 0, x, y \rangle = y$  for all  $x, y \in S$ .

The following lemma is useful for computational purposes.

LEMMA 2.1. *Under the additional condition  $\lambda \neq 1$ , (P2) is equivalent to*

- (P2')  $\langle \alpha, \langle \beta, x, y \rangle, z \rangle = \langle \alpha\beta, x, \langle \alpha(1 - \beta)(1 - \alpha\beta)^{-1}, y, z \rangle \rangle$  for all  $\alpha, \beta \in [0, 1]$ ,  $\alpha\beta \neq 1$ ,  $\alpha \neq 0$ ,  $x, y, z \in S$ .

*Proof.* Consider the transformation

$$\begin{aligned} \alpha &= 1 - (1 - \lambda)(1 - \mu), \\ \beta &= \lambda[1 - (1 - \lambda)(1 - \mu)]^{-1}, \end{aligned}$$

\* Received by the editors September 6, 1979, and in revised form February 20, 1980.

† Department of Mathematics, University of Denver, Denver, Colorado, 80208.

‡ Department of Mathematics, Florida Atlantic University, Boca Raton, Florida, 33432.

with inverse  $\lambda = \alpha\beta$  and  $\mu = \alpha(1-\alpha)(1-\alpha\beta)^{-1}$ . The Jacobians are

$$J(\alpha, \beta; \lambda, \mu) = -(1-\lambda)[1-(1-\lambda)(1-\mu)]^{-1},$$

$$J(\lambda, \mu; \alpha, \beta) = -\alpha(1-\alpha\beta)^{-1}.$$

Since  $(1-\lambda)(1-\mu) = 1$  if and only if  $\alpha = 0$ , and  $\alpha\beta = 1$  if and only if  $\lambda = 1$ , the equivalence of (P2) and (P2') may be obtained by substitution.  $\square$

We shall show later that (P5) follows from the other four postulates. The prototype example of a convex structure is a convex subset  $S_0$  of a real vector space in which  $\langle \lambda, x, y \rangle = \lambda x + (1-\lambda)y$ . In the sequel when we consider such a convex set  $S_0$  we shall always assume that it is equipped with this convex structure.

Although postulates (P1)–(P5) are natural properties for many blending operations, we now argue that for certain applications, some of these postulates must be relaxed. Some of the reasons for this are that the components being blended may interact, or the final blend may depend upon the speed with which components are mixed, or a threshold phenomenon may exist.

Postulate (P1) states that the order in which two components are mixed is immaterial. However, it is well known that if water is poured quickly into a strong acid solution, then a violent reaction occurs producing considerable heat and vapor. On the other hand, if acid is poured into water there is little interaction and the resulting blend is a weaker acid solution. For another example, when mixing a metal with a concentrated complexing agent, if the agent is added to the metal, only one ligand will adhere to each metal atom. However, when mixing in the other order, several ligands may adhere for a relatively long time.

Postulate (P2) states that if  $x$  is mixed with a blend of  $y$  and  $z$  the result is the same as when a blend of  $x$  and  $y$  is mixed with  $z$  at the same concentrations. This may not hold when blending interacting chemicals. If a small amount of oxygen is mixed with a small amount of hydrogen, a reaction occurs creating water and energy. If a large amount of nitrogen is now mixed with the resulting water, one bubbles nitrogen into the water. However, if the same amount of nitrogen is first mixed with the above amount of oxygen and the resulting blend is mixed with the hydrogen, a gaseous substance is obtained. Also, it is well known from cooking recipes that the operation of mixing ingredients need not be associative.

For a blending situation in which (P3) fails, consider octane ratings of gasoline. Suppose  $S$  is the set of possible octane numbers and  $\langle \lambda, x, y \rangle$  is the octane number for a blend of a gasoline with octane number  $x$  and a gasoline with octane number  $y$  in proportion  $\lambda$  to  $(1-\lambda)$ . Experiments have shown that octane ratings do not blend linearly [10]. For example, one can mix a gasoline with octane number 100 with a different gasoline with octane number 100 and get a gasoline with octane number 105. In this case  $\langle \lambda, x, x \rangle \neq x$ . As pointed out in [10], this occurs because  $x \in S$  does not completely specify the gasoline. If other parameters are adjoined to the parameter  $x \in S$ , so that the gasoline is completely specified, then (P3) would hold. The fact remains that in some situations it is impractical or impossible to completely specify the components of blends with a manageably small number of parameters, so (P3) might fail.

Postulate (P4) states that if  $y$  can be substituted for  $z$  in a blend with  $x$  at a certain concentration, then  $y = z$ . In a recipe calling for a small amount of butter, one can frequently substitute margarine with no perceptible difference. However, butter and margarine are certainly different, so (P4) may fail. Postulate (P4) is closely associated with threshold phenomena which we shall consider later.

Contrary to the other four postulates, we see no conceivable circumstance in which (P5) fails. If one blends nothing with  $y$  the result should certainly be  $y$ .

In the sequel we use the notation  $R_{>} = \{x \in R : x > 0\}$ ,  $R_{\geq} = \{x \in R : x \geq 0\}$ . We now give examples which show that each of the above postulates (except (P5)) is independent of the others. That is, (P1), (P2), (P3), and (P4) cannot be derived from the other four postulates. In Example  $n$  below,  $(P_n)$  does not hold while  $(P_m)$ ,  $m \neq n$ , hold,  $n = 1, 2, 3, 4$ ,  $m = 1, 2, 3, 4, 5$ .

*Example 1.* Let  $S = R$  and define

$$\langle \lambda, x, y \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ y, & \text{if } 0 \leq \lambda < 1. \end{cases}$$

*Example 2.* Let  $S = R$  and define

$$\langle \lambda, x, y \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ y, & \text{if } \lambda = 0, \\ x, & \text{if } x = y, \\ x + y, & \text{otherwise.} \end{cases}$$

To show that (P2) fails, suppose  $y = z \neq 0$ ,  $x \neq 0$ ,  $x \neq y$  and  $\lambda, \mu \neq 0, 1$ . Then the left-hand side of (P2) becomes  $x + y$  and the right-hand side is  $x + 2y$ .

*Example 3.* Let  $S = R_{>}$  and define

$$\langle \lambda, x, y \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ y, & \text{if } \lambda = 0, \\ x + y, & \text{otherwise.} \end{cases}$$

*Example 4.* Let  $S = R_{>}$  and define

$$\langle \lambda, x, y \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ y, & \text{if } \lambda = 0, \\ \max(x, y), & \text{otherwise.} \end{cases}$$

In this case  $\langle \frac{1}{2}, 4, 3 \rangle = \langle \frac{1}{2}, 4, 2 \rangle$ , but  $3 \neq 2$ .

The following example shows that (P5) does not follow from (P1), (P3), and (P4).

Let  $S = R$  and  $\langle \lambda, x, y \rangle = (1 - \lambda)x + \lambda y$ .

However, we have the following.

**LEMMA 2.2.** *If a convex prestructure satisfies (P2), (P3), and (P4), then (P5) holds.*

*Proof.* For  $0 < \lambda < 1$ , applying (P2) and (P3) gives

$$\langle \lambda, x, \langle 0, x, y \rangle \rangle = \langle \lambda, \langle 1, x, x \rangle, y \rangle = \langle \lambda, x, y \rangle.$$

Applying (P4) we have  $\langle 0, x, y \rangle = y$ .  $\square$

**3. Representation theorems.** If  $S_1$  and  $S_2$  are convex prestructures, a map  $F : S_1 \rightarrow S_2$  is *affine* if  $F(\langle \lambda, x, y \rangle_1) = \langle \lambda, F(x), F(y) \rangle_2$ . We say that  $S_1$  and  $S_2$  are *isomorphic* if there exists an affine bijection from  $S_1$  to  $S_2$ . An important representation theorem due to M. Stone [12] states that a convex prestructure  $S$  is a convex structure if and only if  $S$  is isomorphic to a convex subset of a real vector space. The following lemma gives a kind of converse to this representation theorem. This result was observed by M. Rusin [10] in a slightly different context.

LEMMA 3.1. *Let  $S$  be an arbitrary nonempty set, let  $S_0$  be a convex subset of a real vector space and suppose there exists a bijection  $F : S \rightarrow S_0$ . If we define  $\langle \lambda, x, y \rangle = F^{-1}[\lambda F(x) + (1 - \lambda)F(y)]$ ,  $(\lambda, x, y) \in [0, 1] \times S \times S$ , then  $(S, \langle \cdot, \cdot, \cdot \rangle)$  is a convex structure and  $S$  and  $S_0$  are isomorphic.*

*Proof.* Direct verification shows that (P1)–(P5) hold. It is clear that  $F : S \rightarrow S_0$  is an isomorphism.  $\square$

As an application of Lemma 3.1, let  $S = R_{>}$ ,  $0 \neq p \in R$ , and define  $\langle \lambda, x, y \rangle = [\lambda x^p + (1 - \lambda)y^p]^{1/p}$ ,  $\lambda \in [0, 1]$ ,  $x, y \in S$ . If we define  $F : S \rightarrow R_{>}$  by  $F(x) = x^p$ , then  $F$  is a bijection and  $\langle \lambda, x, y \rangle = F^{-1}[\lambda F(x) + (1 - \lambda)F(y)]$ . Hence  $(S, \langle \cdot, \cdot, \cdot \rangle)$  is a convex structure.

For another example, let  $S = R_{>}$  and define  $\langle \lambda, x, y \rangle = x^\lambda y^{(1-\lambda)}$ ,  $\lambda \in [0, 1]$ ,  $x, y \in S$ . If we define  $F : S \rightarrow R$  by  $F(s) = \log s$ , then  $F$  is a bijection and again  $\langle \lambda, x, y \rangle = F^{-1}[\lambda F(x) + (1 - \lambda)F(y)]$  so  $(S, \langle \cdot, \cdot, \cdot \rangle)$  is a convex structure. To illustrate the sensitivity of the axioms to slight changes, suppose we now let  $S = R_{\geq}$  and again define  $\langle \lambda, x, y \rangle = x^\lambda y^{(1-\lambda)}$ . Then the above function  $F$  is not defined at 0. We know that (P1)–(P5) hold for  $x, y, z \neq 0$  and it is easy to check that (P1), (P2), (P3), and (P5) hold universally. However, (P4) does not hold since for  $\lambda \neq 0$  and  $y \neq z$  we have  $0^\lambda y^{(1-\lambda)} = 0^\lambda z^{(1-\lambda)}$ .

For a practical application we turn to petroleum engineering. A crude blending model associates a pair of nonnegative numbers  $(n, a)$  with each type of gasoline [10], where  $n$  is the octane number and  $a$  the aromatic hydrocarbon content. A blend of two such components is defined by

$$\langle \lambda, (n_1, a_1), (n_2, a_2) \rangle = (\lambda n_1 + (1 - \lambda)n_2 + k\lambda(1 - \lambda)(a_1 - a_2)^2, \lambda a_1 + (1 - \lambda)a_2),$$

where  $k$  is a fixed positive number. For example, if  $k = 20$ ,  $\lambda = \frac{1}{2}$ ,  $n_1 = n_2 = 100$ ,  $a_1 = 0$ , and  $a_2 = 1$ , then the octane number of the blend is 105. It is easy to check that postulates (P1), (P3), (P4), and (P5) hold. However, the verification of (P2) is quite tedious. This verification can be avoided by considering the map  $F : R_{\geq}^2 \rightarrow R_{\geq}^2$  given by  $F((n, a)) = (n + ka^2, a)$ . Then

$$\begin{aligned} &F^{-1}[\lambda F((n_1, a_1)) + (1 - \lambda)F((n_2, a_2))] \\ &= F^{-1}[(\lambda n_1 + (1 - \lambda)n_2 + \lambda ka_1^2 + (1 - \lambda)ka_2^2, \lambda a_1 + (1 - \lambda)a_2)] \\ &= F^{-1}[(\lambda n_1 + (1 - \lambda)n_2 + k\lambda(1 - \lambda)(a_1 - a_2)^2 \\ &\quad + k(\lambda a_1 + (1 - \lambda)a_2)^2, \lambda a_1 + (1 - \lambda)a_2)] \\ &= \langle \lambda, (n_1, a_1), (n_2, a_2) \rangle. \end{aligned}$$

It follows from Lemma 3.1 that  $(R_{\geq}^2, \langle \cdot, \cdot, \cdot \rangle)$  is a convex structure. More precise hydrocarbon blending models have been constructed using five parameters instead of the two used above [11]. This again becomes a convex structure with base space  $R_{\geq}^5$ .

In what follows we develop more general representation theorems by relaxing some of the postulates for a convex structure. For this it is useful to introduce the concept of precones.

A *precone* is a nonempty set  $K$  together with two operations,  $+$ :  $K \times K \rightarrow K$  and  $\bullet$ :  $R_{\geq} \times K \rightarrow K$  (denoted  $\alpha x$ ,  $\alpha \in R_{\geq}$ ,  $x \in K$ ) and an element  $\theta \in K$  satisfying:

- (C1)  $\alpha(x + y) = \alpha x + \alpha y$ , for all  $\alpha \in R_{\geq}$ ,  $x, y \in K$ ;
- (C2)  $\alpha(\beta x) = (\alpha\beta)x$ , for all  $\alpha, \beta \in R_{\geq}$ ,  $x \in K$ ;
- (C3)  $1x = x$ ,  $0x = \theta$ , for all  $x \in K$ .



A precone  $K$  is *commutative* if

$$(C4) \quad x + y = y + x, \text{ for all } x, y \in K.$$

A precone  $K$  is *associative* if

$$(C5) \quad x + (y + z) = (x + y) + z, \text{ for all } x, y, z \in K.$$

A precone  $K$  is *distributive* if

$$(C6) \quad (\alpha + \beta)x = \alpha x + \beta x, \text{ for all } \alpha, \beta \in \mathbb{R}_{\geq}, x \in K.$$

A precone  $K$  is *cancellative* if

$$(C7) \quad x + y = x + z, \text{ for some } x \in K \text{ implies that } y = z, \text{ for any } y, z \in K.$$

A precone  $K$  has a *zero* if

$$(C8) \quad \theta + x = x, \text{ for all } x \in K.$$

A precone which satisfies (C4)–(C7) is called a *cone*. A cone which satisfies (C8) is a cone with a *zero*. The prototype example of a cone is a subset  $K_0$  of a real vector space satisfying:  $x + y \in K_0$ , for all  $x, y \in K_0$ , and  $\alpha x \in K_0$ , for all  $\alpha \geq 0, x \in K_0$ . If  $K_1$  and  $K_2$  are precones, a map  $T : K_1 \rightarrow K_2$  is *linear* if  $T(\alpha x + \beta y) = \alpha Tx + \beta Ty$ , for all  $\alpha, \beta \in \mathbb{R}_{\geq}, x, y \in K_1$ . We say that  $K_1$  and  $K_2$  are *isomorphic* if there is a linear bijection from  $K_1$  to  $K_2$ . It is implicit in Stone’s work [12] that any cone is isomorphic to a cone in a real vector space.

A precone  $K$  is a convex prestructure under the natural operation  $\langle \lambda, x, y \rangle = \lambda x + (1 - \lambda)y, \lambda \in [0, 1], x, y \in K$ . All precones will be assumed to be equipped with this convex prestructure. A subset  $S$  of a precone  $K$  is *convex* if  $\lambda x + (1 - \lambda)y \in S$  for every  $\lambda \in [0, 1], x, y \in S$ . A convex subset  $S \subseteq K$  is a *base* for  $K$  if  $\theta \notin S$  and for every  $\theta \neq x \in K$  there exists a unique  $s \in S$  and  $\alpha > 0$  such that  $x = \alpha s$ .

Let  $S$  be a convex prestructure. We now associate with  $S$  a natural precone. Define  $S_+ = \{(\alpha, s) : \alpha \in \mathbb{R}_{\geq}, s \in S\}$ . We define  $(\alpha, s) = (\beta, t)$  if  $\alpha = \beta \neq 0$  and  $s = t$  or if  $\alpha = \beta = 0$ , and we employ the notation  $\theta \equiv (0, s)$ , for all  $s \in S$ . For  $x = (\alpha, s), y = (\beta, t) \in S_+$  with  $\alpha + \beta \neq 0$ , define  $x + y = (\alpha + \beta, \langle \alpha/(\alpha + \beta), s, t \rangle)$  and define  $\theta + \theta = \theta$ . For  $\alpha \in \mathbb{R}_{\geq}, x = (\beta, s) \in S_+$ , define  $\alpha x = (\alpha\beta, s)$ .

LEMMA 3.2. *Let  $S$  be a convex prestructure.*

(1)  $S_+$  is a precone and  $S$  is isomorphic to a base for  $S_+$ .

(2) If  $S$  is isomorphic to a base for a precone  $K$ , then  $S_+$  and  $K$  are isomorphic.

(3) If  $S$  is a nonempty set,  $K$  a precone,  $S_0 \subseteq K$  a convex subset, and  $T : S \rightarrow S_0$  a bijection, then  $S$  is a convex prestructure under the operation  $\langle \lambda, s, t \rangle = T^{-1}[\lambda T(s) + (1 - \lambda)T(t)]$  and  $S$  and  $S_0$  are isomorphic.

*Proof.* (1) To verify (C1) we have for  $\beta + \gamma \neq 0, \alpha \neq 0$ ,

$$\begin{aligned} \alpha[(\beta, s) + (\gamma, t)] &= \alpha\left(\beta + \gamma, \left\langle \frac{\beta}{\beta + \gamma}, s, t \right\rangle\right) \\ &= \left(\alpha\beta + \alpha\gamma, \left\langle \frac{\beta}{\beta + \gamma}, s, t \right\rangle\right) = \alpha(\beta, s) + \alpha(\gamma, t). \end{aligned}$$

The cases  $\beta + \gamma = 0, \alpha = 0$  are trivial. To verify (C2) we have

$$\alpha[\beta(\gamma, s)] = (\alpha\beta\gamma, s) = \alpha\beta(\gamma, s).$$

(C3) follows from  $1(\alpha, s) = (\alpha, s)$  and  $0(\alpha, s) = (0, s)$ . Define  $T : S \rightarrow S_+$  by  $Ts = (1, s)$ . Then  $T : S \rightarrow T(S) \subseteq S_+$  is a bijection. To show that  $T$  is affine we have

$$T(\langle \lambda, s, t \rangle) = (1, \langle \lambda, s, t \rangle) = (\lambda, s) + (1 - \lambda, t) = \lambda T(s) + (1 - \lambda)T(t).$$

By the above,  $T(S)$  is convex. Clearly,  $\theta \notin T(S)$  and if  $\theta \neq x = (\alpha, s) \in S_+$ , then  $\alpha > 0$  and  $(\alpha, s) = \alpha(1, s)$ . Hence,  $T(S)$  is a base for  $S_+$ .

(2) Let  $T_1 : S \rightarrow T_1(S) \subseteq K$  be an isomorphism, where  $T_1(S)$  is a base for  $K$ . Define  $F : S_+ \rightarrow K$  by  $F(\alpha, s) = \alpha T_1 s$ . Then  $F$  is a bijection, and  $F$  is linear since

$$F(\alpha(\beta, s)) = F(\alpha\beta, s) = \alpha\beta T_1 s = \alpha F(\beta, s),$$

and for  $\alpha + \beta \neq 0$  we have

$$\begin{aligned} F((\alpha, s) + (\beta, t)) &= F\left(\alpha + \beta, \left\langle \frac{\alpha}{\alpha + \beta}, s, t \right\rangle\right) \\ &= (\alpha + \beta) T_1 \left( \left\langle \frac{\alpha}{\alpha + \beta}, s, t \right\rangle \right) \\ &= (\alpha\beta) [\alpha(\alpha + \beta)^{-1} T_1 s + \beta(\alpha + \beta)^{-1} T_1 t] \\ &= \alpha T_1 s + \beta T_1 t = F(\alpha, s) + F(\beta, t). \end{aligned}$$

Moreover,

$$F(\theta + \theta) = F(\theta) = F(0, s) = 0 T_1 s = \theta,$$

and by (C3) and (C1) we have

$$\theta = 0(x + x) = 0x + 0x = \theta + \theta.$$

(3) The proof is straightforward.  $\square$

Lemma 3.2 shows that any convex prestructure can be represented as a base for a precone. This lemma also showed that  $S_+$  is the unique precone (up to an isomorphism) with this property. For this reason we call  $S_+$  the *canonical* precone associated with  $S$  and the map  $T : S \rightarrow S_+$  defined by  $Ts = (1, s)$  is the *canonical embedding*.

**THEOREM 3.3.** *Let  $S$  be a convex prestructure.*

- (1)  $S$  satisfies (P1) if and only if  $S_+$  is commutative.
- (2)  $S$  satisfies (P2) if and only if  $S_+$  is associative.
- (3)  $S$  satisfies (P3) if and only if  $S_+$  is distributive.
- (4)  $S$  satisfies (P4) if and only if  $S_+$  is cancellative.
- (5)  $S$  satisfies (P5) if and only if  $S_+$  has a zero.

*Proof.* (1) If  $S$  satisfies (P1) and  $\alpha + \beta \neq 0$ , then

$$(\alpha, s) + (\beta, t) = (\alpha + \beta, \langle \alpha(\alpha + \beta)^{-1}, s, t \rangle) = (\beta + \alpha, \langle \beta(\alpha + \beta)^{-1}, t, s \rangle) = (\beta, t) + (\alpha, s).$$

Suppose  $S_+$  is commutative and  $x = (\lambda, s)$ ,  $y = (1 - \lambda, t)$ . Then

$$(1, \langle \lambda, s, t \rangle) = x + y = y + x = (1, \langle 1 - \lambda, t, s \rangle).$$

Hence,  $\langle \lambda, s, t \rangle = \langle 1 - \lambda, t, s \rangle$ .

(2) If  $S$  satisfies (P2) and  $\alpha \neq 0$ ,  $\beta + \gamma \neq 0$ , then

$$\begin{aligned} (\alpha, s) + [(\beta, t) + (\gamma, u)] &= (\alpha, s) + [(\beta + \gamma), \langle \beta(\beta + \gamma)^{-1}, t, u \rangle] \\ &= (\alpha + \beta + \gamma, \langle \alpha(\alpha + \beta + \gamma)^{-1}, s, \langle \beta(\beta + \gamma)^{-1}, t, u \rangle \rangle) \\ &= (\alpha + \beta + \gamma, \langle (\alpha + \beta)(\alpha + \beta + \gamma)^{-1}, \langle \alpha(\alpha + \beta)^{-1}, s, t \rangle, u \rangle) \\ &= (\alpha + \beta, \langle \alpha(\alpha + \beta)^{-1}, s, t \rangle) + (\gamma, u) \\ &= [(\alpha, s) + (\beta, t)] + (\gamma, u). \end{aligned}$$

The other cases are left to the reader. Conversely, suppose  $S_+$  is associative and

$$x = (\lambda, s), \quad y = ((1 - \lambda)(\mu, t), \quad z = ((1 - \lambda)(1 - \mu), v), \quad \text{where } \lambda + (1 - \lambda)\mu \neq 0.$$

Then

$$\begin{aligned} (1, \langle \lambda, s, \langle \mu, t, v \rangle \rangle) &= (\lambda, s) + (1 - \lambda, \langle \mu, t, v \rangle) \\ &= (\lambda, s) + [((1 - \lambda)\mu, t) + ((1 - \lambda)(1 - \mu), v)] \\ &= x + (y + z) = (x + y) + z \\ &= (\lambda + (1 - \lambda)\mu, \langle \lambda[\lambda + (1 - \lambda)\mu]^{-1}, s, t \rangle) + ((1 - \lambda)(1 - \mu), v) \\ &= (1, \langle \lambda + (1 - \lambda)\mu, \langle \lambda[\lambda + (1 - \lambda)\mu]^{-1}, s, t \rangle, v \rangle) \end{aligned}$$

and (P2) follows.

(3) Suppose  $S$  satisfies (P3) and  $\gamma \neq 0, \alpha + \beta \neq 0$ . Then

$$\begin{aligned} \alpha(\gamma, s) + \beta(\gamma, s) &= (\alpha\gamma, s) + (\beta\gamma, s) \\ &= ((\alpha + \beta)\gamma, \langle \alpha\gamma[(\alpha + \beta)\gamma]^{-1}, s, s \rangle) \\ &= ((\alpha + \beta)\gamma, s) = (\alpha + \beta)(\gamma, s). \end{aligned}$$

The other cases are left to the reader. Conversely, suppose  $S_+$  is distributive. Then

$$(1, \langle \lambda, s, s \rangle) = (\lambda, s) + (1 - \lambda, s) = \lambda(1, s) + (1 - \lambda)(1, s) = (1, s).$$

Hence  $\langle \lambda, s, s \rangle = s$ .

(4) If  $S$  satisfies (P4),  $\alpha, \beta, \gamma \neq 0$ , and  $(\alpha, s) + (\beta, t) = (\alpha, s) + (\gamma, v)$ , then

$$(\alpha + \beta, \langle \alpha(\alpha + \beta)^{-1}, s, t \rangle) = (\alpha + \gamma, \langle \alpha(\alpha + \gamma)^{-1}, s, v \rangle).$$

Hence,  $\beta = \gamma$  and

$$\langle \alpha(\alpha + \beta)^{-1}, s, t \rangle = \langle \alpha(\alpha + \beta)^{-1}, s, v \rangle.$$

Therefore,  $t = v$ . The other cases are left to the reader. Conversely, suppose  $S_+$  is cancellative and  $\langle \lambda, s, t \rangle = \langle \lambda, s, v \rangle, \lambda \neq 1$ . Let  $x = (\lambda, s), y = (1 - \lambda, t)$ , and  $z = (1 - \lambda, v)$ . Then  $x + y = x + z$ , so  $y = z$  and  $t = v$ .

(5) If  $S$  satisfies (P5), and  $\alpha \neq 0$ , then

$$(0, t) + (\alpha, s) = (\alpha, \langle 0, t, s \rangle) = (\alpha, s).$$

We have shown in the proof of Lemma 3.2 that  $\theta + \theta = \theta$ . Conversely, if  $S_+$  has a zero, then

$$(1, \langle 0, s, t \rangle) = (0, s) + (1, t) = (1, t).$$

Hence  $\langle 0, s, t \rangle = t$ .  $\square$

**COROLLARY 3.4.** *A convex prestructure  $S$  satisfies (P1) or (P2) or (P3) or (P4) or (P5), respectively, if and only if  $S$  is isomorphic to a base for a precone  $K$  where  $K$  is commutative or associative or distributive or cancellative or has a zero, respectively.*

**COROLLARY 3.5.** *Let  $S$  be a nonempty set,  $S_0$  a base for a precone  $K$  and  $T : S \rightarrow S_0$  a bijection. Define  $\langle \lambda, s, t \rangle = T^{-1}[\lambda Ts + (1 - \lambda)Tt]$ . Then  $(S, \langle \cdot, \cdot, \cdot \rangle)$  satisfies (P1), (P2), (P3), (P4), or (P5), respectively, if and only if  $K$  is commutative, associative, distributive, cancellative or has a zero, respectively.*

One can generalize the concept of a vector space and obtain representation theorems for convex prestructures on these generalized vector spaces. Since the methods and results are similar to those for precones we shall just give a brief outline.

We define a *prevector* space to be a nonempty set  $V$  together with two operations,  $+$ :  $V \times V \rightarrow V$  and  $\bullet$ :  $R \times V \rightarrow V$  (denoted  $\alpha x$ ,  $\alpha \in R$ ,  $x \in V$ ), and an element  $\theta \in V$  satisfying (C1), (C2), and (C3) in which  $R_{\geq}$  is replaced by  $R$  and  $K$  by  $V$ . In a similar way, by replacing  $R_{\geq}$  by  $R$  and  $K$  by  $V$  in (C4)–(C8) we obtain definitions of commutative, associative,  $\dots$  prevector spaces.

If  $S$  is a convex prestructure, we define  $V(S) = \{(x, y) : x, y \in S_+\}$ . Define  $\theta = (\theta, \theta)$  and  $(x, y) + (x', y') = (x + x', y + y')$ . If  $\alpha \geq 0$ , define  $\alpha(x, y) = (\alpha x, \alpha y)$ , and if  $\alpha < 0$ , define  $\alpha(x, y) = (-\alpha y, -\alpha x)$ . It is straightforward to check that  $V(S)$  is a prevector space. Define  $W : S \rightarrow V(S)$  by  $Ws = ((1, s), \theta) = (Ts, \theta)$ . Then  $W$  is affine since

$$\begin{aligned} W\langle \lambda, s, t \rangle &= ((1, \langle \lambda, s, t \rangle), \theta) \\ &= ((\lambda, s) + (1 - \lambda, t), \theta + \theta) = ((\lambda, s), \theta) + ((1 - \lambda, t), \theta) \\ &= \lambda((1, s), \theta) + (1 - \lambda)((1, t), \theta) = \lambda Ws + (1 - \lambda) Wt. \end{aligned}$$

Thus,  $S$  is isomorphic to a base for the prevector space  $V(S)$ , where the obvious definitions are used. Then results similar to Lemma 3.2, Theorem 3.3 and their corollaries hold with the notable exceptions of Theorem 3.3(3) and the distributive law in the corollaries. (For the failure of the distributive law consider  $(\alpha + \beta)(x, y)$ , where  $\alpha\beta < 0$ .)

**4. Equivalence classes, homomorphisms.** In this section we introduce several natural equivalence relations on convex prestructures, and show that there is a nontrivial homomorphism of any convex prestructure satisfying only (P1), (P2), and (P3) onto a convex subset of a real vector space.

Denote the set of affine maps from the convex prestructure  $S_1$  to the convex prestructure  $S_2$  by  $Af(S_1, S_2)$ . We call the elements of  $S^* \equiv Af(S, R)$  affine *functionals*. It is shown in [3], [4] that  $S$  is a convex structure if and only if  $S^*$  separates elements of  $S$ .

We now define three relations on a convex prestructure  $S$ . We write  $x \sim y$  if  $\langle \lambda, x, z \rangle = \langle \lambda, y, z \rangle$ , for all  $\lambda \in [0, 1]$ , and all  $z \in S$ . We write  $x \approx y$  if there exists a  $\lambda \in (0, 1)$  and a  $z \in S$  such that  $\langle \lambda, x, z \rangle = \langle \lambda, y, z \rangle$ . Finally, we write  $x \cong y$  if for all  $f \in S^*$ ,  $f(x) = f(y)$ .

It is easy to show that  $x \sim y$  implies  $x \approx y$  implies  $x \cong y$ , and that  $\sim$  and  $\cong$  are equivalence relations. The statement “ $x \sim y$ ” is interpreted “ $x$  may always be substituted for  $y$ ” and “ $x \approx y$ ” means “ $x$  may sometimes be substituted for  $y$ .”

**LEMMA 4.1.** *If  $S$  is a convex prestructure satisfying (P1) and (P2), then  $\approx$  is an equivalence relation.*

*Proof.* Transitivity is the only nonobvious property. Suppose  $x \approx y$  and  $y \approx z$ . Then there exist  $\lambda, \mu \in (0, 1)$ ,  $v, w \in S$ , such that  $\langle \lambda, x, v \rangle = \langle \lambda, y, v \rangle$ ,  $\langle \mu, y, w \rangle = \langle \mu, z, w \rangle$ . Let  $k = \mu(1 - \lambda)[\mu(1 - \lambda) + \lambda(1 - \mu)]^{-1}$  and  $\sigma = \lambda\mu[\mu(1 - \lambda) + \lambda(1 - \mu) + \lambda\mu]^{-1}$ . Then  $k, \sigma \in (0, 1)$  and

$$\begin{aligned} \langle \sigma, x, \langle k, v, w \rangle \rangle &= \langle \sigma + (1 - \sigma)k, \langle \sigma[\sigma + (1 - \sigma)k]^{-1}, x, v \rangle, w \rangle \\ &= \langle \sigma + (1 - \sigma)k, \langle \lambda, x, v \rangle, w \rangle = \langle \sigma + (1 - \sigma)k, \langle \lambda, y, v \rangle, w \rangle \\ &= \langle \sigma, y, \langle k, v, w \rangle \rangle = \langle \sigma, y, \langle 1 - k, w, v \rangle \rangle \\ &= \langle \sigma + (1 - \sigma)(1 - k), \langle \sigma[\sigma + (1 - \sigma)(1 - k)]^{-1}, y, w \rangle, v \rangle \\ &= \langle \sigma + (1 - \sigma)(1 - k), \langle \mu, y, w \rangle, v \rangle \\ &= \langle \sigma + (1 - \sigma)(1 - k), \langle \mu, z, w \rangle, v \rangle \\ &= \langle \sigma, z, \langle 1 - k, w, v \rangle \rangle = \langle \sigma, z, \langle k, v, w \rangle \rangle. \end{aligned}$$

Hence,  $x \approx z$ .  $\square$

In the sequel, when we speak of the relation  $\approx$  we shall always assume that (P1) and (P2) hold so we get an equivalence relation. We now form the quotient spaces  $S/\sim$ ,  $S/\approx$ ,  $S/\cong$  consisting of the residue classes  $[x]_{\sim}$ ,  $[x]_{\approx}$ ,  $[x]_{\cong}$ ,  $x \in S$ . Denote the natural surjections by

$$i : S \rightarrow S/\sim, \quad j : S \rightarrow S/\approx, \quad k : S \rightarrow S/\cong.$$

**THEOREM 4.2.** *The quotient spaces  $S/\sim$ ,  $S/\approx$ ,  $S/\cong$  are convex prestructures under the operations*

$$\begin{aligned} \langle \lambda, [x]_{\sim}, [y]_{\sim} \rangle &= [\langle \lambda, x, y \rangle]_{\sim}, \\ \langle \lambda, [x]_{\approx}, [y]_{\approx} \rangle &= [\langle \lambda, x, y \rangle]_{\approx}, \\ \langle \lambda, [x]_{\cong}, [y]_{\cong} \rangle &= [\langle \lambda, x, y \rangle]_{\cong}, \end{aligned}$$

and  $i$ ,  $j$ , and  $k$  are affine.

*Proof.* We shall do the proofs for  $\approx$ . Hence, let  $S$  satisfy (P1) and (P2). To show that the operation is well defined, suppose  $x_1 \approx x_2$  and  $y_1 \approx y_2$ . Then there exists a  $\lambda \in (0, 1)$ ,  $z \in S$  such that  $\langle \lambda, z, x_1 \rangle = \langle \lambda, z, x_2 \rangle$ . For any  $\beta \in (0, 1]$ , choose  $\mu = \mu(\beta) = \lambda\beta[1 - \lambda + \lambda\beta]^{-1}$ . (Hence,  $\mu \in (0, \lambda]$ .)

Then we have

$$\begin{aligned} \langle \mu, z, \langle \beta, x_1, y_1 \rangle \rangle &= \langle \mu + (1 - \mu)\beta, \langle \lambda, z, x_1 \rangle, y_1 \rangle \\ &= \langle \mu + (1 - \mu)\beta, \langle \lambda, z, x_2 \rangle, y_1 \rangle = \langle \mu, z, \langle \beta, x_2, y_1 \rangle \rangle. \end{aligned}$$

Hence,  $\langle \beta, x_1, y_1 \rangle \approx \langle \beta, x_2, y_1 \rangle$ , for any  $\beta \in (0, 1]$ . Similarly,  $\langle \beta, x_2, y_1 \rangle \approx \langle \beta, x_2, y_2 \rangle$ , for any  $\beta \in (0, 1]$ .

The case  $\beta = 0$  follows from (P1).

Hence,  $S/\approx$  is a convex prestructure and

$$j(\langle \lambda, x, y \rangle) = [\langle \lambda, x, y \rangle]_{\approx} = \langle \lambda, [x]_{\approx}, [y]_{\approx} \rangle = \langle \lambda, j(x), j(y) \rangle,$$

so  $j$  is affine.  $\square$

**THEOREM 4.3** (1) *The convex prestructure  $S/\approx$  satisfies (P1), (P2), and (P4). If  $S$  satisfies (P3), then  $S/\approx$  is a convex structure.*

(2)  *$S/\cong$  is a convex structure.*

*Proof.* The proof of (1) is straightforward. If  $f \in S^*$ , then  $\hat{f} \in (S/\cong)^*$  where  $\hat{f}([x]_{\cong}) = f(x)$ , and conversely, if  $\hat{g} \in (S/\cong)^*$  then  $\hat{g} \circ k \in S^*$ . Hence, if  $[x]_{\cong} \neq [y]_{\cong}$  there exists an  $f \in S^*$  such that  $f(x) \neq f(y)$ , so  $(S/\cong)^*$  separates elements of  $S/\cong$ . It follows from the remark at the beginning of this section that  $S/\cong$  is a convex structure.  $\square$

**COROLLARY 4.4.** *Let  $S$  satisfy (P1), (P2), and (P3). Then there is an affine surjection  $J$  of  $S$  onto a generating convex subset of a real vector space  $V$ . Furthermore,  $J(x) \neq J(y)$  if and only if there exists a  $\lambda \in (0, 1)$  and a  $z \in S$  such that  $\langle \lambda, x, z \rangle \neq \langle \lambda, y, z \rangle$ ; and  $V$  is unique up to an isomorphism.*

*Proof.* Since  $S$  satisfies (P1), (P2), and (P3),  $S/\approx$  is a convex structure. Hence, there exists an isomorphism  $T : S/\approx \rightarrow S_0$  for some generating convex subset  $S_0$  of a real vector space  $V$ . Then  $T \circ j : S \rightarrow S_0$  is an affine surjection. If  $T_1 : S \rightarrow S_1$  is an affine surjection, where  $S_1$  is a generating convex subset of a real vector space  $W$ , then the natural extension of  $T \circ j \circ T_1^{-1} : W \rightarrow V$  is an isomorphism.  $\square$

Thus, if  $S$  satisfies (P1), (P2), and (P3) it has a nontrivial linear representation and any nonlinear behavior of a convex prestructure occurring through the failure of (P4) gets lost (by indentifications) when making a linear representation of the system. Notice

that any convex prestructure  $S$  admits the trivial homomorphism  $f: S \rightarrow R$  given by  $f(x) = 0$  for all  $x \in S$ .

Since the relation  $\approx$  is playing the major role in this homomorphism theorem, it deserves a final piece of attention. We show that if  $x \approx y$  for a single mixing parameter  $\lambda$ , then  $x \approx y$  is achieved for infinitely many mixing parameters.

LEMMA 4.5. *If  $S$  is a convex prestructure satisfying (P2) and  $\langle \lambda, x, x \rangle \sim x$  for all  $\lambda \in (0, 1)$ , (a weakened version of (P3)), then*

$$\langle \lambda, z, x \rangle = \langle \lambda, z, y \rangle \text{ for some } \lambda \in [0, 1] \text{ implies}$$

$$\langle \alpha, z, x \rangle = \langle \alpha, z, y \rangle \text{ for all } \alpha \geq \lambda.$$

*Proof.* For any  $\rho \in [0, 1]$  we have

$$\begin{aligned} \langle \rho, z, \langle \lambda, z, x \rangle \rangle &= \langle \rho + (1 - \rho)\lambda, \langle \rho[\rho + (1 - \rho)\lambda]^{-1}, z, z \rangle, x \rangle \\ &= \langle \rho + (1 - \rho)\lambda, z, x \rangle. \end{aligned}$$

Hence,

$$\begin{aligned} \langle \rho + (1 - \rho)\lambda, z, y \rangle &= \langle \rho, z, \langle \lambda, z, y \rangle \rangle \\ &= \langle \rho, z, \langle \lambda, z, x \rangle \rangle = \langle \rho + (1 - \rho)\lambda, z, x \rangle. \end{aligned}$$

Thus,  $\langle \alpha, z, x \rangle = \langle \alpha, z, y \rangle$  for all  $\alpha \in [\lambda, 1]$ .  $\square$

We may interpret Lemma 4.5 as follows. If we may substitute  $x$  for  $y$  in a mixture with  $z$  where  $y$  has concentration  $1 - \lambda$ , then we may substitute  $x$  for  $y$  in a mixture with  $z$  for all lower concentrations.

**5. Threshold behavior.** In any application in which one has thresholds for distinguishing between objects or thresholds dividing linear from nonlinear behavior, (P4) seems too strong an axiom. The following experiment in color perception is proposed as a possibility for finding threshold behavior.

Project a colored pattern on a screen for a subject to observe. The pattern consists of a small circle inside a large concentric circle. The inner circle is divided in half; one half being color  $A$  and the other half color  $B$ . The annulus between the two circles is colored with color  $C$ . With a fixed color  $A$  and color  $C$ , change color  $B$  until the subject finds a match with  $A$ . Increase the size of the inner circle slightly and ask the subject if colors  $A$  and  $B$  still match. Continue this process. If at some stage, colors  $A$  and  $B$  do not match, a threshold effect has occurred. To our knowledge, no experiment of this type has been performed.

In view of Lemma 4.5, one could define

$$t_{x,y}^z = \sup_{\lambda \in [0, 1]} \{ \lambda : \langle 1 - \lambda, z, x \rangle = \langle 1 - \lambda, z, y \rangle \},$$

the ‘‘threshold for substituting  $x$  for  $y$  in a mixture with  $z$ .’’ Then  $x$  may be substituted for  $y$  in a mixture with  $z$  whenever  $\lambda < t_{x,y}^z$ . Also, let  $t(x, y) = \inf \{ t_{x,y}^z : z \in S \}$  to get an absolute threshold.

For an example of a threshold effect, let  $S = \{x, y, z\}$  and define

$$\langle \lambda, x, y \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ y, & \text{if } \lambda < 1, \end{cases}$$

$$\langle \lambda, y, z \rangle = \begin{cases} y, & \text{if } \lambda = 1, \\ z, & \text{if } \lambda < 1, \end{cases}$$

$$\langle \lambda, x, z \rangle = \begin{cases} x, & \text{if } \lambda = 1, \\ z, & \text{if } \lambda < 1. \end{cases}$$

Define the operation with the other permutations of  $x, y, z$  so that (P1), (P3) and (P5) hold. Then  $(S, \langle \cdot, \cdot, \cdot \rangle)$  satisfies all the postulates of a convex structure except for (P4). Postulate (P4) fails since for  $\lambda \neq 1$

$$\langle \lambda, z, x \rangle = \langle \lambda, z, y \rangle = z,$$

and yet  $x \neq y$ . Notice, however, that  $x \approx y$ . If we think of  $x, y, z$  as vertices of a triangle, we can get a nonlinear “coloring” of the solid triangle as follows. If a point of the triangle has barycentric coordinates  $(\lambda, \mu), \lambda, \mu \in [0, 1]$ , “color” the point with the blend  $\langle \lambda, \langle \mu, x, y \rangle, z \rangle$ . We then obtain the following nonlinear coloring:

$$\langle \lambda, \langle \mu, x, y \rangle, z \rangle = \begin{cases} x, & \text{if } \lambda = \mu = 1, \\ y, & \text{if } \lambda = 1, \mu < 1, \\ z, & \text{if } \lambda < 1. \end{cases}$$

We call the above an *extreme* threshold since the threshold occurs only at the extreme points.

The following theorem shows that two nonextreme thresholds cannot exist for interconnected pairs of elements in a convex prestructure satisfying (P1) and (P2).

**THEOREM 5.1.** *Let  $S$  be a convex prestructure satisfying (P1) and (P2). Let  $x, y, z \in S$  satisfy*

$$\begin{aligned} \langle \lambda, x, y \rangle &= \begin{cases} y, & \text{if } \lambda < \lambda_0, \\ \neq y, & \text{if } \lambda \geq \lambda_0, \end{cases} \\ \langle \mu, y, z \rangle &= \begin{cases} z, & \text{if } \mu < \mu_0, \\ \neq z, & \text{if } \mu \geq \mu_0, \end{cases} \end{aligned}$$

where  $\lambda_0, \mu_0 > 0$ . Then either  $\lambda_0$  or  $\mu_0$  equals 1.

*Proof.* Assume  $\lambda_0 < 1$  and  $\mu_0 < 1$ . Then there exists a triple of numbers  $\lambda, \mu, \delta \in [0, 1]$  satisfying the inequalities:  $\lambda < \lambda_0, \delta < \mu_0$ ,

$$\begin{aligned} \mu(1-\lambda)(1-\mu\lambda)^{-1} &< \mu_0, \mu\lambda[\mu\lambda + (1-\mu\lambda)\delta]^{-1} < \lambda_0, \\ \mu\lambda + (1-\mu\lambda)\delta &< \mu_0 \leq \mu. \end{aligned}$$

We then have

$$\begin{aligned} z \neq \langle \mu, y, z \rangle &= \langle \mu, \langle \lambda, x, y \rangle, z \rangle \\ &= \langle \mu\lambda, x, \langle \mu(1-\lambda)(1-\mu\lambda)^{-1}, y, z \rangle \rangle \\ &= \langle \mu\lambda, x, z \rangle = \langle \mu\lambda, x, \langle \delta, y, z \rangle \rangle \\ &= \langle \mu\lambda + (1-\mu\lambda)\delta, \langle \mu\lambda[\mu\lambda + (1-\mu\lambda)\delta]^{-1}, x, y \rangle, z \rangle \\ &= \langle \mu\lambda + (1-\mu\lambda)\delta, y, z \rangle = z. \end{aligned}$$

Since this is a contradiction,  $\lambda_0$  or  $\mu_0$  must equal 1.  $\square$

The following theorem gives further constraints on nonextreme thresholds.

**THEOREM 5.2.** *Let  $S$  be a convex prestructure satisfying (P2) and (P3). Suppose*

$$\langle \lambda, x, y \rangle = \begin{cases} y, & \text{if } \lambda < \lambda_0 < 1, \\ x, & \text{for } \lambda = \text{some } \alpha, \quad \lambda_0 \leq \alpha < 1. \end{cases}$$

*Then  $x = y$ .*

*Proof.* Suppose  $x = \langle \alpha, x, y \rangle = \langle \beta, x, y \rangle$ . Then

$$\begin{aligned} x &= \langle \alpha, x, y \rangle = \langle \alpha, \langle \beta, x, y \rangle, y \rangle \\ &= \langle \alpha\beta, x, \langle \alpha(1-\beta)(1-\alpha\beta), y, y \rangle \rangle \\ &= \langle \alpha\beta, x, y \rangle. \end{aligned}$$

Thus  $x = \langle \alpha^2, x, y \rangle$  and by induction  $x = \langle \alpha^n, x, y \rangle$ . But  $\alpha^n < \lambda_0$  for some  $n$  and hence,  $x = \langle \alpha^n, x, y \rangle = y$ .  $\square$

In particular, (P2) and (P3) rule out step-function behavior:

$$\langle \lambda, x, y \rangle = \begin{cases} y, & \text{if } \lambda < \lambda_0, \\ x, & \text{if } \lambda \geq \lambda_0, \end{cases}$$

except for extreme thresholds  $\lambda_0 = 1$ . Lemma 6.2 in the next section yields further constraints on threshold behavior when (P2) and (P3) hold. It thus seems that the proper framework for studying threshold behavior is a convex prestructure satisfying (P2) and (P5).

If  $x, y$  exhibit threshold behavior of the form

$$\langle \lambda, x, y \rangle = \begin{cases} y, & \text{if } \lambda < \lambda_0, \\ \neq y, & \text{if } \lambda > \lambda_0, \end{cases}$$

then if (P3) holds we have  $\langle \lambda, y, y \rangle = \langle \lambda, x, y \rangle$  for  $\lambda < \lambda_0$ , so  $x \approx y$ . If in addition (P1) and (P2) hold, Corollary 4.4 applies and this type of threshold behavior will vanish in a well-controlled manner under linear representations of the system.

**6. The natural topology.** Let  $S$  be a convex prestructure. In order to discuss convergence and continuity on  $S$  we define a topology which is compatible with the convexity structure on  $S$ . In this section we shall always assume that  $S$  satisfies (P2). Although it is not absolutely necessary to assume (P2), the topology becomes much simpler when this postulate holds.

For  $x \in S$  and  $0 < \varepsilon \leq 1$ , we define the  $\varepsilon$ -neighborhood of  $x$  to be  $N_x(\varepsilon) = \{ \langle \lambda, y, x \rangle : y \in S, \lambda < \varepsilon \} \cup \{x\}$ . We may think of  $N_x(\varepsilon)$  as the set of mixtures of  $x$  with other elements  $y$  where the concentration of  $y$  is less than  $\varepsilon$ . Notice that if (P3) or (P5) holds then  $x \in \{ \langle \lambda, y, x \rangle : y \in S, \lambda < \varepsilon \}$ , so the description of  $N_x(\varepsilon)$  simplifies.

LEMMA 6.1. *The collection  $\{N_x(\varepsilon) : x \in S, 0 < \varepsilon \leq 1\}$  is a basis for a topology on  $S$ .*

*Proof.* Let  $z \in N_x(\varepsilon)$ . We show that there exists an  $\varepsilon_1 > 0$  such that  $N_z(\varepsilon_1) \subseteq N_x(\varepsilon)$ . If  $z = x$ , choose  $\varepsilon_1 = \varepsilon$ . If  $z \neq x$ , there exist  $t \in S, \lambda < \varepsilon$ , such that  $z = \langle \lambda, t, x \rangle$ . Let  $\varepsilon_1 = \varepsilon - \lambda$ . Suppose  $u \in N_z(\varepsilon_1), u \neq z$ . There exist  $v \in S, \mu < \varepsilon_1$  such that

$$\begin{aligned} u &= \langle \mu, v, z \rangle = \langle \mu, v, \langle \lambda, t, x \rangle \rangle \\ &= \langle \mu + (1-\mu)\lambda, \langle \mu[\mu + (1-\mu)\lambda]^{-1}, v, t \rangle, x \rangle. \end{aligned}$$

But,

$$\mu + (1-\mu)\lambda = \lambda + \mu(1-\lambda) < \lambda + (\varepsilon - \lambda)(1-\lambda) = \varepsilon - \lambda(\varepsilon - \lambda) < \varepsilon.$$

Hence,  $u \in N_x(\varepsilon)$  and  $N_z(\varepsilon_1) \subseteq N_x(\varepsilon)$ . Suppose  $z \in N_x(\varepsilon_1) \cap N_y(\varepsilon_2)$ . There exist  $\varepsilon'_1, \varepsilon'_2$  such that  $N_z(\varepsilon'_1) \subseteq N_x(\varepsilon_1), N_z(\varepsilon'_2) \subseteq N_y(\varepsilon_2)$ . Letting  $\varepsilon = \min(\varepsilon'_1, \varepsilon'_2)$  gives  $N_z(\varepsilon) \subseteq N_x(\varepsilon_1) \cap N_y(\varepsilon_2)$ .  $\square$

We call the topology generated by the neighborhood basis  $\{N_x(\varepsilon)\}$  the *natural topology* on  $S$ . In considerations involving the real line we shall always assume that  $R$  is endowed with its usual topology.



LEMMA 6.2. *If  $S$  satisfies (P3), then the function  $\lambda \mapsto \langle \lambda, x, y \rangle$  is continuous on  $[0, 1]$  for all  $x, y \in S$ .*

*Proof.* Let  $\varepsilon > 0$  and  $\lambda_0 \in [0, 1]$ . If  $|\lambda - \lambda_0| < |1 - \lambda_0|\varepsilon$ , and  $\mu = (\lambda - \lambda_0)(1 - \lambda_0)^{-1}$ , then

$$\begin{aligned} \langle \lambda, x, y \rangle &= \langle \mu + (1 - \mu)\lambda_0, x, y \rangle \\ &= \langle \mu + (1 - \mu)\lambda_0, \langle \mu[\mu + (1 - \mu)\lambda_0]^{-1}, x, x \rangle, y \rangle \\ &= \langle \mu, x, \langle \lambda_0, x, y \rangle \rangle. \end{aligned}$$

Since  $\mu < \varepsilon$ ,  $\langle \lambda, x, y \rangle$  is in the  $\varepsilon$ -neighborhood of  $\langle \lambda_0, x, y \rangle$ .  $\square$

In view of Lemma 6.2, threshold jumps may only occur at extreme thresholds if (P2) and (P3) hold.

LEMMA 6.3. *If  $S$  satisfies (P3), then the function  $x \mapsto \langle \lambda, x, y \rangle$  is continuous for all  $y \in S, \lambda \in (0, 1]$ .*

*Proof.* Let  $\varepsilon > 0$  be given. We shall show that there exists a  $\delta > 0$  such that  $z \in N_x(\delta)$  implies that  $\langle \lambda, z, y \rangle \in N_{\langle \lambda, x, y \rangle}(\varepsilon)$ . Since  $\lambda(1 - \rho)[1 - \lambda\rho]^{-1} \rightarrow \lambda$  as  $\rho \rightarrow 0$ , by Lemma 6.2 there exists a  $\delta_1$  such that  $\rho < \delta$ , implies

$$\langle \lambda(1 - \rho)[1 - \lambda\rho]^{-1}, x, y \rangle = \langle \sigma, u, \langle \lambda, x, y \rangle \rangle,$$

for some  $u \in S, \sigma < \varepsilon/2$ . Choose  $\delta = \min(\varepsilon/2\lambda, \delta_1)$ . Then  $z \in N_x(\delta)$  implies  $z = \langle \rho, t, x \rangle, \rho < \delta, t \in S$ , and

$$\begin{aligned} \langle \lambda, z, y \rangle &= \langle \lambda, \langle \rho, t, x \rangle, y \rangle \\ &= \langle \lambda\rho, t, \langle \lambda(1 - \rho)(1 - \lambda\rho)^{-1}, x, y \rangle \rangle \\ &= \langle \lambda\rho, t, \langle \sigma, \mu, \langle \lambda, x, y \rangle \rangle \rangle \\ &= \langle \lambda\rho + (1 - \lambda\rho)\sigma, \langle \lambda\rho[\lambda\rho + (1 - \lambda\rho)\sigma]^{-1}, t, u \rangle, \langle \lambda, x, y \rangle \rangle, \end{aligned}$$

where

$$\lambda\rho + (1 - \lambda\rho)\sigma < \lambda\rho + \frac{(1 - \lambda\rho)\varepsilon}{2} < \lambda\rho + \frac{\varepsilon}{2} < \varepsilon,$$

which completes the proof.  $\square$

COROLLARY 6.4. *If  $S$  satisfies (P1) and (P3), then the function  $y \mapsto \langle \lambda, x, y \rangle$  is continuous for all  $x \in S, \lambda \in [0, 1]$ .*

We say that  $f \in S^*$  is *bounded* if there exists an  $M \geq 0$  such  $|f(x)| \leq M$ , for all  $x \in S$ .

THEOREM 6.5.  *$f \in S^*$  is bounded if and only if  $f$  is continuous in the natural topology.*

*Proof.* Suppose  $|f(x)| \leq M$ , for all  $x \in S$ . Let  $x_\alpha$  be a net converging to  $x \in S$  in the natural topology and let  $0 < \varepsilon \leq 1$ . Then there exists a  $\beta$  such that  $\alpha \geq \beta$  implies that  $x_\alpha \in N_x(\varepsilon)$ . Hence,  $x_\alpha = \langle \lambda_\alpha, y_\alpha, x \rangle$  (or  $x_\alpha = x$  and set  $\lambda_\alpha = 0$ ) for  $\alpha \geq \beta$ , where  $\lambda_\alpha < \varepsilon$ . We then obtain

$$f(x_\alpha) = \lambda_\alpha f(y_\alpha) + (1 - \lambda_\alpha)f(x),$$

and hence,

$$|f(x_\alpha) - f(x)| = \lambda_\alpha |f(y_\alpha) - f(x)| \leq 2M\lambda_\alpha < 2M\varepsilon.$$

It follows that  $f(x_\alpha) \rightarrow f(x)$ . Conversely, suppose that  $f \in S^*$  is not bounded. Then there exist  $x_n \in S$  such that  $|f(x_n)| > n^2, n = 1, 2, \dots$ . Now  $\langle n^{-1}, x_n, x \rangle \rightarrow x$ , but

$$|f(\langle n^{-1}, x_n, x \rangle) - f(x)| = |n^{-1}f(x_n) - n^{-1}f(x)| > n^{-1}|f(x_n)| - n^{-1}|f(x)| > n - n^{-1}|f(x)|.$$

Since the right-hand side of the inequality goes to  $\infty$  as  $n \rightarrow \infty$ ,  $f$  is not continuous at  $x$ .  $\square$

Let  $S_b^*$  denote the set of bounded affine functions and let  $\|f\| = \sup \{|f(x)| : x \in S\}$ , for  $f \in S_b^*$ . It is not hard to show that  $(S_b^*, \|\cdot\|)$  is a Banach space.

LEMMA 6.6. *If  $F \in Af(S_1, S_2)$ , then  $F$  is continuous for the natural topologies.*

*Proof.* Let  $x_\alpha \in S_1$  be a net converging to  $x \in S_1$  and let  $0 < \varepsilon \leq 1$ . Then there exists a  $\beta$  such that  $\alpha \geq \beta$  implies that  $x_\alpha \in N_x(\varepsilon)$ . Hence for  $\alpha \geq \beta$ ,  $x_\alpha \neq x$ , there exists a  $\lambda_\alpha < \varepsilon$ , and a  $y_\alpha \in S_1$  such that  $x_\alpha = \langle \lambda_\alpha, y_\alpha, x \rangle_1$ . Then  $F(x_\alpha) = \langle \lambda_\alpha, F(y_\alpha), F(x) \rangle_2$  and  $F(x_\alpha) \in N_{F(x)}(\varepsilon)$ . Hence,  $F(x_\alpha) \rightarrow F(x)$  in  $S_2$ .  $\square$

We use the notation  $Af(S) = Af(S, S)$ .

COROLLARY 6.7. *If  $F \in Af(S)$ , then  $F$  is continuous in the natural topology.*

We now extend the natural topology from  $S$  to  $S_+$ . For simplicity of notation we suppress the canonical embedding and look upon  $S$  as a base for the precone  $S_+$ . Hence, for any  $\theta \neq x \in S_+$ , there exist unique  $\alpha > 0$ ,  $s \in S$ , such that  $x = \alpha s$ , and  $\langle \lambda, s, t \rangle = \lambda s + (1 - \lambda)t$ , for all  $\lambda \in [0, 1]$ ,  $s, t \in S$ . For  $x = \alpha s \in S_+$ ,  $\alpha > 0$ ,  $s \in S$ , and  $0 < \varepsilon \leq 1$ , define the  $\varepsilon$ -neighborhood of  $x$  to be

$$N_x(\varepsilon) = \{\beta t \in S_+ : |\beta - \alpha| < \varepsilon, t \in N_s(\varepsilon)\}.$$

We also define

$$N_\theta(\varepsilon) = \{\beta t \in S_+ : |\beta| < \varepsilon\}.$$

Thus,  $\alpha_\delta s_\delta \rightarrow \alpha s$  if and only if  $\alpha_\delta \rightarrow \alpha$  and  $s_\delta \rightarrow s$  and  $\alpha_\delta s_\delta \rightarrow \theta$  if and only if  $\alpha_\delta \rightarrow 0$ . As before,  $\{N_x(\varepsilon)\}$  is a basis for a topology on  $S_+$  which we again call the *natural topology*.

If  $f \in S^*$ , define  $\hat{f}: S_+ \rightarrow R$  by  $\hat{f}(\alpha s) = \alpha f(s)$ .

LEMMA 6.8.  *$\hat{f}$  is the unique linear extension of  $f$  to  $S_+$ .*

*Proof.* Let  $\theta \neq x = \alpha s$ ,  $\theta \neq y = \beta t$ . Then

$$\begin{aligned} \hat{f}(\beta x) &= \hat{f}(\beta \alpha s) = \beta \alpha f(s) = \beta \hat{f}(\alpha s) = \beta \hat{f}(x) \\ \hat{f}(x + y) &= \hat{f}(\alpha s + \beta t) = \hat{f}[(\alpha + \beta)(\alpha(\alpha + \beta)^{-1}s + \beta(\alpha + \beta)^{-1}t)] \\ &= (\alpha + \beta)f[\alpha(\alpha + \beta)^{-1}s + \beta(\alpha + \beta)^{-1}t] \\ &= (\alpha + \beta)[\alpha(\alpha + \beta)^{-1}f(s) + \beta(\alpha + \beta)^{-1}f(t)] \\ &= \alpha f(s) + \beta f(t) = \hat{f}(\alpha s) + \hat{f}(\beta t) = \hat{f}(x) + \hat{f}(y). \end{aligned}$$

The other cases are left to the reader. For uniqueness, let  $g$  be a linear extension of  $f$  to  $S_+$ . Then

$$g(x) = g(\alpha s) = \alpha g(s) = \alpha f(s) = \hat{f}(\alpha s) = \hat{f}(x), \quad \text{for all } x \in S_+. \quad \square$$

If  $F \in Af(S)$ , define  $\hat{F}: S_+ \rightarrow S_+$  by  $\hat{F}(\alpha s) = \alpha \hat{F}(s)$ .

LEMMA 6.9.  *$\hat{F}$  is the unique linear extension of  $F$  to  $S_+$ .*

*Proof.* Similar to the proof of Lemma 6.8.  $\square$

Define the linear functional  $\hat{\tau}$  on  $S_+$  by  $\hat{\tau}(x) = \hat{\tau}(\alpha s) = \alpha$ . Denote the set of linear functionals on  $S_+$  by  $S_+^*$  and the set of linear maps from  $S_+$  to  $S_+$  by  $L(S_+)$ . We say that  $f \in S_+^*$  is *bounded* if there exists an  $M \geq 0$  such that  $|f(x)| \leq M\hat{\tau}(x)$ , for all  $x \in S_+$ . We say that  $F \in L(S_+)$  is bounded if there exists an  $M \geq 0$  such that  $\hat{\tau}[F(x)] \leq M\hat{\tau}(x)$ , for all  $x \in S_+$ . The next theorem is an extension of Theorem 6.5 and Lemma 6.6.

THEOREM 6.10.

(1) *If  $f \in S_+^*$ , then the following statements are equivalent: (a)  $f$  is continuous in the natural topology, (b)  $f$  is bounded, (c)  $f|S \in S_b^*$ .*

(2) *If  $F \in Af(S)$ , then  $\hat{F}$  is bounded.*

(3) If  $F \in L(S_+)$  and  $F : S \rightarrow S$ , then  $F$  is bounded.

(4) If  $F \in L(S_+)$ , then  $F$  is continuous in the natural topology if and only if  $F$  is bounded.

*Proof.* (1) If  $f$  is continuous, then  $f|S$  is continuous, so by Theorem 5.5 there exists an  $M \geq 0$  such that  $|f(s)| \leq M$ , for all  $s \in S$ . If  $\theta \neq x \in S_+$ , then  $[\hat{\tau}(x)]^{-1}x \in S$  so  $|f(x/\hat{\tau}(x))| \leq M$  and  $|f(x)| \leq M\hat{\tau}(x)$ . Hence (a) implies (b). That (b) implies (c) is trivial. If  $f|S \in S_b^*$ , then by Theorem 6.5,  $f|S$  is continuous. Now suppose  $\alpha_\delta s_\delta \rightarrow \alpha s$ . Then  $\alpha_\delta \rightarrow \alpha$  and  $s_\delta \rightarrow s$ . Hence,

$$f(\alpha_\delta s_\delta) = \alpha_\delta f(s_\delta) \rightarrow \alpha f(s),$$

so (c) implies (a).

(2) If  $F \in Af(S)$ , then for  $x = \alpha s = S_+$ , we have

$$\hat{\tau}[\hat{F}(x)] = \hat{\tau}[\hat{F}(\alpha s)] = \hat{\tau}[\alpha F(s)] = \alpha = \alpha \hat{\tau}(s) = \hat{\tau}(\alpha s) = \hat{\tau}(x).$$

(3) In this case  $F = (F|S)^\wedge$  and the result follows from (2).

(4) Define  $f(x) = \hat{\tau}[F(x)]$ . Then  $f \in S_+^*$  and the result follows from (1).  $\square$

The above results can also be extended to the prevector space  $V(S)$ . We leave the details of this to the reader.

**7. Observables and instruments.** In this section we begin a development of a generalized probability theory in the framework of convex prestructures. This theory may have applications in quantum mechanics [5] and possibly other fields such as mathematical economics [14].

Let  $S$  be a convex prestructure. In this section we think of  $S$  as a set of “states” describing a physical or mathematical system. For example, in traditional quantum mechanics [1], [7],  $S$  is the set of positive traceclass operators with trace one on a complex Hilbert space, and in probability theory,  $S$  is the set of probability measures on a probability space. In both these examples the convex structure is defined in terms of the usual linear operations.

Equip  $S^*$  with the usual pointwise order, and let  $\tau \in S^*$  be defined by  $\tau(x) = 1$ , for all  $x \in S$ . An *effect* is a function  $f \in S^*$  satisfying  $0 \leq f \leq \tau$ , where  $0 \in S^*$  is defined by  $0(x) = 0$  for every  $x \in S$ . If  $\mathcal{E}(S)$  denotes the set of effects, then  $\mathcal{E}(S)$  is a closed convex subset of the Banach space  $S_b^*$ . For  $f \in \mathcal{E}(S)$ , we interpret  $f(x)$  as the probability that the effect  $f$  is observed when the system is in state  $x$ . Notice that  $0$  and  $\tau$  are the effects which are never observed and always observed, respectively. In traditional quantum mechanics, an effect is given by an operator  $A$  satisfying  $0 \leq A \leq I$ , and if  $s$  is a state, then  $A(s) = \text{tr}(As)$ . In probability theory, an effect is given by a random variable  $g$  satisfying  $0 \leq g \leq 1$ , and if  $\mu$  is a state, then  $g(\mu) = \int g d\mu$ . Proponents of hidden variable theories for quantum mechanics claim that there are not enough effects to separate states [5]. If this is the case, as we have seen earlier,  $S$  cannot be a convex structure. Hence, convex prestructures might be a convenient framework in which to study hidden variable theories.

Denote the Borel subsets of  $R$  by  $B(R)$ . An *observable* is a map  $Q : B(R) \rightarrow \mathcal{E}(S)$  satisfying  $Q(R) = \tau$  and  $Q(\cup E_i)x = \sum Q(E_i)x$ , for any disjoint sequence  $E_i \in B(R)$  and any  $x \in S$ . In short, an observable is an effect-valued measure. We interpret  $Q(E)$  as the effect observed when  $Q$  has a value in  $E \in B(R)$ . In traditional quantum mechanics, a self-adjoint operator  $A$  gives the observable  $P^A$  where  $P^A$  is the resolution of the identity for  $A$ . In probability theory, a random variable  $g$  gives the observable  $E \rightarrow \chi_{g^{-1}(E)}$ .

Whereas the elements of  $S$  are interpreted as states, we think of the elements of  $S_+$  as “unnormalized” states. The set of *effects*  $\mathcal{E}(S_+)$  on  $S_+$  are defined to be the set of

functions  $f \in S_+^*$  satisfying  $\hat{0} \leq f \leq \hat{\tau}$ . The map  $f \rightarrow \hat{f}$  gives a natural bijection between  $\mathcal{E}(S)$  and  $\mathcal{E}(S_+)$ . We can also extend the concept of observable to  $S_+$ . An *observable* on  $S_+$  is a map  $Q: B(R) \rightarrow \mathcal{E}(S_+)$  such that  $Q(R) = \hat{\tau}$  and  $Q(\cup E_i)x = \sum Q(E_i)x$ , for any disjoint sequence  $E_i \in B(R)$  and any  $x \in S_+$ .

An *operation* is a map  $F \in L(S_+)$  satisfying  $\hat{\tau}[F(x)] \leq \hat{\tau}(x)$  for every  $x \in S_+$ . We denote the set of operations by  $\mathcal{O}(S_+)$ . Notice that elements of  $\mathcal{E}(S_+)$  and  $\mathcal{O}(S_+)$  are bounded and hence continuous in the natural topology (we assume in the sequel that (P2) holds). We denote the set of linear operators on the real linear space  $S_+^*$  by  $L(S_+^*)$ . If  $F \in L(S_+)$ , we define  $F^*: S_+^* \rightarrow S_+^*$  by  $[F^*(f)](x) = f[F(x)]$ . We then have  $F^* \in L(S_+^*)$ . If  $F \in \mathcal{O}(S_+)$ , the effect  $f$  associated with  $F$  is defined as  $f = F^*\hat{\tau}$ . Notice that  $f \in \mathcal{E}(S_+)$  since

$$0 \leq (F^*\hat{\tau})(x) = \hat{\tau}[F(x)] \leq \hat{\tau}(x), \quad \text{for all } x \in S_+.$$

Elements of  $\mathcal{O}(S_+)$  are thought of as conditioning operations in the following sense. If  $x \in S_+$ ,  $F \in \mathcal{O}(S_+)$ , then  $F(x)$  is the (unnormalized) state conditioned by an observation of the effect associated with  $F$ . In other words, if  $x$  is the original state of the system and the effect associated with  $F$  is observed, then the resulting state is  $F(x)$ .

Although every operation is associated with a unique effect, an effect may have many operations with which it is associated. For example, suppose (P3) holds and  $f \in \mathcal{E}(S_+)$ . Let  $s_0 \in S$  and define  $F: S_+ \rightarrow S_+$  by  $F(x) = f(x)s_0$ . Then  $F \in \mathcal{O}(S_+)$  since  $F \in L(S_+)$  and

$$\hat{\tau}[F(x)] = \hat{\tau}[f(x)s_0] = f(x) \leq \hat{\tau}(x).$$

Moreover,  $f$  is associated with  $F$  since

$$(F^*\hat{\tau})(x) = \hat{\tau}[F(x)] = \hat{\tau}[f(x)s_0] = f(x), \quad \text{for all } x \in S_+.$$

An *instrument* is a map  $\mathcal{J}: B(R) \rightarrow L(S_+)$  such that  $\hat{\tau}[\mathcal{J}(R)x] = \hat{\tau}(x)$ , for every  $x \in S_+$  and  $\mathcal{J}(\cup E_i)x = \sum \mathcal{J}(E_i)x$ , for any sequence of disjoint  $E_i \in B(R)$  and every  $x \in S_+$  where convergence of the sum is in the natural topology. Notice that if  $\mathcal{J}$  is an instrument, then  $\mathcal{J}(E) \in \mathcal{O}(S_+)$ , for every  $E \in B(R)$ . Indeed, if  $E'$  is the complement of  $E$ ,

$$\hat{\tau}[\mathcal{J}(E)x] \leq \hat{\tau}[\mathcal{J}(E)x] + \hat{\tau}[\mathcal{J}(E')x] = \hat{\tau}[\mathcal{J}(R)x] = \hat{\tau}(x).$$

Thus, an instrument is an operation-valued measure. We interpret  $\mathcal{J}(E)$  as the operation resulting from a measurement with the instrument  $\mathcal{J}$  giving a value in  $E$ .

The observable  $Q$  associated with an instrument  $\mathcal{J}$  is defined by  $Q(E) = \mathcal{J}(E)^*\hat{\tau}$  for all  $E \in B(R)$ . To show that  $Q$  is an observable, note that  $Q(E) \in \mathcal{E}(S_+)$  and

$$Q(R)x = [\mathcal{J}(R)^*\hat{\tau}](x) = \hat{\tau}[\mathcal{J}(R)x] = \hat{\tau}(x),$$

so  $Q(R) = \hat{\tau}$ . Finally, if  $E_i \in B(R)$  are disjoint,

$$\begin{aligned} Q(\cup E_i)x &= [\mathcal{J}(\cup E_i)^*\hat{\tau}](x) = \hat{\tau}[\mathcal{J}(\cup E_i)x] \\ &= \hat{\tau}[\sum \mathcal{J}(E_i)x] = \sum \hat{\tau}[\mathcal{J}(E_i)x] \\ &= \sum [\mathcal{J}(E_i)^*\hat{\tau}](x) = \sum Q(E_i)x. \end{aligned}$$

Notice that  $Q$  is the unique observable satisfying  $Q(E)x = \hat{\tau}[\mathcal{J}(E)x]$ , for all  $E \in B(R)$ ,  $x \in S_+$ .

An observable  $Q$  may be associated with many instruments. For example, suppose (P3) holds, let  $s_0 \in S$  and define  $\mathcal{J}(E)x = [Q(E)x]s_0$ . Then  $\mathcal{J}(E) \in L(S_+)$ , for all  $E \in$

$B(R)$  and

$$\hat{\tau}[\mathcal{F}(R)x] = \hat{\tau}[[Q(R)x]s_0] = \hat{\tau}[\hat{\tau}(x)s_0] = \hat{\tau}(x).$$

Moreover, if  $E_i \in B(R)$  are disjoint and  $x \in S_+$ , then

$$\mathcal{F}(\cup E_i)x = [Q(\cup E_i)x]s_0 = \{\sum [Q(E_i)x]\}s_0 = \sum [Q(E_i)x]s_0 = \sum \mathcal{F}(E_i)x.$$

Hence,  $\mathcal{F}$  is an instrument, and  $Q$  is associated with  $\mathcal{F}$  since

$$\hat{\tau}[\mathcal{F}(E)x] = \hat{\tau}[[Q(E)x]s_0] = Q(E)x.$$

If  $F, G \in \mathcal{O}(S_+)$ , then  $FG \in \mathcal{O}(S_+)$  since

$$\hat{\tau}[FG(x)] \leq \hat{\tau}[G(x)] < \hat{\tau}(x).$$

We thus see that  $\mathcal{O}(S_+)$  is a semigroup. If  $\mathcal{F}, \mathcal{G}$  are instruments on  $B(R)$  and there exists an instrument  $\mathcal{H}$  on  $B(R^2)$  such that  $\mathcal{H}(E \times F) = \mathcal{F}(E)\mathcal{G}(F)$ , for every  $E, F \in B(R)$ , then  $\mathcal{H}$  is called the *composition of  $\mathcal{F}$  following  $\mathcal{G}$*  and is denoted  $\mathcal{F} \circ \mathcal{G}$ .

Let  $\mathcal{F}, \mathcal{G}$  be instruments with associated observables  $Q, P$ , respectively. The observable  $E \mapsto \mathcal{F}(R)^*[Q(E)]$  is called the observable  $Q$  *conditioned by the measurement of  $P$  with instrument  $\mathcal{F}$* . If  $\mathcal{F} \circ \mathcal{G}$  exists, then the observable  $T$  based on  $B(R^2)$  defined by  $T(\Delta) = [\mathcal{F} \circ \mathcal{G}(\Delta)]^*(\hat{\tau})$ ,  $\Delta \in B(R^2)$ , is called the *joint distribution of  $\mathcal{F}$  following  $\mathcal{G}$* . The following lemma shows that the joint distribution has the correct marginal distributions.

LEMMA 7.1.  $T(R \times E) = P(E)$ ,  $T(E \times R) = \mathcal{F}(R)^*[Q(E)]$ .

*Proof.* For any  $x \in S_+$  we have

$$\begin{aligned} T(R \times E)x &= \{[\mathcal{F} \circ \mathcal{G}(R \times E)]^* \hat{\tau}\}x \\ &= \hat{\tau}\{[\mathcal{F} \circ \mathcal{G}(R \times E)]x\} = \hat{\tau}[\mathcal{F}(R)\mathcal{G}(E)x] \\ &= \hat{\tau}[\mathcal{F}(E)x] = [\mathcal{F}(E)^* \hat{\tau}]x = P(E)x, \end{aligned}$$

and

$$\begin{aligned} T(E \times R)x &= \hat{\tau}[\mathcal{F}(E)\mathcal{G}(R)x] = \{[\mathcal{F}(E)\mathcal{G}(R)]^* \hat{\tau}\}x \\ &= [\mathcal{F}(R)^* \mathcal{F}(E)^* \hat{\tau}]x = \mathcal{F}(R)^*Q(E)x, \end{aligned}$$

so the results follow.  $\square$

The above is just the beginning of a probability theory on convex prestructures. This work is a generalization of the ‘‘operational’’ approach to quantum mechanics [1], [2], [8], [9].

REFERENCES

[1] E. DAVIES, *Quantum Theory of Open Systems*, Academic Press, New York, 1976.  
 [2] E. DAVIES AND J. LEWIS, *An operational approach to quantum probability*, Comm. Math. Phys., 17 (1970), pp. 239–260.  
 [3] S. GUDDER, *Convex structures and operational quantum mechanics*, Comm. Math. Phys., 29 (1973), pp. 249–264.  
 [4] ———, *Convexity and mixtures*, SIAM Rev., 19 (1977), pp. 221–240.  
 [5] ———, *Stochastic Methods in Quantum Mechanics*, Elsevier North-Holland, New York, 1979.  
 [6] M. HAUSNER, *Multidimensional utilities*, in Decision Processes, R. M. Thrall, C H. Coombs, and R L. Davis, eds., John Wiley, New York, 1954.  
 [7] J. JAUCH, *Foundations of Quantum Mechanics*, Addison-Wesley, Reading MA, 1968.  
 [8] B. MIELNIK, *Geometry of quantum states*, Comm. Math. Phys., 9 (1968), pp. 55–80.  
 [9] ———, *Theory of filters*, Comm. Math. Phys., 15 (1969), pp. 1–46.

- [10] M. RUSIN, *The structure of nonlinear blending models*, Chem. Eng. Sci., 30 (1975), pp. 937-944.
- [11] ———, *A new method for representing non-linear blending problems in a linear programming format*, preprint, American Petroleum Institute, Washington, D.C., 1978.
- [12] M. STONE, *Postulates for a barycentric calculus*, Ann. of Math., 29 (1949), pp. 25-30.
- [13] R. THRALL, *Applications of multidimensional utility*, in Decision Processes, R. M. Thrall, C. H. Coombs, and R. L. Davis, eds., John Wiley, New York, 1954.
- [14] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton NJ, 1944.

## A REMARK ON EQUATIONS $y'' + p_1y' + p_2y = 0$ AT AN IRREGULAR SINGULAR POINT\*

J. F. COLOMBEAU† AND A. MERIL†

**Abstract.** We extend some classical results on the solutions of the differential equation  $y'' + p_1y' + p_2y = 0$  at infinity in a sector of the complex plane to a much more general class of differential equations still of the above type. The proof uses "scales of Banach spaces" and is inspired by that of the linear "Ovcyannikov theorem".

**1. Introduction.** The origin of this work lies in a study of the existence and asymptotic expansions of solutions of systems of partial differential equations of the Cauchy-Kovalewsky type in the neighborhood of singular points. We had the idea to combine both the classical study of systems of ordinary differential equations at an irregular singular point and the theorems of existence of the Ovcyannikov type. We show here in a rather general setting that this combination is indeed possible in some cases and in this way yields new results of existence of solutions that are represented in a natural way by asymptotic expansions that one may easily compute. In order to make the paper shorter we give only one theorem which appears as a generalization of a classical theorem of Hoheisel [3] on the solutions of the scalar equation  $y'' + p_1(z)y' + p_2(z)y = 0$  in a sector of the complex plane and for  $z$  tending to infinity.

There are a number of works starting from the Ovcyannikov theorem and its applications to Cauchy problems at ordinary points of differential equations valued in a scale of Banach spaces (Ovcyannikov [4], Trèves [5] [6]) then applying this method for regular singular points (an example of this is in Baouendi-Goulaouic [1]). In this paper we consider the more complicated case of irregular singular points.

For this we chose the method of Hoheisel [3] because it is directly adaptable to combining with the Ovcyannikov type majorizations. The closely related method of Erdelyi [2] (in the real case) would fit as well. Our proof relies upon the convergence of some series which give the solution, hence we approximate it with explicit majorizations.

**2. Notation.** We only need some classical definitions, but we prefer to recall them. A scale of Banach spaces  $E = \bigcup_{0 < \alpha < 1} E_\alpha$  is a family  $(E_\alpha)_{0 < \alpha < 1}$  of Banach spaces such that, if  $\alpha < \alpha'$ ,  $E_{\alpha'}$  is contained in  $E_\alpha$  with (natural) injection of norm  $\leq 1$ . A classical example (Trèves [6]) is the usual scale of germs of holomorphic functions at a compact set  $K$  of  $\mathbb{C}^n$ .  $E$  is obviously a vector space, and we denote by  $l(E)$  the vector space of all linear maps from  $E$  into  $E$ . We denote by  $L(E_{\alpha'}, E_\alpha)$  the Banach space of linear continuous maps from  $E_{\alpha'}$  to  $E_\alpha$  equipped with its usual strong norm denoted by  $\|\cdot\|_{L(E_{\alpha'}, E_\alpha)}$ . If  $l$  is in  $l(E)$  we denote by  $l|_{E_\alpha}$  the restriction of  $l$  to  $E_\alpha \subset E$ . We shall consider functions valued in a Banach space and then the asymptotic expansions are defined exactly as usual in the scalar case (Wasow [7]).

**3. Statement of the abstract result.** Let  $E$  be a scale of Banach spaces and  $S$  be the sector of the complex plane defined by  $|z| > r_0 > 0$  and  $\varphi_1 < \arg z < \varphi_2$ .

We consider the  $E$ -valued differential equation

$$(*) \quad y''(z) + P_1(z)y'(z) + P_2(z)y(z) = 0,$$

\* Received by the editors June 29, 1979, and in revised form February 5, 1980.

† Université de Bordeaux I, U.E.R. de Mathématiques et d'Informatique, 33405 Talence, France.

where if  $\nu = 1, 2$ ,  $P_\nu$  maps  $S$  into  $l(E)$  and has the following form:

$$P_\nu(z) = a_{\nu,0} \text{id} + z^{-1} a_{\nu,1} \text{id} + z^{-2} L_\nu(z),$$

(here  $\text{id}$  is the identity map on  $E$ ,  $a_{\nu,0}$  and  $a_{\nu,1}$  are complex numbers, and  $L_\nu(z)$  is in  $l(E)$ ).

We assume that  $(a_{1,0})^2 \neq 4a_{2,0}$  and that, for every  $\alpha \in ]0, 1[$  and  $\varepsilon \in ]0, \alpha[$ , every  $z \in S$ , we have

$$L_\nu(z)_{/E_\alpha} \in L(E_\alpha, E_{\alpha-\varepsilon}),$$

and

$$\|L_\nu(z)_{/E_\alpha}\|_{L(E_\alpha, E_{\alpha-\varepsilon})} \leq \frac{C}{\varepsilon},$$

where  $C$  is a constant independent of  $\alpha$ ,  $\varepsilon$  and  $z$ . Furthermore, we assume that the map  $z \rightarrow L_\nu(z)$  from  $S$  to  $L(E_\alpha, E_{\alpha-\varepsilon})$  is holomorphic, for every  $\alpha > 0$ , and every  $0 < \varepsilon < \alpha$ .

Let  $\Sigma$  be any subsector of  $S$  defined by  $\Psi_1 < \arg z < \Psi_2$  with  $0 < \Psi_2 - \Psi_1 < \pi$ . Let  $\alpha_0$ ,  $\alpha$  be given with  $0 < \alpha < \alpha_0 < 1$ .

**THEOREM.** *There is a number  $R > 0$  and complex numbers  $\sigma, \rho$  such that for every given element  $\chi_0$  of  $E_{\alpha_0}$  there exists a solution  $y$  of (\*) defined in  $\Sigma$  for  $|z| > R$ , valued in  $E_\alpha$  and such that,*

$$\lim_{\substack{z \rightarrow \infty \\ z \in \Sigma}} e^{(a_{1,0}/2)z} \cdot e^{-\sigma z} \cdot z^{-\rho} y(z) = \chi_0.$$

*Remark 1.*  $R$  is explicitly computed in the proof: there exist  $R_1 > 0$  and  $M > 0$  such that  $R = \max(R_1, Me/(\alpha_0 - \alpha))$  where  $\log e = 1$ .  $\sigma$  is defined by:

$$\sigma^2 = \frac{(a_{1,0})^2}{4} - a_{2,0}, \quad \frac{\pi - \Psi_1 + \Psi_2}{2} \leq \arg \sigma \leq \frac{3\pi - \Psi_1 + \Psi_2}{2},$$

(eventually 2 solutions),

$$\rho = -\frac{a_{1,1}}{2} - \frac{2a_{2,1} - a_{1,0} \cdot a_{1,1}}{4\sigma}.$$

*Remark 2.* If we assume that the functions  $L_\nu$  admit asymptotic expansions as  $z \rightarrow \infty$ ,  $z \in S$ , then  $y$  admits an asymptotic expansion as  $z \rightarrow \infty$ ,  $z \in \Sigma$ , that may be (formally) computed easily according to classical calculations (see Erdelyi [2], Wasow [7]).

*Remark 3.* When the spaces  $E_\alpha$  are Banach algebras we may find a second type of solutions according to Hoheisel [3].

**4. Motivations and applications.** In [1], Baouendi and Goulaouic study partial differential operators of the type:

$$t^k D_t^m + \sum_{p < m} \sum_{|\beta| \leq m-p} c_{p,\beta}(x, t) D_t^p D_x^\beta,$$

with the usual notations:  $t \in \mathbb{R}$ ;  $n, k, m$  and  $p \in \mathbb{N}$ ;  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ ;

$$\beta = (\beta_1, \dots, \beta_n) \in \mathbb{N}^n; \quad |\beta| = \beta_1 + \dots + \beta_n; \quad D_t^p = \frac{\partial^p}{\partial t^p}; \quad D_x^\beta = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \dots \partial x_n^{\beta_n}},$$

where the functions  $c_{p,\beta}$  are smooth enough (analytic for example) and satisfy various other assumptions (see [1]) and where a fundamental assumption is  $k \leq m$ . These



authors give some generalizations of the Cauchy-Kovalewsky theorem in the case of the above operators and with initial data on a characteristic hypersurface.

A motivation of our abstract theorem was to study the case  $k > m$  using a technique of “irregular singular point”. Of course, as it is well known for ordinary differential equations, we obtain in this case results of a different kind (instead of solutions defined in a neighborhood of a point we obtain solutions in a bounded sector and having certain asymptotic expansions). We restrict ourselves to the case  $m = 2$  in the sequel. We obtain the following results.

Let the following partial differential operator be given (in which  $k \in \mathbb{R}$  and  $k > 2$ ):

$$\mathcal{P}_k = t^k D_t^2 + \sum_{\rho < 2} \sum_{|\beta| \leq 1} c_{\rho, \beta}(x, t) D_x^\beta D_t^\rho,$$

where  $t$  ranges over a complex sector

$$S = \left\{ t \in \mathbb{C} \text{ such that } \Psi_1 < \arg t < \Psi_2 \text{ with } 0 < \Psi_2 - \Psi_1 < \min \left( \frac{2\pi}{k-2}, 2\pi \right) \text{ and } |t| < r \right\},$$

where  $x$  ranges over a bounded open subset  $V$  of  $\mathbb{C}^n$  and where the functions  $c_{\rho, \beta}$  are holomorphic in  $V \times S$ .

Such an operator can be written (choosing a determination of the logarithm in  $S$ ):

$$\begin{aligned} \mathcal{P}_k = t^k D_t^2 + (c_1 t^{k/2} + c_2 t^{k-1}) D_t + c_3 + c_4 t^{[(k/2)-1]} + \sum_{|\beta| \leq 1} t^{[(3/2)k-2]} b_{1, \beta}(x, t) D_x^\beta D_t \\ + \sum_{|\beta| \leq 1} t^{k-2} b_{2, \beta}(x, t) D_x^\beta. \end{aligned}$$

We assume that there is a choice of complex numbers  $c_i$  for which  $(c_1)^2 - 4c_3 \neq 0$  and for which the functions  $b_{1, \beta}$  and  $b_{2, \beta}$  are bounded on  $V \times S$ .

We define two complex numbers  $\sigma$  and  $\rho$  by:

$$\sigma^2 = \frac{(c_1)^2 - 4c_3}{(2-k)^2}, \quad \frac{\pi}{2} + \frac{(k-2)}{4} (\Psi_1 + \Psi_2) \leq \arg \sigma \leq \frac{3\pi}{2} + \frac{(k-2)}{4} (\Psi_1 + \Psi_2),$$

and

$$\rho = \frac{k - 2c_2}{2(2-k)} + \frac{2c_1 c_2 - k c_1 - 4c_4}{2\sigma(2-k)^2}.$$

**THEOREM.** *There exists  $R > 0$  (small enough) such that if  $\Omega$  is a bounded open subset of  $\mathbb{C}^n$  with  $\bar{\Omega} \subset V$  and if  $f$  is a holomorphic function on  $V$  there exists a solution  $u(x, t)$  of*

$$\mathcal{P}_k u(x, t) = 0,$$

*holomorphic if  $t \in S$ ,  $|t| < R$  and  $x \in \Omega$ , such that:*

$$\lim_{\substack{t \rightarrow 0 \\ t \in S}} e^{[(c_1/(2-k)) - \sigma] t^{(-k/2)+1}} t^{\rho[(k/2)-1]} u(x, t) = f(x),$$

*uniformly in  $x \in \Omega$ .*

*If furthermore the functions  $t \rightarrow b_{\nu, \beta}(x, t)$  ( $\nu = 1, 2$ ) admit asymptotic expansions when  $t \rightarrow 0$ ,  $t \in S$  (with uniform majorizations in  $x \in V$ ) then this solution admits an asymptotic expansion if  $t \rightarrow 0$ ,  $t \in S$ ,*

$$u(x, t) \sim e^{[\sigma - c_1/(2-k)] t^{(-k/2)+1}} t^{-\rho[(k/2)-1]} \left[ f(x) + \sum_{n \geq 1} f_n(x) t^{n[(k/2)-1]} \right],$$

(with uniform majorizations in  $x \in \Omega$ ) and where the computation of the functions  $f_n$  is easy (and classical—see the proof).

*Remark 1.* We may obviously restrict ourselves to  $x \in \mathbb{R}^n$  and functions  $c_{p,\beta}$  which are analytic in  $x$ . But if  $t$  is real our abstract theorem cannot be applied in the form given in § 3. Nevertheless, using the proof given in Erdelyi [2] in place of that of Hoheisel [3] (in fact the proof in [2] is an adaptation to the real case of the proof in [3]) the interested reader may adapt the proof of the abstract theorem to the real case and the functions  $c_{p,\beta}$  may be  $C^\infty$  in  $t$  (and analytic in  $x$ ).

*Proof of the theorem.* Let  $\Omega_s = \bigcup_{x \in \Omega} B(x, s)$  where

$$B(x, s) = \left\{ \zeta \in \mathbb{C}^n \text{ such that } \sup_{1 \leq i \leq n} |x_i - \zeta_i| < s \right\}.$$

We denote by  $E_s(\Omega)$  the Banach space of the continuous functions on the closure  $\bar{\Omega}_s$  of  $\Omega_s$ , which are holomorphic in  $\Omega_s$ , and we equip this space with the norm

$$\|f\|_s = \sup_{x \in \bar{\Omega}_s} |f(x)|.$$

If  $s_0 > 0$  is such that  $\bar{\Omega}_{s_0} \subset V$  we consider the scale of Banach spaces  $(E_s(\Omega))_{0 < s < s_0}$ . We set  $z = t^{(-k/2)+1}$  and  $y(z, x) = u(t, x)$ . The equation  $\mathcal{P}_k u(x, t) = 0$  becomes:

$$\left[ \frac{\partial^2}{\partial z^2} + \left\{ \frac{2c_1}{2-k} + \frac{2c_2-k}{2-k} \frac{1}{z} + \frac{1}{z^2} \frac{2}{2-k} \sum_{|\beta| \leq 1} b_{1,\beta}(x, z^{2/(2-k)}) D_x^\beta \right\} \frac{\partial}{\partial z} + \frac{4c_3}{(2-k)^2} + \frac{4c_4}{(2-k)^2} \frac{1}{z} + \frac{1}{z^2} \frac{4}{(2-k)^2} \sum_{|\beta| \leq 1} b_{2,\beta}(x, z^{2/(2-k)}) D_x^\beta \right] y(z, x) = 0.$$

We set

$$a_{1,0} = \frac{2c_1}{2-k}, \quad a_{1,1} = \frac{2c_{2-k}}{2-k},$$

$$a_{2,0} = \frac{4c_3}{(2-k)^2}, \quad a_{2,1} = \frac{4c_4}{(2-k)^2},$$

$$L_\nu(z) = \left( \frac{2}{2-k} \right)^\nu \sum_{|\beta| \leq 1} b_{\nu,\beta}(x, z^{2/(2-k)}) D_x^\beta, \quad \text{if } \nu = 1, 2.$$

It suffices now to apply the abstract theorem (the majorizations on  $L_\nu(z)$  come from Cauchy's inequalities: Lemma 1 of [1]).

*Remark.* We may also work with the dual scale (if  $\Omega$  is convex balanced  $\mathcal{H}(\Omega_s)$  is dense in  $\mathcal{H}(\Omega_{s'})$  for  $s > s'$ ; if not, do like [1, p. 459]). In this last case  $f_0$  is an element of  $\mathcal{H}'(V)$ ; in particular  $f_0$  may be a  $C^\infty$  function on  $V$ , and for each  $t$ ,  $u(x, t)$  is no longer a function in the variable  $x$  but an analytic functional on  $\Omega$  (i.e.,  $u(\cdot, t) \in \mathcal{H}'(\Omega)$ ). This method is developed in Trèves [6].

**5. Proof of the theorem.** If  $\sigma$  is a complex square root of  $((a_{1,0})^2/4) - a_{2,0}$ , the line  $\text{Re } \sigma z = 0$  is called a critical line. After a rotation in the variable  $z$  one may assume that the half-line  $\left\{ \lim_{z=0}^{\text{Re } z \leq 0} \right\}$  is in  $\Sigma$ , that  $\Sigma$  is contained in the set  $\{z \in S \text{ such that } \pi/2 < \Psi'_1 < \arg z < \Psi'_2 < 3\pi/2\}$  and that  $\text{Re } z = 0$  is not a critical line.

Now let  $\sigma$  be defined by

$$(0) \quad \begin{cases} \sigma^2 = \frac{(a_{1,0})^2}{4} - a_{2,0}, \\ \sigma' = \operatorname{Re} \sigma > 0. \end{cases}$$

We do the change of unknown function

$$y(z) = e^{-(a_{1,0}/2)z} \cdot \bar{y}(z).$$

Then (\*) becomes

$$(1) \quad \bar{y}''(z) + (a_{1,1}z^{-1} \operatorname{id} + z^{-2}L_1(z))\bar{y}'(z) + G(z)\bar{y}(z) = 0,$$

with

$$G(z) = \left( -\frac{(a_{1,0})^2}{4} + a_{2,0} \right) \operatorname{id} + (a_{2,1} - \frac{1}{2}a_{1,0} \cdot a_{1,1})z^{-1} \operatorname{id} + z^{-2}L_2(z) - \frac{a_{1,0}}{2} z^{-2}L_1(z);$$

hence, with obvious new notations, (1) becomes:

$$(1') \quad \bar{y}''(z) + [c_{1,1}z^{-1} \operatorname{id} + z^{-2}L_1(z)]\bar{y}'(z) + [c_{2,0} \operatorname{id} + c_{2,1}z^{-1} \operatorname{id} + z^{-2}L_2'(z)]\bar{y}(z) = 0.$$

We set

$$(1'') \quad \rho = -\frac{c_{1,1}}{2} - \frac{c_{2,1}}{2\sigma},$$

and

$$b = -\frac{c_{2,1}}{\sigma}.$$

Setting  $\bar{y}(z) = e^{\sigma z}(z)^\rho v(z)$  we obtain from (1'') the new equation

$$(2) \quad v''(z) + (2\sigma + z^{-1}b)v'(z) = z^{-2}E(v),$$

with

$$-E(v) = L_1(z)v'(z) + \bar{L}_2(z)v(z),$$

and

$$\bar{L}_2(z) = [\rho(\rho - 1) + \rho c_{1,1}] \operatorname{id} + (\sigma + \rho z^{-1})L_1(z) + L_2'(z).$$

If  $\chi$  is a function of  $z \in \Sigma$  valued in  $E_\alpha$  we set (when possible)

$$(3) \quad (T\chi)(z) = \int_{-\infty}^z e^{-2\sigma\xi}\xi^{-b} \int_{-\infty}^\xi e^{2\sigma\lambda}\lambda^{b-2}[E(\chi)](\lambda) \, d\lambda \, d\xi.$$

Then we have the following lemma.

LEMMA 1. *Let  $\chi$  be a holomorphic function from  $\Sigma$  to  $E_\alpha$  such that for some  $A > 0$  and some  $n \in \mathbb{N}^*$  (let  $\|\cdot\|_\alpha$  denote the norm in  $E_\alpha$ )*

$$(4) \quad \begin{cases} \|\chi(z)\|_\alpha \leq A|z|^{-(n-1)}, \\ \|\chi'(z)\|_\alpha \leq A|z|^{-(n-1)}, \\ \|\chi''(z)\|_\alpha \leq A(n-1)|z|^{-(n-1)}. \end{cases}$$

Then for some  $R_1 > 0$  (large enough and independent of  $\chi$ )  $T\chi$  is defined for  $|z| > R_1$  and

$z \in \Sigma$ . Furthermore for every  $\varepsilon \in ]0, \alpha[$ ,

$$(5) \quad \begin{cases} \|(T\chi)(z)\|_{\alpha-\varepsilon} \leq \frac{AM|z|^{-n}}{n\varepsilon}, \\ \|(T\chi)'(z)\|_{\alpha-\varepsilon} \leq \frac{AM|z|^{-n}}{n\varepsilon}, \\ \|(T\chi)''(z)\|_{\alpha-\varepsilon} \leq \frac{AM|z|^{-n}}{\varepsilon}, \end{cases}$$

for some constant  $M > 0$  independent of  $\chi, A, \alpha, \varepsilon, n$ .

*Proof.* From the assumptions on  $L_1$  and  $L_2$ ,  $E(\chi)$  is holomorphic from  $\Sigma$  to  $E_{\alpha-\varepsilon}$  and

$$(6) \quad \|(E(\chi))(z)\|_{\alpha-\varepsilon} \leq \frac{CA}{\varepsilon} |z|^{-(n-1)}, \quad \text{for some constant } C > 0.$$

We set

$$(7) \quad \varphi(\xi) = \int_{-\infty}^{\xi} e^{2\sigma\lambda} \lambda^{b-2} E(\chi)(\lambda) d\lambda,$$

with the path of integration given by the union of two straight line segments,  $[-\infty, a] \cup [a, \xi]$ , where  $a$  is a (negative) real number in  $\Sigma$  and  $-\infty$  is the point at infinity of the negative half real line in  $\mathbb{C}$ .

From (0) (6), Cauchy's integral theorem and if  $\xi = \xi' + i\xi''$ ,  $\xi', \xi'' \in \mathbb{R}$ , one has

$$\varphi(\xi) = \int_{-\infty+i\xi''}^{\xi'+i\xi''} e^{2\sigma\lambda} \lambda^{b-2} (E(\chi))(\lambda) d\lambda,$$

and

$$(8) \quad \|\varphi(\xi)\|_{\alpha-\varepsilon} \leq \frac{CA}{\varepsilon} e^{-2\sigma''\xi''} J(\xi'),$$

with

$$(9) \quad J(\xi') = \int_{-\infty}^{\xi'} e^{2\sigma'\lambda'} (\lambda'^2 + \xi''^2)^{(b'-(n+1))/2} d\lambda',$$

where  $b = b' + ib''$ ,  $b', b'' \in \mathbb{R}$ .

The function  $\lambda' \rightarrow e^{2\sigma'\lambda'} (\lambda'^2 + \xi''^2)^{b'/2}$  has its derivative of the sign of  $\lambda'^2 + (b'/2\sigma')\lambda' + \xi''^2$  which is positive if  $|\xi''| \geq |b'|/(4\sigma')$ ; if  $|\xi''| < |b'|/(4\sigma')$ , there is some  $x_0 > 0$  such that  $\lambda'^2 + (b'/2\sigma')\lambda' + \xi''^2 \geq 0$ , if  $\lambda' < -x_0$ . We set  $R_1 = (x_0^2 + b'^2/4\sigma'^2)^{1/2}$ . Every point  $z = \lambda' + i\xi''$ ,  $\lambda', \xi'' \in \mathbb{R}$ , in the subsector  $|z| > R_1$ , of  $\Sigma$  has the property that  $|\xi''| \geq |b'|/4\sigma'$ , or  $\lambda' < -x_0$ .

From now on we only work in  $\Sigma$  for  $|z| > R_1$ , hence

$$(10) \quad J(\xi') \leq e^{2\sigma'\xi'} (\xi'^2 + \xi''^2)^{b'/2} \int_{-\infty}^{\xi'} (\lambda'^2 + \xi''^2)^{-(n+1)/2} d\lambda'.$$

Now the fact that  $\Sigma \subset \{z \in S \text{ such that } \pi/2 < \Psi'_1 < \arg z < \Psi'_2 < 3\pi/2\}$  and a straightforward calculation give:

$$(11) \quad J(\xi') \leq e^{2\sigma'\xi'} (\xi'^2 + \xi''^2)^{(b'-n)/2} \frac{C_1}{n}, \quad C_1 \text{ constant } > 0.$$

Hence from (8)

$$(12) \quad \|\varphi(\xi)\|_{\alpha-\varepsilon} \leq \frac{C_1 CA}{\varepsilon n} e^{2\sigma'\xi'-2\sigma''\xi''} (\xi'^2 + \xi''2)^{(b'-n)/2}.$$

Since

$$(12') \quad (T\chi)'(\xi) = e^{-2\sigma\xi}\xi^{-b}\varphi(\xi),$$

we have

$$(13) \quad \|(T\chi)'(\xi)\|_{\alpha-\varepsilon} \leq \frac{AM}{n\varepsilon} |\xi|^{-n},$$

and from (2), (6)  $\|(T\chi)''(z)\|_{\alpha-\varepsilon} \leq (AM/\varepsilon)|z|^{-n}$ , for a new larger constant  $M$ , which give the two last inequalities of Lemma 1.

(9) may be rewritten

$$J(\xi') = \int_{-\infty}^{\xi'} e^{\sigma'\lambda'} (\lambda'^2 + \xi''2)^{[b'-(n+1)]/2} e^{\sigma'\lambda'} d\lambda'.$$

As before we have

$$J(\xi') \leq \frac{e^{2\sigma'\xi'}}{\sigma'} (\xi'^2 + \xi''2)^{[b'-(n+1)]/2},$$

so that

$$\|\varphi(\xi)\|_{\alpha-\varepsilon} \leq \frac{CA e^{2\sigma'\xi'} e^{-2\sigma''\xi''}}{\varepsilon\sigma'} (\xi'^2 + \xi''2)^{[b'-(n+1)]/2},$$

hence from (12')  $\|(T\chi)'(z)\|_{\alpha-\varepsilon} \leq (AM/\varepsilon)|z|^{-n-1}$  (a new larger constant  $M$ ). Now  $(T\chi)(z) = \int_{-\infty}^z (T\chi)'(\xi) d\xi$  where the integration may be done on the half-line  $\rho e^{i \arg z}$ , hence

$$\|(T\chi)(z)\|_{\alpha-\varepsilon} \leq \frac{AM}{n\varepsilon} |z|^{-n}. \quad \square$$

Let  $\chi_0$  be a given element of  $E_{\alpha_0}$ ; we shall define by induction a sequence  $(\chi_n)_{n \in \mathbb{N}}$  of functions by

$$\begin{aligned} \chi_0(z) &= \chi_0, \\ \chi_{n+1}(z) &= (T\chi_n)(z). \end{aligned}$$

LEMMA 2. For every  $n \in \mathbb{N}$ ,  $|z| > R_1$  and  $z \in \Sigma$ ,  $d \in ]0, \alpha_0[$ ,  $\chi_n(z)$  is well defined and  $\|\chi_n(z)\|_{\alpha_0-d} \leq \|\chi_0\|_{\alpha_0} (M^n e^n / d^n) |z|^{-n}$  ( $M$  is the constant given in Lemma 1,  $\log e = 1$ ).

Proof. The proof follows an "Ovcyannikov method" of majorization. We shall prove by induction that for every  $d \in ]0, \alpha_0[$  and if  $\alpha' = \alpha_0 - d$  we have

$$(14) \quad \|\chi_n(z)\|_{\alpha'+d/(n+1)} \leq \frac{M^n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{3}\right)^2 \cdots \left(1 - \frac{1}{n+1}\right)^n}.$$

For this we prove by induction that, for every  $d \in ]0, \alpha_0[$ ,

$$\|\chi_n(z)\|_{\alpha'+d/(n+1)} \quad \text{and} \quad \|\chi'_n(z)\|_{\alpha'+d/(n+1)} \leq \frac{M^n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n+1}\right)^n},$$

$$\|\chi''_n(z)\|_{\alpha'+d/(n+1)} \leq \frac{M^n n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n+1}\right)^n}.$$

This is clearly true for  $n = 0$ , so that we may assume it is true for  $n - 1$  and apply Lemma 1 with  $\chi = \chi_{n-1}$ ,  $\alpha = \alpha' + d/n$ ,  $\varepsilon = d/n$  and

$$A = \frac{M^{n-1} \|\chi_0\|_{\alpha_0}}{d^{n-1} \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n}\right)^{n-1}}.$$

It follows that

$$\|\chi_n(z)\|_{\alpha'} \quad \text{and} \quad \|\chi'_n(z)\|_{\alpha'} \leq \frac{M^n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n}\right)^{n-1}},$$

$$\|\chi''_n(z)\|_{\alpha'} \leq \frac{M^n n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n}\right)^{n-1}}.$$

Remember that  $\alpha' = \alpha_0 - d$ ,  $0 < d < \alpha_0$ . We set  $d = d'(1 - 1/(n+1))$ , hence  $\alpha' = \alpha_0 - d' + d'/(n+1)$ ; for  $d' \in ]0, \alpha_0[$  (hence we use only  $0 < d < [n/(n+1)]\alpha_0$ ) we set  $\alpha'' = \alpha_0 - d' > 0$ . Hence for every  $d' \in ]0, \alpha_0[$ ,

$$\|\chi_n(z)\|_{\alpha''+d'/(n+1)} \quad \text{and} \quad \|\chi'_n(z)\|_{\alpha''+d'/(n+1)} \leq \frac{M^n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d'^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n+1}\right)^n},$$

$$\|\chi''_n(z)\|_{\alpha''+d'/(n+1)} \leq \frac{M^n n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d'^n \left(1 - \frac{1}{2}\right) \cdots \left(1 - \frac{1}{n+1}\right)^n},$$

which proves (14).

Since

$$\left(1 - \frac{1}{n+1}\right)^n \geq \frac{1}{e}, \quad \|\chi_n(z)\|_{\alpha'+d/(n+1)} \leq \frac{e^n M^n \|\chi_0\|_{\alpha_0} |z|^{-n}}{d^n}.$$

Since the injection  $E_{\alpha'+[d/(n+1)]} \rightarrow E_{\alpha'}$  is of norm  $\leq 1$ ,

$$\|\chi_n(z)\|_{\alpha_0-d} \leq \|\chi_0\|_{\alpha_0} \frac{M^n e^n}{d^n} |z|^{-n}.$$

*End of the proof of the theorem.* From Lemma 2 the series  $v(z) = \sum_{n=0}^{+\infty} \chi_n(z)$  converges uniformly in  $E_{\alpha_0-d}$  if  $|z| > Me/d$ .  $v$  is solution of (2) and  $y(z) = e^{(-\alpha_1/2)z} e^{\sigma z} (z)^\rho v(z)$  is the desired solution. The proofs of the remarks are quite analogous to the classical case (Hoheisel [3], Erdelyi [2]).

**Acknowledgment.** The authors are indebted to the referee for valuable criticism.

## REFERENCES

- [1] M. S. BAOUENDI AND C. GOULAOUIC, *Cauchy problems with characteristic initial hypersurface*, Comm. Pure Appl. Math., 36 (1973), pp. 433–473.
- [2] A. ERDELYI, *Asymptotic Expansions*, Dover, New York, 1956.
- [3] G. HOHEISEL, *Asymptotische integration linearer Differential-gleichungen*, J. Reine Angew. Math., 153 (1924), pp. 228–244.
- [4] L. V. OVCYANNIKOV, *A singular operator in a scale of Banach spaces*, Dokl. Akad. Nauk. SSSR, 163 (1965), pp. 819–822; Soviet Math. Dokl, 6 (1965), pp. 1025–1028.
- [5] F. TRÈVES, *On the theory of linear partial differential operators with analytic coefficients*, Trans. Amer. Math. Soc., 137 (1969), pp. 1–20.
- [6] ———, *Ovcyannikov theorem and hyperdifferential operators*, I.M.P.A., Rio de Janeiro, 1968.
- [7] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

## KILLING TENSORS AND VARIABLE SEPARATION FOR HAMILTON-JACOBI AND HELMHOLTZ 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 show how to characterize those involutive families of Killing tensors that correspond to orthogonal separation, and for  $n = 3$  those families that correspond to nonorthogonal separation.

**1. Introduction.** We are concerned with the separation of variables problem for the Hamilton-Jacobi equation

$$(1.1) \quad g^{ij} \partial_{x^i} W \partial_{x^j} W = E, \quad g^{ij} = g^{ji}, \quad 1 \leq i, j \leq n,$$

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  $\{x^i\}$ . (Here we allow all coordinates and tensors to be complex, and adopt the tensor notation in Eisenhart's book [1].) Full understanding of the separation problem for the Helmholtz equation

$$(1.2) \quad \frac{1}{\sqrt{g}} \partial_{x^i} (\sqrt{g} g^{ij} \partial_{x^j} \psi) = E \psi,$$

$$g = \det (g_{ij}),$$

depends on an understanding of the corresponding problem for (1.1). Indeed, it is rather easy to show that any coordinate system yielding (product) separation of (1.2) also yields (additive) separation of (1.1); see, e.g., [2]–[5]. For orthogonal coordinates it was shown by Eisenhart [6] that a separable system  $\{x^i\}$  for (1.1) yields separation of (1.2) if and only if  $R_{ij} = 0$ ,  $1 \leq i < j \leq n$ , where  $R_{hi}$  is the Ricci tensor (thus the nondiagonal elements of the Ricci tensor vanish). For nonorthogonal coordinates the conditions that separation of (1.1) yields separation of (1.2) are much more complicated; for  $n = 3$  and 4 these conditions are given in [2], [3]. (For  $n = 2$ , (1.1) and (1.2) separate in precisely the same systems. Furthermore, the authors have shown that for constant curvature spaces in dimensions 3 and 4 the two equations also separate in the same systems.)

To study the relation between variable separation for (1.1) and Killing tensors for  $V_n$  we use the natural symplectic structure on the cotangent bundle  $\tilde{V}_n$  of the manifold. Corresponding to local coordinates  $\{x^i\}$  on  $V_n$ , we have coordinates  $\{x^i, p_i\}$  on  $\tilde{V}_n$ . (If  $\{\hat{x}^k(x^i)\}$  is another local coordinate system on  $V_n$  then it corresponds to  $\{\hat{x}^k, \hat{p}_k\}$ , where  $\hat{p}_k = p_i \partial x^i / \partial \hat{x}^k$ .) The *Poisson bracket* of two functions  $F(x^i, p_i)$ ,  $G(x^i, p_i)$  on  $\tilde{V}_n$  is defined by

$$(1.3) \quad [F, G] = \partial_{x^i} F \partial_{p_i} G - \partial_{p_i} F \partial_{x^i} G.$$

Let

$$(1.4) \quad H = g^{ij} p_i p_j.$$

\* Received by the editors October 25, 1979, and in final revised form April 3, 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.



A linear function  $L$  in the momenta  $p_j$ ,

$$(1.5) \quad L = \xi^j(x^l)p_j,$$

is a symmetry for (1.1) if

$$(1.6) \quad [L, H] = 0.$$

In this case we say that  $\{\xi^j\}$  is a *Killing vector*. It is straightforward to show that (1.6) is equivalent to Killing's equation

$$(1.7) \quad \xi_{i,j} + \xi_{j,i} = 0,$$

where  $\xi_{i,j}$  is the  $j$ th covariant derivative of  $\xi_i$  [1]. Similarly the quadratic function

$$(1.8) \quad A = a^{ij}(x^l)p_i p_j, \quad a^{ij} = a^{ji},$$

is a symmetry of (1.1) provided  $[A, H] = 0$ , and this is equivalent to

$$(1.9) \quad a_{ij,k} + a_{ki,j} + a_{jk,i} = 0.$$

We say that  $\{a^{ij}\}$  (or  $\{a_{ij}\}$ ) is a *Killing tensor of order 2*. Note that the condition that two quadratic functions  $A$  and  $B = b^{ij}p_i p_j$  are in involution, i.e., that  $[A, B] = 0$ , is

$$(1.10) \quad a_{ij,l}b^l_k + a_{ki,l}b^l_j + a_{jk,l}b^l_i = b_{ij,l}a^l_k + b_{ki,l}a^l_j + b_{jk,l}a^l_i.$$

The basic link between separation of variables for (1.1) and Killing tensors is now easy to state: To every orthogonal coordinate system  $\{y^l\}$  which permits additive separation of variables in (1.1), there correspond  $n - 1$  second order Killing tensors  $A_1, \dots, A_{n-1}$ , which are in involution and such that  $\{H, A_1, \dots, A_{n-1}\}$  is linearly independent. The separable solutions  $W = \sum_{k=1}^n W^{(k)}(y^k)$  are characterized by the relations

$$(1.11) \quad H(y^j, p_j) = E, \quad A_l(y^j, p_j) = \lambda_l, \quad l = 1, \dots, n - 1, \quad p_j = \partial_{y^j} W,$$

where  $\lambda_1, \dots, \lambda_{n-1}$  are the separation constants. See [7] and [8] for definitions and discussions of the proof. For nonorthogonal separable coordinates the characterization is the same except that one or more of the  $A_l$  are first order Killing tensors, i.e., Killing vectors. For  $n \leq 4$  all possible separable systems and their corresponding Killing tensors have been computed [2], [3].

In the language of Hamiltonian mechanics, Killing tensors are “constants of the motion”. The basic link mentioned above states that if (1.1) is separable, then the corresponding Hamiltonian system is “completely integrable” [9].

A fundamental difficulty remaining in this theory is that, whereas to every separable system there corresponds a family of  $n - 1$  Killing tensors in involution, there also exist families of Killing tensors in involution that are not related to separable systems, see, e.g., [10], [11]. For a truly satisfactory theory a decision process is needed to determine, for a given family of  $n - 1$  commuting Killing tensors, whether or not that family characterizes a system of separable coordinates and, if so, to compute these coordinates from the given Killing tensors. (Thus, we need to show which constants of the motion lead to variable separation.)

In this paper we develop the decision process to characterize for all  $n \geq 2$  those families of Killing tensors that correspond to orthogonal, i.e., Stäckel type, coordinates, and for  $n = 3$  those families that correspond to nonorthogonal separation. Basic to our theory is the algebraic classification of pairs of quadratic forms under conjugacy transformations [12].

In § 2 we study the first and second order Killing tensors corresponding to a 2-dimensional Riemannian manifold, and show how to characterize those Killing tensors that define variable separation. The case  $n = 2$  is especially simple, because every manifold  $V_2$  is conformally flat and because variable separation is determined by a single Killing tensor. Thus one is able to compute much more explicit and detailed results than is possible for larger  $n$ .

In § 3 we determine the maximum dimension of the space of second order Killing tensors for a general Riemannian manifold  $V_n$ , and characterize the families of  $n - 1$  Killing tensors that determine orthogonal variable separation for (1.1). Here the conditions (1.10) expressing the fact that two Killing tensors are in involution become important.

Finally, in § 4 we show how to characterize for  $n = 3$  those Killing tensors that define nonorthogonal separation.

The techniques of this paper are related to many subjects of physical and practical importance. Separation of variables is one of the most powerful tools for obtaining complete integrals of (1.1) and explicit solutions of (1.2). (In many problems one adds a potential  $V(x)$  to the left-hand sides of these equations. Addition of a potential merely places another restriction on the possible coordinate systems permitting separation.) Most of the special functions of mathematical physics arise as solutions of (1.2) in appropriate separable coordinate systems. For many examples and applications see [13].

**2. Separable systems for  $V_2$ .** For two-dimensional Riemannian spaces there are only three distinct ways that variables can separate in (1.1), corresponding to the number of ignorable variables, i.e., variables associated with Killing vectors:

(1) 2 ignorable variables. The metric is

$$(2.1) \quad ds^2 = (dx^1)^2 + (dx^2)^2 = g_{ij} dx^i dx^j,$$

and the Killing tensor is  $A = p_1^2$ . These are just Cartesian coordinates in flat space.

(2) 1 ignorable variable. The metric is

$$(2.2) \quad ds^2 = f(x^2)[(dx^1)^2 + (dx^2)^2],$$

and the Killing tensor is  $A = p_1^2$ .

(3) No ignorable variables. The metric is

$$(2.3) \quad ds^2 = (\sigma_1(x^1) + \sigma_2(x^2))[(dx^1)^2 + (dx^2)^2],$$

and the corresponding Hamilton-Jacobi equation is

$$\frac{1}{\sigma_1 + \sigma_2} [(\partial_{x^1} W)^2 + (\partial_{x^2} W)^2] = E.$$

The associated Killing tensor is

$$(2.4) \quad A = \frac{1}{\sigma_1 + \sigma_2} (\sigma_2 p_1^2 - \sigma_1 p_2^2).$$

Note that (1) can be considered as a degenerate case of (2), which in turn is a degenerate case of (3). Furthermore, all separable coordinate systems for  $n = 2$  are orthogonal.

In order to determine exactly when the above cases can arise, we study the space of all second order Killing tensors for a given local manifold  $V_2$ . In particular, we examine

the pairs of forms

$$(2.5) \quad \begin{aligned} ds^2 &= \varphi = g_{ij} dx^i dx^j, & 1 \leq i, j \leq 2, \\ \psi &= a_{ij} dx^i dx^j, \end{aligned}$$

where the  $a_{ij}$  are components of a Killing tensor, i.e., satisfy equations (1.9). Using the standard algebraic classification of pairs of quadratic forms under conjugacy transformations [12] we can at a fixed point  $P \in V_2$  choose the pair in one of the following canonical forms:

$$(i) \quad [11]: \quad \begin{aligned} \varphi &= g_{ij} dx^i dx^j = c_1(dx^1)^2 + c_2(dx^2)^2, \\ \psi &= a_{ij} dx^i dx^j = c_1\rho_1(dx^1)^2 + c_2\rho_2(dx^2)^2. \end{aligned}$$

Here  $c_1, c_2$  are constants and the  $\rho_i$  are distinct roots of the equation

$$(2.6) \quad \det(a_{ij} - \rho g_{ij}) = 0.$$

$$(ii) \quad [(11)]: \quad \text{This is type (i) with } \rho_1 = \rho_2.$$

$$(iii) \quad [2]: \quad \begin{aligned} \varphi &= 2 dx^1 dx^2 \\ \psi &= c_1(dx^1)^2 + 2\rho_1 dx^1 dx^2. \end{aligned}$$

Here  $\rho_1$  is a double root of (2.6). The standard notation for these forms is explained in [12].

It is easy to show that the roots  $\rho_i$  and the classification into canonical types are independent of coordinates. Furthermore, except for some singular cases that are not of interest here, a Killing tensor will maintain its canonical type in some neighborhood of  $P$ .

Now we explicitly compute the possible Killing tensors admitted by a given space and classify the possibilities according to their canonical types. To simplify the computations we note that every  $V_2$  is conformally flat; i.e., there exist coordinates  $\{x^1, x^2\}$  such that

$$ds^2 = Q(x^1, x^2)[(dx^1)^2 + (dx^2)^2] = g_{ij} dx^i dx^j.$$

Equations (1.9) then become ( $\partial_i Q \equiv Q_i$ ):

$$(2.7) \quad \begin{aligned} a_{11,1} &= \partial_1 a_{11} + a_{12} \frac{Q_2}{Q} - a_{11} \frac{Q_1}{Q} = 0, \\ a_{22,2} &= \partial_2 a_{22} + a_{12} \frac{Q_1}{Q} - a_{22} \frac{Q_2}{Q} = 0, \\ 2a_{12,2} + a_{22,1} &= 2\partial_2 a_{12} - 3a_{12} \frac{Q_2}{Q} + \partial_1 a_{22} - (a_{11} - 2a_{22}) \frac{Q_1}{Q} = 0, \\ 2a_{12,1} + a_{11,2} &= 2\partial_1 a_{12} - 3a_{12} \frac{Q_1}{Q} + (a_{22} - 2a_{11}) \frac{Q_2}{Q} + \partial_2 a_{11} = 0. \end{aligned}$$

Writing  $a_{ij} = Q\bar{a}_{ij}$  and substituting the first two equations (2.7) in the last two, we obtain

$$(2.8) \quad \begin{aligned} 2\partial_x^2(\bar{a}_{12}/Q) + \partial_x^1((\bar{a}_{22} - \bar{a}_{11})/Q) &= 0, \\ 2\partial_x^1(\bar{a}_{12}/Q) + \partial_x^2((\bar{a}_{11} - \bar{a}_{22})/Q) &= 0. \end{aligned}$$

The integrability conditions for (2.8) imply

$$(2.9) \quad [\partial_{x^1 x^1} + \partial_{x^2 x^2}] \left( \frac{\bar{a}_{11} - \bar{a}_{22}}{Q} \right) = 0,$$

i.e.,

$$(2.10) \quad \frac{\bar{a}_{11} - \bar{a}_{22}}{Q} = f(x^1 + ix^2) + h(x^1 - ix^2), \quad \frac{2\bar{a}_{12}}{Q} = -i(f - h),$$

where  $f, h$  are analytic functions. The remaining equations (2.7) reduce to

$$(2.11) \quad \begin{aligned} \partial_{x^1} \bar{a}_{11} &= \frac{i}{2} Q_2 (f - h), \\ \partial_{x^2} \bar{a}_{22} &= \frac{i}{2} Q_1 (f - h). \end{aligned}$$

The integrability condition for (2.11) and the first of equations (2.10) is

$$\partial_{x^1 x^2} [Q(f + h)] + \frac{i}{2} \partial_{x^1} [Q_1(f - h)] - \frac{i}{2} \partial_{x^2} [Q_2(f - h)] = 0.$$

If we choose new coordinates  $z^1 = x^1 + ix^2, z^2 = x^1 - ix^2$ , then this last condition becomes

$$(2.12) \quad 2(f(z^1)Q_{z^1 z^1} - h(z^2)Q_{z^2 z^2}) + 3(f'Q_{z^1} - h'Q_{z^2}) + Q(f'' - h'') = 0,$$

where the metric is

$$(2.13) \quad ds^2 = Q dz^1 dz^2.$$

If we regard  $V_2$ , hence  $Q$ , as given, and find functions  $f(z^1), h(z^2)$  satisfying (2.12), we can then employ (2.1) and (2.11) to determine the matrix  $(a_{ij})$  to within the addition of a constant times  $Q\delta_{ij} = g_{ij}$ , where  $\delta_{ij}$  is the Kronecker delta.

Recall that  $V_2$  is a space of constant curvature if and only if

$$(2.14) \quad \partial_{z^1 z^2} \ln Q = kQ,$$

for some constant  $k$ . The case  $k = 0$  corresponds to flat space.

We now restrict attention to spaces for which  $Q(z^1, z^2)$  is analytic in a neighborhood of the point  $(z_0^1, z_0^2)$ , which without loss of generality we can take as  $(0, 0)$ , and consider the vector space of all second order Killing tensors  $a_{ij}$  on  $V_2$ , analytic in a neighborhood of  $(0, 0)$ . Let  $D$  be the dimension of this vector space.

**THEOREM 1.**

- (1)  $1 \leq D \leq 6$ .
- (2)  $D = 6$  if and only if  $V_2$  is a space of constant curvature.
- (3) If  $V_2$  is not a space of constant curvature, then  $D \leq 4$ .

*Proof.* We write (2.12) in the form

$$(2.15) \quad f'' + \frac{3f'Q_1}{Q} + \frac{2fQ_{11}}{Q} = h'' + \frac{3h'Q_2}{Q} + \frac{2hQ_{22}}{Q}.$$

Prescribing the values of  $h(0), h'(0), h''(0), f(0), f'(0)$  we can use (2.15) to compute  $f''(0)$ . Differentiating (2.15) successively with respect to  $z^1$ , we can compute  $f^{(k)}(0)$  for all  $k \geq 2$ . Similarly, successive differentiation of (2.15) with respect to  $z^2$  allows us to compute  $h^{(l)}(0)$  for  $l \geq 3$ . Thus any solution  $\{f(z^1), h(z^2)\}$  is uniquely determined by the above five prescribed values. Once  $f$  and  $h$  are given, the Killing tensor  $a_{ij}$  is determined

to within addition of an arbitrary multiple of  $g_{ij}$ . Hence  $1 \leq D \leq 6$ , and assertion (1) is verified.

For spaces of constant curvature  $D = 6$ . Indeed, for flat space and Cartesian coordinates the functions

$$(2.16) \quad L_1 = p_1, \quad L_2 = p_2, \quad L_3 = x^1 p_2 - x^2 p_1$$

form a basis for the Lie algebra of Killing vectors. Clearly the set  $\{L_j L_k : 1 \leq j \leq k \leq 3\}$  forms a basis for a 6-dimensional subspace of second order Killing tensors. It follows immediately from (1) that this subspace is in fact the full space of Killing tensors and  $D = 6$ . A similar argument holds for Riemannian spaces of nonzero constant curvature.

To finish the proof of (2) we note that  $D = 6$  if and only if the integrability conditions for (2.15) are satisfied identically. Suppose  $D = 6$ . Applying the operator  $\partial_2 \cdot x^2$  to both sides of (2.15), we obtain

$$(2.17) \quad \begin{aligned} &\partial_2 \left( \frac{3Q_1}{Q} \right) f'' + 2\partial_2 \left( \frac{Q_{11}}{Q} \right) f' + \partial_{12} \left( \frac{3Q_1}{Q} \right) f' + \partial_{12} \left( \frac{2Q_{11}}{Q} \right) f \\ &= \partial_1 \left( \frac{3Q_2}{Q} \right) h'' + 2\partial_1 \left( \frac{Q_{22}}{Q} \right) h' + \partial_{12} \left( \frac{3Q_2}{Q} \right) h' + \partial_{12} \left( \frac{2Q_{22}}{Q} \right) h. \end{aligned}$$

If  $D = 6$ , then this condition on  $f$  and  $h$  cannot be independent of (2.15). Hence, either the coefficients of  $f''$ ,  $f'$ ,  $f$ ,  $h''$ ,  $h'$ , and  $h$  vanish identically, in which case  $\partial_{12} \ln Q = 0$  and  $V_2$  is flat, or  $\partial_{12} \ln Q \neq 0$  and (2.16) is obtained from (2.15) through multiplication by  $\partial_{12} \ln Q$ . In the second case one verifies easily that

$$\partial_1 \left( \frac{\partial_{12} \ln Q}{Q} \right) = \partial_2 \left( \frac{\partial_{12} \ln Q}{Q} \right) = 0,$$

hence that  $\partial_{12} \ln Q = kQ$ ,  $k \neq 0$ . Thus,  $V_2$  is a space of nonzero constant curvature.

If  $D < 6$ , then  $\partial_{12} \ln Q \neq 0$  and (2.16) is independent of (2.15). Then we can eliminate  $f''$  and  $h''$  between these two equations and obtain a condition relating only  $f'$ ,  $f$ ,  $h'$  and  $h$ . It follows that  $D \leq 4$ . Q.E.D.

Before proceeding further it is useful to recall the classical work of Stäckel and Eisenhart characterizing *orthogonal* separable coordinates  $\{y^i\}$  on a manifold  $V_n$ . Stäckel [14] showed that the orthogonal coordinates permit separation if and only if the metric

$$(2.18) \quad ds^2 = H_1^2 (dy^1)^2 + \dots + H_n^2 (dy^n)^2$$

is in *Stäckel form*, i.e.,

$$(2.19) \quad H_j^2 = \frac{S}{S^{j1}}, \quad S = \det (\varphi_{ij}(y^i)) \neq 0,$$

where  $S^{i1}$  is the cofactor of  $\varphi_{i1}$  in  $S$ . Eisenhart, [6] and [1, Appendix 13], found a more intrinsic characterization of Stäckel form. His basic result is:

**THEOREM 2.** *A necessary and sufficient condition that the metric  $ds^2 = g_{ij} dx^i dx^j$  on  $V_n$  can be given the Stäckel form is that*

- (1) *the space admits  $n - 1$  Killing tensors  $a_{ij}^{(\alpha)}$ ,  $\alpha = 1, \dots, n - 1$ , such that the  $n$  tensors  $\{g_{ij}, a_{ij}^{(\alpha)}\}$  form a linearly independent set;*
- (2) *the roots  $\rho^{(\alpha)}$  for each of the characteristic equations  $\det (a_{ij}^{(\alpha)} - \rho^{(\alpha)} g_{ij}) = 0$  are simple;*

(3) *there is a coordinate system  $\{y^i\}$  on  $V_n$  such that*

$$(2.20) \quad (a_{ij}^{(\alpha)} - \rho_h^{(\alpha)} g_{ij}) \lambda_{(h)}^i = 0, \quad h = 1, \dots, n, \quad \alpha = 1, \dots, n - 1,$$

where  $\rho_1^{(\alpha)}, \dots, \rho_n^{(\alpha)}$  are the roots of  $a_{ij}^{(\alpha)}$ . Here,  $\lambda_{(h)}^i = \partial x^i / \partial y^h$ .

The coordinates  $\{y^i\}$  are those for which the metric assumes Stäckel form. Note that condition (3) requires that the vector fields  $\lambda_{(1)}^i, \dots, \lambda_{(h)}^i$  be normal, and that they satisfy (2.20) for all  $\alpha$ .

Of course the normality conditions on the vector fields are very difficult to check for general  $n$ , and Theorem 2 is not very useful as a practical tool. (We will formulate and prove a practical version of this theorem in § 3.) Moreover, for  $n > 2$  nonorthogonal separation may occur, and this was not considered by Eisenhart.

On the other hand, for  $n = 2$  Theorem 2 simplifies greatly:

**COROLLARY 1.** *Every Killing tensor  $a_{ij}$  that is linearly independent of  $g_{ij}$  and of type [11] in a coordinate neighborhood defines a separable coordinate system for the Hamilton-Jacobi equation on  $V_2$ . Conversely, every separable coordinate system arises in this manner.*

Indeed it is simple to show that the two equations (2.20), ( $\alpha = 1, h = 1, 2$ ) must always admit normal vector fields as solutions. Furthermore, for  $n = 2$  all separable systems are orthogonal. (Note that the explicitly given Killing tensors defining variable separation on  $V_2$ , e.g., (2.4) are all of type [11].)

Now we return to the explicit computation of second order Killing tensors on the manifold with metric (2.13). Every tensor  $\psi = a_{ij} dx^i dx^j$  obtained by solving (2.10)–(2.12) defines a separation of variables, provided it is not a multiple of  $g_{ij}$  and it has elementary divisors of type [11]. This latter condition, that of unequal roots, takes the form

$$(2.21) \quad \Phi \equiv (a_{11} - a_{22})^2 + 4a_{12}^2 \neq 0.$$

If  $\psi$  is of type [(11)], then  $\Phi = 0$ , and in fact  $a_{11} = a_{22}, a_{12} = 0$ . It follows easily from (2.10)–(2.12) that then  $a_{11} = a_{22} = \lambda Q, \lambda \in \mathbb{C}$ . Thus there are no nontrivial type [(11)] Killing tensors.

Now suppose  $\psi$  is of type [2]. Then  $\Phi = 0$ , so

$$(2.22) \quad a_{11} - a_{22} = \pm 2ia_{12} \neq 0.$$

Without loss of generality we can assume that the plus sign holds in (2.22). Then the integrability condition (2.12) reduces to

$$(2.23) \quad 2f \frac{Q_{11}}{Q} + 3f' \frac{Q_1}{Q} + f'' = 0.$$

It follows immediately from this expression that type [2] tensors either do not occur or form a subspace of dimension 2 or 3 (from which the one-dimensional subspace corresponding to  $f \equiv 0$  must be deleted).

**THEOREM 3.**

(1)  $V_2$  admits a 3-dimensional subspace of type [2] Killing tensors if and only if it is flat.

(2)  $V_2$  admits a 2-dimensional subspace of type [2] Killing tensors if and only if  $Q_{11} = g(z^1)Q_1 + \frac{1}{3}(g' - 2g^2)Q$  for some analytic function  $g$ .

(3) If  $V_2$  is of nonzero constant curvature, it admits no type [2] Killing tensors.

To prove (1), we note that if the subspace of type [2] Killing tensors is of dimension 3 then the integrability conditions for (2.23) must all be satisfied identically. Differentiating (2.23) with respect to  $z^2$ , we obtain a condition on  $f$  and  $f'$  alone, which must be

trivial. Hence  $\partial_{12} \ln Q = 0$  and  $V_2$  is flat. Conversely, if  $V_2$  is flat we can choose  $Q \equiv 1$  so that (2.23) becomes  $f''' = 0$  and the subspace of type [2] tensors is 3 dimensional. Parts (2) and (3) of the theorem are proved by similar but slightly more involved computations. QED

For  $n = 2$  we now have a complete solution to our problem. Type [11] Killing tensors independent of the metric tensor always define separable coordinates, whereas type [2] tensors are never associated with separation. The manifolds admitting type [2] tensors can be computed explicitly, and include flat space but not spaces of nonzero constant curvature. A real Riemannian manifold with positive definite metric never admits a type [2] tensor.

The most important example of a type [2] Killing tensor is

$$(2.24) \quad a^{ij} p_i p_j = L_3(L_1 - iL_2)$$

in flat space, where we are using the notation (2.16). Here

$$(2.25) \quad \begin{aligned} \varphi &= ds^2 = (dx^1)^2 + (dx^2)^2, \\ \psi &= -x^2(dx^1)^2 + (x^1 + ix^2) dx^1 dx^2 - ix^1(dx^2)^2. \end{aligned}$$

In [10], [13] the authors point out that (2.24) does not correspond to variable separation. Moreover, as follows from [13, p. 60], this example is unique. Every type [2] Killing tensor in flat space is, to within addition of a scalar multiple of  $p_1^2 + p_2^2$ , an element of the orbit of (2.24) under the adjoint action of the complex Euclidean group  $E(2)$ .

**3. Orthogonal separable systems.** Let  $V_n$  be a complex  $n$ -dimensional Riemannian manifold. We choose a local coordinate system  $\{x^i\}$  on this manifold and consider the vector space of Killing tensors  $\psi = a_{ij} dx^i dx^j$  analytic in a neighborhood of  $\theta$ , ( $x^j = 0, j = 1, \dots, n$ ). Let  $D$  be the dimension of this space.

THEOREM 4.

- (1)  $1 \leq D \leq n(n+1)^2(n+2)/12$ .
- (2) If  $V_n$  is flat then  $D = n(n+1)^2(n+2)/12$ .

*Proof.* The condition that  $\psi$  is a Killing tensor is (1.9), which we can write in the form

$$(3.1) \quad \partial_k a_{ij} + \partial_j a_{ki} + \partial_i a_{jk} = R_{ijk}(x^l, a_{hm}),$$

where the terms  $R_{ijk}$  are linear in  $a_{hm}$ . Clearly,  $\psi$  will be uniquely determined by the constants  $a_{hm}(\theta)$  and all possible derivatives of  $a_{hm}$  evaluated at  $\theta$ . These constants are not independent of one another, since they are constrained by (3.1) and all possible derivatives of (3.1) evaluated at  $\theta$ .

Now the number of constants  $a_{hm}(\theta)$  is  $B_0 = n(n+1)/2$ . The number of terms  $\partial_k a_{ij}(\theta)$  is  $B_1 = n^2(n+1)/2$ , and there are  $C_1 = n(n+1)(n+2)/6$  equations (3.1) giving linear restrictions on these terms. Similarly, there are  $B_2 = n^2(n+1)^2/4$  terms  $\partial_{ki} a_{ij}(\theta)$ , and by differentiating (3.1) we obtain  $C_2 = n^2(n+1)(n+2)/6$  conditions on these terms. Finally there are  $B_3 = n^2(n+1)^2(n+2)/12$  terms  $\partial_{kth} a_{ij}(\theta)$ , and by differentiating (3.1) twice we obtain  $C_3 = n^2(n+1)^2(n+2)/12$  conditions on these third derivative terms. We will show that the  $C_1 + C_2 + C_3$  conditions are linearly independent. This implies that, since  $B_3 = C_3$ , we can solve the  $C_3$  third order equations uniquely for the  $B_3$  third order derivatives:

$$(3.2) \quad \partial_{kth} a_{ij}(\mathbf{x}) = S_{kthij},$$

where  $S$  depends linearly on the tensor components of  $\psi$  and their first two derivatives. From (3.2) we can recursively compute all higher derivatives of  $a_{ij}$ , evaluated at  $\theta$ . Thus  $1 \leqq D \leqq B_0 + B_1 + B_2 - C_1 - C_2 = n(n+1)^2(n+2)/12$ .

In general the integrability conditions for (3.2) will impose additional constraints on the constants, and  $D$  will not achieve this upper limit. However, if  $V_n$  is flat and we choose Cartesian coordinates, then  $R_{ijk} \equiv 0$  in (3.1), and (3.2) reduces to

$$(3.3) \quad \partial_{kjh} a_{ij}(\mathbf{x}) = 0.$$

The integrability conditions for (3.3) are satisfied identically; hence,  $D = n(n+1)^2(n+2)/12$ .

To finish the proof of (1), it is sufficient to show that the  $C_3$  conditions

$$(3.4) \quad \partial_{lhh} a_{ij} + \partial_{lhi} a_{jk} + \partial_{lhj} a_{ki} \sim 0$$

are linearly independent. This is equivalent to showing that these conditions imply the  $B_3$  equations  $\partial_{lhh} a_{ij} \sim 0$ , where “ $\sim$ ” denotes equality up to linear terms in the components of  $\psi$  and their first two derivatives.

Denoting the left-hand side of (3.4) by  $M_{lh,kij}$  we have

$$M_{lh,kij} + M_{kl,hij} - M_{jl,kih} - M_{ij,lkh} = 2(\partial_{lhh} a_{ij} - \partial_{ijh} a_{hk}) \sim 0.$$

Substituting this result and the equivalent forms  $\partial_{khl} a_{ij} \sim \partial_{ijk} a_{lh} \sim \partial_{ijh} a_{lk}$  into  $M_{ij,khl} \sim 0$ , we find  $\partial_{khl} a_{ij} \sim 0$ . Q.E.D.

This result is already known (see [15]–[17]). We have included it here because our proof (though similar to [16]) contains a degree of novelty which enables us to very easily obtain the following two corollaries.

If  $V_n$  is flat, then in Cartesian coordinates  $\{x^i\}$  the Hamilton-Jacobi equation (1.1) becomes

$$(3.5) \quad \sum_{i=1}^n p_i^2 = E, \quad p_i = \partial_i W.$$

The symmetry algebra of this equation is  $\mathcal{G}(n)$  with basis

$$(3.6) \quad p_i, \quad 1 \leqq i \leqq n, \quad m_{jk} = x^j p_k - x^k p_j, \quad 1 \leqq j < k \leqq n.$$

**COROLLARY 2.** *If  $V_n$  is flat,  $n \geqq 2$ , a basis for the space of second order Killing tensors is*

- (a)  $p_i p_j, \quad 1 \leqq i \leqq j \leqq n,$
- (b)  $m_{ij} p_k, m_{ik} p_j, \quad 1 \leqq i < j < k \leqq n,$
- (c)  $m_{ij} p_j, \quad i \neqq j,$
- (d)  $m_{ij} m_{ij}, \quad i < j,$
- (e)  $m_{ik} m_{il}, \quad i \neqq k, l, \quad 1 \leqq k < l \leqq n,$
- (f)  $m_{ij} m_{kl}, m_{ik} m_{jl}, \quad 1 \leqq i < j < k < l \leqq n.$

In particular, all flat space second order Killing tensors are expressible as polynomials of order two in the Killing vectors. Corollary 2 follows easily from consideration of the conditions on the first and second derivatives of the  $a_{ij}$  as obtained in Theorem 4. It also follows from the proof of that theorem that if  $V_n$  admits a space of Killing tensors of maximal dimension  $D = n(n+1)^2(n+2)/12$ , then a basis for this space is  $\{\varphi_{(\alpha)}; 1 \leqq \alpha \leqq D\}$ ; here  $\varphi = A_{(\alpha)} + B_{(\alpha)}$ , the  $A_{(\alpha)}$  run over the flat space basis elements listed in Corollary 2, and

$$B_{(\alpha)} = b_{(\alpha)}^{ij} p_i p_j,$$



with the order of each component  $b_{(\alpha)}^{ij}$  in the variables  $\{x^k\}$  strictly greater than the order (0, 1, or 2) of the components of  $A_{(\alpha)}$ .

If  $V_n$  is a space of nonzero constant curvature (which we normalize as  $K = 4$ ), then there exist coordinates  $\{x^k\}$  such that (see [1]),

$$g_{ij} = \frac{\delta_{ij}}{(1+r^2)^2}, \quad r^2 = \delta_{ab}x^a x^b.$$

The symmetry algebra of the Hamilton-Jacobi equation for  $V_n$  is  $o(n+1)$  with basis

$$(3.7) \quad \begin{aligned} m_{jk} &= x^j p_k - x^k p_j, & 1 \leq j < k \leq n \\ q_l &= (2(x^l)^2 - r^2 + 1)p_l - 2x^l \sum_{s \neq l} x^s p_s, & 1 \leq l \leq n. \end{aligned}$$

Clearly, all products of pairs of Killing vectors are Killing tensors for  $V_n$ , and the tensors  $p_i p_j, m_{ij} p_k, m_{ij} m_{kl}$  for flat space Cartesian coordinates agree respectively (up to terms of lowest order) with the tensors  $q_i q_j, m_{ij} q_k, m_{ij} m_{kl}$  on  $V_n$ .

**COROLLARY 3.** *If  $V_n$  is a space of nonzero constant curvature,  $n \geq 2$ , then  $D = n(n+1)^2(n+2)/12$  and a basis for the space of second order Killing tensors is*

- (a)  $q_i q_j, \quad 1 \leq i \leq j \leq n,$
- (b)  $m_{ij} q_k, m_{ik} q_j, \quad 1 \leq i < j < k \leq n,$
- (c)  $m_{ij} q_j, \quad i \neq j,$
- (d)  $m_{ij} m_{ij}, \quad i < j,$
- (e)  $m_{ik} m_{il}, \quad i \neq k, l, 1 \leq k < l \leq n,$
- (f)  $m_{ij} m_{kl}, m_{ik} m_{jl}, \quad 1 \leq i < j < k < l \leq n.$

Now we consider the problem of characterizing orthogonal separable coordinate systems for general Riemannian manifolds  $V_n$ . Although Eisenhart's result, Theorem 2, gives such a characterization, there remains the considerable practical difficulty of determining when the vector fields  $\{\lambda^i_{(h)}\}$  defined by (2.20) are *normalizable*; i.e., when there exists an orthogonal coordinate system  $\{y^i\}$  such that  $\{\lambda^i_{(h)}\}$  is orthogonal to the coordinate surface  $y^h = \text{const.}$ , for each  $h = 1, \dots, n$ .

The conditions for normalizability are classical and can be expressed in terms of the invariants  $\gamma_{lhk}$ , the *coefficients of rotation*:

$$\gamma_{lhk} = \lambda_{(l)i} \lambda^i_{(h)} \lambda^j_{(k)}, \quad 1 \leq l, h, k \leq n;$$

see, e.g., [1, p. 97]. Then a necessary and sufficient condition that there exist coordinates  $\{y^h\}$  and nonzero invariant functions  $f_h$  such that  $\lambda^i_{(h)} = (\partial x^i / \partial y^h) f_h$  (no sum on  $h$ ), is

$$(3.8) \quad \gamma_{hkl} = 0, \quad 1 \leq h, k, l \leq n, \quad h, k, l \text{ distinct,}$$

see [1, p. 117]. Now suppose  $a_{ij}$  is a tensor field with  $n$  roots  $\rho_1, \dots, \rho_n$ , not necessarily distinct, and let  $\{\lambda^i_{(h)}\}$  be a corresponding orthonormal set of eigenvectors:

$$(3.9) \quad (a_{ij} - \rho_h g_{ij}) \lambda^i_{(h)} = 0, \quad h = 1, \dots, n,$$

$$(3.10) \quad \lambda^i_{(h)} \lambda_{(k)i} = \delta_{hk}, \quad 1 \leq h, k \leq n.$$

Differentiating (3.9) covariantly with respect to  $x^k$  and employing (3.7) and (3.10), we find

$$(3.11) \quad a_{ij,k} \lambda^i_{(h)} \lambda^j_{(l)} \lambda^k_{(m)} = (\rho_h - \rho_l) \gamma_{hlms}, \quad h \neq m.$$

Now suppose the roots of  $a_{ij}$  are simple, i.e., pairwise distinct. From (3.8) we obtain

**THEOREM 5** (Eisenhart, [1, p. 118]). *If  $a_{ij}$  has simple roots  $\rho_1, \dots, \rho_n$  then a necessary and sufficient condition that the vector fields  $\{\lambda^i_{(h)}\}$  be normalizable is*

$$(3.12) \quad a_{ij,k} \lambda^i_{(h)} \lambda^j_{(l)} \lambda^k_{(m)} = 0, \quad 1 \leq h, l, m \leq n, \quad h, l, m \text{ distinct.}$$

We are now ready to prove our fundamental result. Let  $H$ , (1.4), be the Hamiltonian on  $V_n$ .

**THEOREM 6.** *Necessary and sufficient conditions for the existence of an orthogonal separable coordinate system  $\{y^j\}$  for the Hamilton-Jacobi equation (1.1) are that there exist  $n - 1$  quadratic functions  $A^{(\alpha)}$ , (1.8), satisfying:*

(1) *The  $\{A^{(\alpha)}\}$  are constants of the motion, i.e.,  $[H, A^{(\alpha)}] = 0$ ,  $\alpha = 1, \dots, n - 1$ , where  $[\cdot, \cdot]$  is the Poisson bracket (1.3).*

(2) *The  $\{A^{(\alpha)}\}$  are in involution:  $[A^{(\alpha)}, A^{(\beta)}] = 0$ ,  $1 \leq \alpha, \beta \leq n - 1$ .*

(3) *The set  $\{H, A^{(1)}, \dots, A^{(n-1)}\}$  is linearly independent (as  $n$  quadratic forms).*

(4) *At least one of the quadratic forms, say  $A^{(1)}$ , has simple roots.*

(5) *In a local coordinate system  $\{x^l\}$  the quadratic forms satisfy the algebraic commutation property,*

$$(3.13) \quad a^{(\alpha)}_{ij} a^{(\beta)j}_k = a^{(\beta)}_{ij} a^{(\alpha)j}_k.$$

(This property is clearly independent of the coordinates chosen.)

*Proof.* Suppose conditions (1)–(5) are satisfied. Conditions (4) and (5) imply that the quadratic forms can be simultaneously diagonalized by a family of orthonormal vector fields. In the local coordinates  $\{x^j\}$  we have

$$(3.14) \quad (a^{(\alpha)}_{ij} - \rho^{(\alpha)}_h g_{ij}) \lambda^i_{(h)} = 0, \quad h = 1, \dots, n, \quad \alpha = 1, \dots, n - 1,$$

where  $\rho^{(\alpha)}_1, \dots, \rho^{(\alpha)}_n$  are the roots of  $a^{(\alpha)}_{ij}$  and  $\lambda^i_{(h)} \lambda_{(k)i} = \delta_{hk}$ . Setting  $\rho^{(n)}_h = 1$ , for  $h = 1, \dots, n$  we can express condition (3) as

$$(3.14') \quad \det(\rho^{(l)}_m) \neq 0, \quad l, m = 1, \dots, n.$$

Using (1.10), (3.11), and (3.14), we see that condition (1) implies

$$(3.15) \quad \begin{vmatrix} \rho^{(\alpha)}_l & \rho^{(\alpha)}_h & \rho^{(\alpha)}_m \\ 1 & 1 & 1 \\ \gamma_{mhl} & \gamma_{lmh} & \gamma_{hlm} \end{vmatrix} = 0, \quad 1 \leq \alpha \leq n - 1, \quad h, l, m \text{ distinct}$$

and condition (2) implies

$$(3.16) \quad \begin{vmatrix} \rho^{(\alpha)}_l & \rho^{(\alpha)}_h & \rho^{(\alpha)}_m \\ \rho^{(\beta)}_l & \rho^{(\beta)}_h & \rho^{(\beta)}_m \\ \gamma_{hlm} + \gamma_{lmh} & \gamma_{hlm} + \gamma_{mhl} & \gamma_{mhl} + \gamma_{lmh} \end{vmatrix} = 0, \quad 1 \leq \alpha < \beta \leq n - 1.$$

From (3.14'), (3.15) we find  $\gamma_{mhl} = \gamma_{lmh} = \gamma_{hlm}$ . Substituting this result into (3.16) and using (3.14') again, we obtain  $\gamma_{mhl} = \gamma_{lmh} = \gamma_{hlm} = 0$ . Hence, by (3.8) the vector fields  $\{\lambda^i_{(h)}\}$  are normalizable. Theorem 2 can now be invoked to show that the constants of the motion  $A^{(1)}, \dots, A^{(n-1)}$  determine an orthogonal separable coordinate system  $\{y^j\}$ .

Conversely, if we are given an orthogonal separable coordinate system  $\{y^j\}$  for (1.1) we can apply Theorem 2 and reverse the preceding argument to show that conditions (1)–(5) hold. Q.E.D.

We emphasize the importance of the algebraic property (3.13). In flat space with Cartesian coordinates this property simply means that the matrices of the quadratic forms pairwise commute.

As a simple example of the application of Theorem 6 we consider 3-dimensional flat space. A basis for the vector space of second order Killing tensors can be read off from Corollary 2. Now consider the Killing tensor  $m_{12}p_3$ . It is easily verified that the subspace of Killing tensors in involution with  $m_{12}p_3$  is spanned by  $m_{12}p_3$ ,  $m_{12}m_{12}$  and  $p_3p_3$ . In Cartesian coordinates  $\{x^i\}$ , the matrices of these quadratic forms are

$$(3.17) \quad \begin{aligned} m_{12}p_3 &\sim \begin{pmatrix} 0 & 0 & -x^2 \\ 0 & 0 & x^1 \\ -x^2 & x^1 & 0 \end{pmatrix}, & m_{12}m_{12} &\sim \begin{pmatrix} (x^2)^2 & -x^1x^2 & 0 \\ -x^1x^2 & (x^1)^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ p_3p_3 &\sim \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Now  $m_{12}p_3$  has simple roots, but its matrix does not commute with the matrices of  $m_{12}m_{12}$  or  $p_3p_3$ . Hence,  $m_{12}p_3$  does not correspond to a separable coordinate system. On the other hand, the matrices of  $m_{12}m_{12}$  and  $p_3p_3$  do commute and, say,  $m_{12}m_{12} + p_3p_3$  has simple roots. Thus  $m_{12}m_{12}$  and  $p_3p_3$  define a separable coordinate system (cylindrical coordinates).

In his proof of Theorem 2, Eisenhart essentially proves, but does not state, the following result:

**COROLLARY 4.** *Let  $a_{ij}$  be a Killing tensor with simple roots and suppose the associated eigenvector fields  $\{\lambda^i_{(h)}\}$  are normalizable; i.e., suppose condition (3.12) is satisfied. Then there is a unique orthogonal separable coordinate system for (1.1) associated with  $a_{ij}$ .*

For example one can verify that  $m_{12}m_{12} + p_3p_3$  satisfies the hypotheses of Corollary 3 but that  $m_{12}p_3$  violates condition (3.12).

Another interesting example concerns the complex unit sphere  $S_3$ :  $(z^1)^2 + (z^2)^2 + (z^3)^2 + (z^4)^2 = 1$ . We choose complex coordinates

$$(3.18) \quad \mathbf{z} = (\cos x^1 \cos x^2, \cos x^1 \sin x^2, \sin x^1 \cos x^3, \sin x^1 \sin x^3),$$

in which case the metric is

$$(3.19) \quad \varphi = g_{ij} dx^i dx^j = (dx^1)^2 + \cos^2 x^1 (dx^2)^2 + \sin^2 x^1 (dx^3)^2.$$

Consider the Killing tensors

$$(3.20) \quad \begin{aligned} A^{(1)} &= (p_2 + p_3)^2 = a^{ij} p_i p_j, \\ A^{(2)} &= (p_2 - p_3)^2 = b^{ij} p_i p_j. \end{aligned}$$

Here  $[A^{(1)}, A^{(2)}] = 0$ . We claim that these tensors do not define a separation of variables for the Hamilton-Jacobi equation

$$(3.21) \quad p_1^2 + \cos^{-2} x^1 p_2^2 + \sin^{-2} x^1 p_3^2 = E, \quad p_i = \frac{\partial W}{\partial x^i}.$$

(Note that  $p_2$  and  $p_3$  are Killing vectors, hence they belong to the symmetry algebra  $O(4)$ ). We have

$$(3.22) \quad \begin{aligned} \psi_1 &= a_{ij} dx^i dx^j = (\cos^2 x^1 dx^2 + \sin^2 x^1 dx^3)^2, \\ \psi_2 &= b_{ij} dx^i dx^j = (\cos^2 x^1 dx^2 - \sin^2 x^1 dx^3)^2, \end{aligned}$$

and a direct computation shows that the algebraic condition (3.13) is violated. By

considering the 4-dimensional vector space of all Killing tensors in involution with  $A^{(1)}$ , we can similarly show that  $A^{(1)}$  cannot correspond to any orthogonal separable system. Another way to see this is to make use of Eisenhart's results on normalizability of eigenvector fields associated with tensors whose roots are not simple. For this example

$$\det(a_{ij} - \rho g_{ij}) = -\sin^2 x^1 \cos^2 x^1 \rho^2 (\rho - 1),$$

and it is easy to show that  $a_{ij}$  is of type [(11)1]; see the following section. Eisenhart proved that the necessary and sufficient conditions for normalizability of a type [(11)1] tensor are

$$(3.23) \quad \begin{aligned} a_{ij,k} \lambda_{(h)}^i \lambda_{(l)}^j \lambda_{(m)}^k &= 0, & h, l, m \text{ distinct,} \\ a_{ij,k} \lambda_{(1)}^i (\lambda_{(3)}^j \lambda_{(3)}^k - \lambda_{(2)}^j \lambda_{(2)}^k) &= 0, \end{aligned}$$

where  $\lambda_{(1)}^i$  is a unit vector corresponding to the simple root, and  $\lambda_{(2)}^i, \lambda_{(3)}^i$  are mutually orthogonal unit vectors corresponding to the double root [18]. Choosing

$$\lambda_{(1)} = (0, 1, -1), \quad \lambda_{(2)} = (1, 0, 0), \quad \lambda_{(3)} = \frac{1}{\sin x^1 \cos x^1} (1, \sin^2 x^1, -\cos^2 x^1),$$

we find

$$a_{ij,k} \lambda_{(1)}^i \lambda_{(3)}^j \lambda_{(2)}^k = -\cos 2x^1 \neq 0.$$

This shows that  $A^{(1)}$  can never be associated with a separable orthogonal coordinate system for the Hamilton-Jacobi equation (3.21).

**4. Separable systems in 3 variables.** For the Hamilton-Jacobi equation on 3-dimensional Riemannian spaces, both orthogonal and nonorthogonal separable coordinate systems occur. The characterization of orthogonal separable coordinates follows from our general Theorem 6. Nonorthogonal systems have been classified in [2]. (The classification of nonorthogonal systems for  $n = 4$  can be found in [3]. For general  $n$  the classification is very complicated and has not yet been worked out.) The nonorthogonal systems are of two types:

(1) 2 ignorable variables. The metric is

$$(4.1) \quad ds^2 = g_{ij}(x^3) dx^i dx^j,$$

where not all of the terms  $g_{st}(x^3)$ ,  $s \neq t$ , vanish. (Here we regard two separable coordinate systems  $\{\bar{x}^j\}$  and  $\{x^j\}$  as equivalent if

$$(4.2) \quad \begin{aligned} \bar{x}^1 &= ax^1 + bx^2, & \bar{x}^2 &= cx^1 + dx^2, & \bar{x}^3 &= x^3, \\ \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\neq 0; \end{aligned}$$

or

$$(4.3) \quad \bar{x}^1 = x^1 + h(x^3), \quad \bar{x}^2 = x^2 + l(x^3), \quad \bar{x}^3 = f(x^3);$$

or if  $\{\bar{x}^j\}$  is obtained from  $\{x^j\}$  by a succession of these transformations, since all transformations preserve additive separation. Furthermore, if under such transformations  $\{x^j\}$  is equivalent to an orthogonal separable system, then we do not regard  $\{x^j\}$  as truly nonorthogonal.) Each such system is characterized by a pair of Killing vectors  $L_1, L_2$  which are in involution with one another and with the Hamiltonian  $H$ . Here  $L_1 = p_1, L_2 = p_2$ . It follows that these systems can be classified by determining the conjugacy classes of 2-dimensional Abelian subalgebras of the Lie symmetry algebra of the Hamilton-Jacobi equation.

(2) 1 ignorable variable. The metric is

$$(4.4) \quad ds^2 = [U(x^2) + V(x^3)][B(x^3)(dx^2)^2 + 2 dx^1 dx^2 + (dx^3)^2],$$

and the Hamilton-Jacobi equation takes the form

$$(4.5) \quad (U + V)^{-1}[-Bp_1^2 + 2p_1p_2 + p_3^2] = E, \quad p_i = \frac{\partial W}{\partial x^i}.$$

This system is characterized by the Killing vector-Killing tensor pair

$$(4.6) \quad L_1 = p_1, \quad A = (U + V)^{-1}[U(p_3^2 - Bp_1^2) - 2Vp_1p_2].$$

Here,

$$(4.7) \quad [L_1, H] = [A, H] = [L_1, A] = 0.$$

In order to derive results analogous to those of § 2, we recall the standard forms of two quadratic forms in three variables, corresponding to the various possible elementary divisors [12]. Any pair of quadratic differential forms can be reduced to one of these types at a given point if it has the corresponding elementary divisors at that point.

$$\begin{aligned} [111]: \quad \varphi &= g_{ij} dx^i dx^j = c_1(dx^1)^2 + c_2(dx^2)^2 + c_3(dx^3)^2, \\ \psi &= a_{ij} dx^i dx^j = c_1\rho_1(dx^1)^2 + c_2\rho_2(dx^2)^2 + c_3\rho_3(dx^3)^2. \end{aligned}$$

Here the  $\rho_i$  are the (distinct) roots of the equation  $\det(a_{ij} - \rho g_{ij}) = 0$ .

$$[(11)1]: \quad \text{Same as [111] but with } \rho_1 = \rho_2.$$

$$[[111]]: \quad \text{Same as [111] but with } \rho_1 = \rho_2 = \rho_3.$$

$$\begin{aligned} [21]: \quad \psi &= a(dx^2)^2 + 2 dx^1 dx^2 + b(dx^3)^2, \\ \psi &= A(dx^2)^2 + 2\rho_1 dx^1 dx^2 + \rho_2 b(dx^3)^2. \end{aligned}$$

Here the roots are  $\rho_1, \rho_1, \rho_2$ , but the forms cannot be simultaneously diagonalized. (It is possible to take  $a = 0$ , see [12], but the above expressions are more convenient for our purposes.)

$$[21]: \quad \text{Same as [21] but with } \rho_1 = \rho_2.$$

$$\begin{aligned} [3]: \quad \varphi &= 2 dx^1 dx^2 + a(dx^3)^2, \\ \psi &= 2\rho_1 dx^1 dx^2 + 2 dx^2 dx^3 + \rho_1 a(dx^3)^2. \end{aligned}$$

It is clear that Killing tensors corresponding to orthogonal separable systems are of types [111] and [(11)1], whereas those corresponding to nonorthogonal systems (2) above are of types [21] and [(11)1]. The only type [(111)] Killing tensors are constant multiples of  $\{g\}$ . Killing tensors of other types do not correspond to variable separation.

**THEOREM 7.** *Let  $H = g^{ij}p_i p_j$ ,  $A = a^{ij}p_i p_j$  and  $L_1 = \xi^i p_i$ . Necessary and sufficient conditions that a nonorthogonal separable coordinate system for*

$$g^{ij} \frac{\partial W}{\partial x^i} \frac{\partial W}{\partial x^j} = E$$

*be associated with the pair  $L_1, A$  are*

$$(1) [L_1, H] = [L_1, A] = 0.$$

(2) *There exist linear functions  $L_2 = \eta^i p_i$  and  $L_3 = \mu^i p_i$  such that  $\{L_1, L_2, L_3\}$  is linearly independent and  $[L_i, L_j] = 0, i, j = 1, 2, 3$ .*

(3)  $L_1$  is null, i.e.,  $L_1 \cdot L_1 = \xi^i g_{ij} \xi^j = 0$ .

(4)  $L_1$  is an eigenvector of the dual pair  $\hat{A}, \hat{H}$  corresponding to eigenvalue

$$\lambda_3 : (a_{ij} - \lambda_3 g_{ij}) \xi^i = 0, \quad i = 1, 2, 3.$$

(5)  $L_3$  is an eigenvector of  $\hat{A}, \hat{H}$  corresponding to eigenvalue

$$\lambda_2 (\lambda_3 \neq \lambda_2) : (a_{ij} - \lambda_2 g_{ij}) \mu^j = 0, \quad i = 1, 2, 3.$$

(6)  $L_2 \cdot L_3 = \eta^i g_{ij} \mu^j = 0$ .

(7)  $[H, A] = 0$ .

*Proof.* It is easily verified that the pair (4.6) and  $L_2 = p_2, L_3 = p_3$  satisfy conditions (1)–(7). Conversely, suppose conditions (1)–(7) are satisfied. From condition (2) we can introduce new coordinates  $\{x^i\}$  such that  $L_i = p_i, i = 1, 2, 3$ . It is easy to verify that conditions (1)–(6) imply

$$(4.8) \quad (g_{ij}) = \begin{pmatrix} 0 & g_{12} & 0 \\ g_{21} & g_{22} & 0 \\ 0 & 0 & g_{33} \end{pmatrix}, \quad (a_{ij}) = \begin{pmatrix} 0 & \lambda_3 g_{12} & 0 \\ \lambda_3 g_{21} & a_{22} & 0 \\ 0 & 0 & \lambda_2 g_{33} \end{pmatrix},$$

where the matrix elements are independent of  $x^1$ . Then condition (7) is equivalent to

$$(4.9) \quad \begin{aligned} (\alpha) \quad & \partial_3(\lambda_3 g_{12}) + (\lambda_2 - 2\lambda_3) \partial_3 g_{12} = 0, \\ (\beta) \quad & \partial_2 a_{22} + \lambda_3 \left( \frac{2g_{22}}{g_{12}} \partial_2(g_{12}) - \partial_2 g_{22} \right) - 2 \frac{a_{22}}{g_{12}} \partial_2 g_{12} = 0, \\ (\gamma) \quad & \partial_3 a_{22} - 2\lambda_3 g_{12} \partial_3 \left( \frac{g_{22}}{g_{12}} \right) - 2 \frac{a_{22}}{g_{12}} \partial_3 g_{12} + \lambda_2 \partial_3 g_{22} = 0, \\ (\delta) \quad & \partial_3 \lambda_2 = 0, \quad \partial_2 \lambda_3 = 0, \\ (\varphi) \quad & \partial_2(\lambda_2 g_{23}) + (\lambda_3 - 2\lambda_2) \partial_2 g_{33} = 0. \end{aligned}$$

Equation  $(\delta)$  implies  $\lambda_2 = \lambda_2(x^2), \lambda_3 = \lambda_3(x^3)$ , and then  $(\alpha)$  implies  $g_{12} = f(x^2)(\lambda_2 - \lambda_3)$ . Similarly,  $(\varphi)$  implies  $g_{33} = h(x^3)(\lambda_2 - \lambda_3)$ . Substituting these values in  $(\beta)$  we find  $a_{22} - \lambda_3 g_{22} = (g_{12})^2 S(x^3)$ . Then substitution of this expression into  $(\gamma)$  yields

$$\frac{g_{22}}{\lambda_2 - \lambda_3} + S(x^3) f^2(x^2) = K(x^2).$$

Making an appropriate change of variable  $x^1 \rightarrow x^1 + q(x^2)$  followed by a change of scale in  $x^2, x^3$  we can assume  $K \equiv 0, h \equiv 1, f \equiv 1$ . Thus

$$ds^2 = (\lambda_2 - \lambda_3) [2 dx^1 dx^2 - S(x^3)(dx^2)^2 + (dx^3)^2],$$

which is the same as (4.4). Q.E.D.

Note that the conditions of Theorem 7 can be checked in practice. Indeed, if  $L_1, H$ , and  $A$  are given, we first check (1), (3), (4), and (7). Then we use (5) and  $[L_1, L_3] = 0$  to see if a suitable  $L_3$  can be constructed. If successful, we then try to construct  $L_2$  subject to  $[L_2, L_j] = 0, j = 1, 3$ , and condition (6). It is not difficult to construct examples showing that the theorem is false unless condition (6) holds.

In [2] all nonorthogonal flat space separable systems were constructed for  $n = 3$ , and it was shown that these systems all correspond to systems that separate the heat equation in 2-dimensional spacetime. It was also shown that for  $n = 3$  a space of nonzero constant curvature possesses exactly one nonorthogonal separable metric (4.1), and no metrics (4.4).

*Note added in proof.* The principal facts presented in our discussion of the structure of the space of second order Killing tensors for  $V_2$  can all be found in the note, *Sur les géodésiques à intégrals quadratiques*, by M. G. Koenigs (in G. Darboux, *Théorie générale des surfaces*, Vol. IV, 1896, pp. 368–404, reprinted by Chelsea, Bronx NY, 1972). Koenigs considers only the case  $n = 2$ .

## REFERENCES

- [1] L. P. EISENHART, *Riemannian Geometry*, Princeton University Press, Princeton, 1949.
- [2] E. G. KALNINS AND W. MILLER, JR., *Separable coordinates for three-dimensional complex Riemannian spaces*, J. Differential Geom., to appear.
- [3] C. P. BOYER, E. G. KALNINS, AND W. MILLER, JR., *Separable coordinates for four-dimensional Riemannian spaces*. Comm. Math. Phys., 59 (1978), pp. 285–302.
- [4] ———, *R-separable coordinates for three-dimensional complex Riemannian spaces*, Trans. Amer. Math. Soc., 242 (1978), pp. 355–376.
- [5] E. G. KALNINS AND W. MILLER, JR., *R-separation of variables for the Four-dimensional flat space Laplace and Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 244 (1978), pp. 241–261.
- [6] L. P. EISENHART, *Separable systems of Stäckel*, Ann. Math., 35 (1934), pp. 284–305.
- [7] T. H. KOORNWINDER, *A precise definition of separation of variables*, Proceedings of the Scheveningen Conference on Differential Equations (August 1979), Lecture Notes in Mathematics, Springer-Verlag, 1980.
- [8] S. BENENTI AND M. FRANCAVIGLIA, *The theory of separability of the Hamilton-Jacobi equation and its applications to general relativity*, in General Relativity and Gravitation, Vol. 1, A. Held, ed., Plenum, New York, 1980.
- [9] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Graduate Text in Mathematics, Vol. 60, Springer-Verlag, New York, 1978, (translated from the 1974 Russian edition by K. Vogtmann and A. Weinstein).
- [10] E. G. KALNINS, *On the separation of variables for the Laplace equation in two- and three-dimensional Minkowski space*, SIAM J. Math. Anal., 6 (1975), pp. 340–374.
- [11] E. G. KALNINS, W. MILLER, JR., AND P. WINTERNITZ, *The group  $O(4)$ , separation of variables and the hydrogen atom*, SIAM J. Appl. Math., 30 (1976), pp. 630–664.
- [12] T. J. I. BROMWICH, *Quadratic Forms and their Classification by Means of Invariant-Factors*, Hafner, New York (reprint), 1906.
- [13] W. MILLER, JR., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.
- [14] P. STÄCKEL, *Habilitationsschrift*, Halle, 1891.
- [15] T. Y. THOMAS, *The fundamental theorem on quadratic first integrals*, Proc. Nat. Acad. Science, 32 (1946), pp. 10–15.
- [16] G. H. KATZIN AND J. LEVINE, *Quadratic first integrals of the geodesics in spaces of constant curvature*, Tensor, 16 (1965), pp. 97–104.
- [17] I. HAUSER AND R. J. MALHIOT, *Structural equations for Killing tensors of order two, II*, J. Math. Phys., 16 (1975), pp. 1625–1629.
- [18] L. P. EISENHART, *Orthogonal systems of hypersurfaces in a general Riemannian space*, Trans. Amer. Math. Soc., 25 (1923), pp. 297–306.

## ON $q$ -BINOMIAL COEFFICIENTS AND SOME STATISTICAL APPLICATIONS\*

B. R. HANDA† AND S. G. MOHANTY‡

**Abstract.** An expression for the number of lattice paths lying between two arbitrary boundaries and having a given area below it, is obtained which is a determinant involving Gaussian polynomials (or  $q$ -binomial coefficients). Some statistical applications are pointed out. A  $q$ -analogue of the Vandermonde type identity is established for a class of coefficients possessing the  $q$ -additive property.

### 1. Introduction and summary. A $q$ -binomial coefficient is defined by

$$(1) \quad \binom{x}{n}_q = \frac{(q^x - 1)(q^{x-1} - 1) \cdots (q^{x-n+1} - 1)}{(q^n - 1)(q^{n-1} - 1) \cdots (q - 1)},$$

where  $x$  and  $q$  are real numbers,  $n$  is an integer, and  $\binom{x}{0}_q = 1$  and  $\binom{x}{n}_q = 0$  when  $x < n$  and  $x$  is a nonnegative integer, or when  $n < 0$ . Clearly, as  $q \rightarrow 1$ ,  $\binom{x}{n}_q$  becomes the usual binomial coefficient  $\binom{x}{n}$ . Expression (1) is known as a Gaussian polynomial. (See [2, p. 51] for historical references.)

When  $x$  is an integer with  $x \geq n$ , Pólya [9] has given a combinatorial interpretation of  $\binom{x}{n}_q$  in terms of lattice paths. Let  $A_l$  denote the number of lattice paths from  $(0, 0)$  to  $(m, n)$  such that the area below each path in the positive quadrant of the  $XY$  plane is  $l$ . The generating function of  $A_l$  is then shown [9] to be

$$(2) \quad \sum_{l=0}^{mn} A_l q^l = \binom{m+n}{n}_q.$$

(For another combinatorial interpretation of  $q$ -binomial coefficients in terms of finite vector spaces, see [10, p. 240]). In § 2 of this paper, our main purpose is to derive an expression for the generating function of the number of paths which lie between two arbitrary boundaries having the area below the path equal to  $l$ . It is observed that the expression is a determinant involving  $q$ -binomial coefficients. Since (2) is a special case of the determinantal expression, the determinant is a generalization of the  $q$ -binomial coefficient, in this lattice path context. The consideration of restriction on paths by boundaries arises in applications which are pointed out.

The following  $q$ -analogue of the Vandermonde convolution formula for binomial coefficients is well known (see [1], [3]), and is a special case of a formula on basic hypergeometric functions due to Heine [5]:

$$(3) \quad \sum_{k=0}^n \binom{x}{k}_q \binom{y}{n-k}_q q^{k(y-n+k)} = \binom{x+y}{n}_q.$$

In an earlier paper [7], the authors generalized the Vandermonde convolution identity for a general class of coefficients with the so-called additive property (defined in [7]). In § 3, a  $q$ -analogue of the generalized Vandermonde type identity is established for a class of coefficients possessing a  $q$ -analogue of the additive property called the  $q$ -additive property.

\* Received by the editors June 1, 1978 and in final revised form March 10, 1980. This work was supported in part by the National Research Council of Canada.

† Department of Mathematics, Indian Institute of Technology-Delhi, New Delhi-29, India.

‡ Department of Mathematics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.



For completeness we state below a few known results [4] on  $q$ -binomial coefficients, some of which will be needed in our discussion:

$$(4) \quad \binom{n}{k}_q = \binom{n}{n-k}_q, \quad 0 \leq k \leq n, \quad n \text{ being a nonnegative integer;}$$

$$(5) \quad \binom{x+1}{k}_q = \binom{x}{k}_q q^k + \binom{x}{k-1}_q = \binom{x}{k}_q + q^{x-k+1} \binom{x}{k-1}_q;$$

$$(6) \quad \binom{-x}{n} = (-1)^n q^{-nx - n(n-1)/2} \binom{x+n-1}{n}_q;$$

$$(7) \quad \prod_{j=0}^{n-1} (1 + sq^j) = \sum_{k=0}^n q^{k(k-1)/2} \binom{n}{k}_q s^k.$$

**2. A counting problem and applications.** Using the usual representation [8] for a minimal lattice path from  $(0, 0)$  to  $(m, n)$  by means of a nondecreasing vector  $(x_1, x_2, \dots, x_n)$  of nonnegative integers where  $x_i$  is the distance of the path from  $(m, n - i)$ , denote by  $A_l(\mathbf{b}; \mathbf{a})$  the number of paths from  $(0, 0)$  to  $(m, n)$  never crossing the paths with vectors  $\mathbf{a} = (a_1, a_2, \dots, a_n)$  and  $\mathbf{b} = (b_1, b_2, \dots, b_n)$ .  $b_i \leq a_i$  for all  $i$  such that the area below the paths in the positive quadrant of the  $XY$  plane is  $l$ . We prove

THEOREM 1.

$$(8) \quad \sum_{l=\beta}^{\alpha} A_l(\mathbf{b}; \mathbf{a}) q^l = \det_{n \times n} \left\| \binom{a_i - b_j + 1}{j - i + 1}_q \right\| q^{(j-i)(j-i+1)/2 + (j-i+1)b_j},$$

where

$$\alpha = \sum_{i=1}^n a_i, \quad \beta = \sum_{i=1}^n b_i, \quad \text{and} \quad \binom{x}{y}_{q^+} = \binom{\max(x, 0)}{y}_q.$$

*Proof.* Since the area below a path with vector  $(x_1, x_2, \dots, x_n)$  is  $\sum_{i=1}^n x_i$ , we have

$$(9) \quad \sum_{l=\beta}^{\alpha} A_l(\mathbf{b}, \mathbf{a}) q^l = \sum_{x_1=b_1}^{a_1} \sum_{x_2=y_2}^{a_2} \dots \sum_{x_n=y_n}^{a_n} q^{x_1 + \dots + x_n},$$

where  $y_i = \max(x_{i-1}, b_i)$ ,  $i = 2, 3, \dots, n$ .

Since the upper triangular determinant

$$\det_{n \times n} \left\| \binom{x_i - b_j}{j - i}_q \right\| q^{(j-i)(j-i+1)/2 + (j-i)b_j}$$

is equal to 1, the right-hand side of (9) can be written as

$$(10) \quad \sum_{x_1=b_1}^{a_1} \sum_{x_2=y_2}^{a_2} \dots \sum_{x_n=y_n}^{a_n} \det_{n \times n} \left\| \binom{x_i - b_j}{j - i}_q \right\| q^{x_i - b_j - j + i} q^{(j-i)(j-i+1)/2 + (j-i+1)b_j}.$$

From the summation formula

$$\sum_{x=p}^l q^{x-h} \binom{x}{h}_q = \binom{l+1}{h+1}_q - \binom{p}{h+1}_q$$

(which easily follows from (5)), we have, for  $i \leq j$ ,

$$(11) \quad \sum_{x_i=y_i}^{a_i} q^{x_i - b_j - j + i} \binom{x_i - b_j}{j - i}_q = \binom{a_i - b_j + 1}{j - i + 1}_q - \binom{x_{i-1} - b_j}{j - i + 1}_q,$$

since  $b_i \leq b_j$ .

Consider the summation on  $x_i$ , which only appears in the  $i$ th row of the determinant in (10). Using (11), we obtain the  $j$ th element equal to

$$(12) \quad q^{(j-i)(j-i+1)/2+(j-i+1)b_j} \left[ \binom{a_i - b_j + 1}{j - i + 1}_{q^+} - \binom{x_{i-1} - b_j}{j - i + 1}_{q^+} \right] \quad \text{for } i \leq j,$$

and equal to 0 for  $i > j$ .

Now, if we add  $q^{-x_{i-1}}$  times the  $(i - 1)$ st row to the  $i$ th row determined by (12), the  $i$ th row so obtained is equal to the  $i$ th row of (8). Proceeding in this manner and summing over  $x_n, x_{n-1}, \dots, x_1$  successively, we get the desired result (8). This completes the proof.

In particular, if we set  $b_i = 0$ , and  $a_i = m$  for all  $i$  in (8), we should obtain (2).

From Theorem 1, we have

$$\sum_{l=0}^{mn} A_l = \det_{n \times n} \left\| \binom{m + 1}{j - i + 1}_q q^{(j-i)(j-i+1)/2} \right\|.$$

In order to simplify the above determinant, we replace the first row of the determinant by

$$\sum_{i=1}^n \binom{-(m + 1)}{i - 1}_q q^{(i-1)(-2)/2+(i-1)(m+1)} \times (\text{row } i).$$

The new first row of the determinant then becomes

$$\left[ 0, 0, \dots, 0, q^{n(n-1)/2+n(m+1)} \binom{-(m + 1)}{n}_q \right]$$

by way of (3). Now expanding the determinant by the first row gives the value of the determinant as

$$(-1)^{n+2} \binom{-(m + 1)}{n}_q q^{n(n-1)/2+n(m+1)},$$

which simplifies to  $\binom{m + n}{n}_q$  with the help of (6). This checks (2).

The special case when  $q = 1$  gives the number of paths lying between **a** and **b**. This result was first obtained by Kreweras [6].

Letting  $D_0 = 1$  and

$$D_j = \sum_{l=\beta_j}^{\alpha_j} A_l(b_1, \dots, b_j; a_1, \dots, a_j) q^l,$$

where

$$\alpha_j = \sum_{i=1}^j a_i \quad \text{and} \quad \beta_j = \sum_{i=1}^j b_i, \quad j = 1, \dots, n,$$

we remark that Theorem 1 is equivalent to the following recurrence relation on the  $D_j$ 's which is derived by expanding the determinant by its last column:

$$(13) \quad D_n = \sum_{i=0}^{n-1} (-1)^i \binom{a_{n-i} - b_n + 1}{i + 1}_{q^+} q^{i(i+1)/2+(i+1)b_n} D_{n-i-1}.$$

A direct combinatorial proof of relation (13) or its equivalent

$$(14) \quad \sum_{i=0}^n (-1)^{n-i} \binom{a_{i+1} - b_n + 1}{n-i}_{q^+} q^{(n-i)(n-i-1)/2+(n-i)b_n} D_i = 0,$$

as suggested by the referee, is provided below.

Note that when  $j = n$ ,

$$\binom{a_{n+1} - b_n + 1}{n-j}_{q^+} = 1,$$

and therefore the value of  $a_{n+1}$  is of no consequence.

Consider the sets  $S_j = \{(x_1, \dots, x_n) : x_1 \leq \dots \leq x_j, x_{j+1} > \dots > x_n, b_i \leq x_i \leq a_i \text{ for all } i\}$ ,  $j = 0, 1, \dots, n$ , and assign the weight  $(-1)^{n-j} q^{x_1 + \dots + x_n}$  to each sequence in  $S_j$ , for all  $j$ . We want to sum these weights for all sequences. It can be seen that the sum of the weights for sequences in  $S_j$  can be written as

$$\left( \sum_{R_1} q^{x_1 + \dots + x_j} \right) \left( \sum_{R_2} (-1)^{n-j} q^{x_{j+1} + \dots + x_n} \right),$$

where

$$R_1 = \{(x_1, \dots, x_j) : x_1 \leq \dots \leq x_j, b_i \leq x_i \leq a_i, i = 1, \dots, j\}$$

and

$$R_2 = \{(x_{j+1}, \dots, x_n) : x_{j+1} > \dots > x_n, b_i \leq x_i \leq a_i, i = j+1, \dots, n\} \\ = \{(x_{j+1}, \dots, x_n) : a_{j+1} \geq x_{j+1} > \dots > x_n \geq b_n\}.$$

But

$$\sum_{R_1} q^{x_1 + \dots + x_j} = D_j,$$

and

$$\sum_{R_2} (-1)^{n-j} q^{x_{j+1} + \dots + x_n} = (-1)^{n-j} \binom{a_{j+1} - b_n + 1}{n-j}_{q^+} q^{(n-j)(n-j-1)/2+(n-j)b_n}.$$

This checks with the left-hand side of (14).

We derive the sum of the weights in another way. Observe that any sequence  $(x_1, \dots, x_n)$  in  $S_j$  belongs either to  $S_{j-1}$  if  $x_j > x_{j+1}$  or to  $S_{j+1}$  if  $x_j \leq x_{j+1}$ . Therefore, the weights for any given  $(x_1, \dots, x_n)$  cancel, occurring twice with opposite signs. Hence the sum of the weights over all sequences equals zero, which is the right-hand side of (14). This completes the proof.

As an application of Theorem 1, we assert that it provides an expression for the probability generating function of the joint probability distribution of the two-sided Kolmogorov-Smirnov statistic and the Wilcoxon-Mann-Whitney statistic. The Wilcoxon-Mann-Whitney statistic is a linear function of the rank sum statistic  $U = \sum_{i=1}^m R_i$ , where  $R_i$  is the number of  $Y$  observations that precede the  $i$ th largest  $X$  observation in the two independent samples  $(X_1, \dots, X_m)$  and  $(Y_1, Y_2, \dots, Y_n)$  obtained from a continuous distribution. Represent the  $j$ th element of the combined ordered sample by a horizontal unit if it is an  $X$  or by a vertical unit if it is a  $Y$ . Then the combined ordered sample is represented by a path from  $(0, 0)$  to  $(m, n)$  with  $m$  horizontal unit steps and  $n$  vertical unit steps. In that case the rank sum statistic is the area under the corresponding

path. Also, for the Kolmogorov-Smirnov statistic

$$D_{m,n} = \sup_{-\infty < x < \infty} |F_m(x) - G_n(x)|,$$

where  $F_m$  and  $G_n$  are the empirical distribution functions, it is well known that the event  $\{D_{m,n} \leq c\}$  corresponds to the set of paths from  $(0, 0)$  to  $(m, n)$  which do not cross the lines  $nx = my \pm mnc$ .

Hence if  $(a_1, \dots, a_n)$  and  $(b_1, \dots, b_n)$  are the vectors corresponding to the path boundaries  $nx = my + mnc$  and  $nx = my - mnc$ , we have, by Theorem 1,

$$P(D_{m,n} \leq c, U = l) = \frac{A_l(\mathbf{b}; \mathbf{a})}{\binom{m+n}{n}}$$

This proves our assertion.

It is well known that any partition of  $l$  into at most  $n$  parts can be represented by a nondecreasing vector  $(x_1, \dots, x_n)$  of nonnegative integers with the property  $\sum_{i=1}^n x_i = l$ .  $A_l(\mathbf{b}; \mathbf{a})$  then represents the number of partitions  $(x_1, \dots, x_n)$  of  $l$  into at most  $n$  parts such that  $b_j \leq x_j \leq a_j, j = 1, \dots, n$ . Thus Theorem 1 provides an expression for the generating function of such partitions.

The following result provides another application of Theorem 1.

Let  $X_1, X_2, \dots, X_n$  be a random sample from the geometric distribution  $P(X = x) = pq^x, x = 0, 1, \dots, p + q = 1, 0 < p, q < 1$ . Then for any nondecreasing nonnegative integer vectors  $(a_1, \dots, a_n)$  and  $(b_1, \dots, b_n)$  such that  $b_i \leq a_i$  for all  $i$ ,

$$\begin{aligned} &P(b_i \leq X_i \leq a_i, i = 1, 2, \dots, n, 0 \leq X_1 \leq X_2 \leq \dots \leq X_n) \\ (15) \quad &= p^n \det_{n \times n} \left\| \binom{a_i - b_j + 1}{j - i + 1}_{q^+} q^{(j-i)(j-i+1)/2 + (j-i+1)b_j} \right\|. \end{aligned}$$

The result follows immediately, because the required probability is given by the right-hand side of (9) multiplied by  $p^n$ .

If the  $x$ 's in (9) are real numbers instead of integers, we get

THEOREM 2.

$$\begin{aligned} (16) \quad &\int_{x_1=b_1}^{a_1} \int_{x_2=y_2}^{a_2} \dots \int_{x_n=y_n}^{a_n} q^{x_1+\dots+x_n} dx_1 \dots dx_n \\ &= \det_{n \times n} \left\| q^{b_j(j-i+1)} \left( \frac{q^{(a_i-b_j)_+} - 1}{\log q} \right)^{j-i+1} / (j-i+1)! \right\|, \end{aligned}$$

where  $(x)_+ = \max(0, x)$ .

*Proof.* As in Theorem 1, notice that the integrand  $q^{x_1+\dots+x_n}$  can be multiplied by the triangular determinant

$$\det_{n \times n} \left\| q^{b_j(j-i)} \left( \frac{q^{(x_i-b_j)_+} - 1}{\log q} \right)^{j-i} / (j-i)! \right\|,$$

which has value unity. Proceeding in the same manner as in the proof of Theorem 1, we

can get (16) by observing that, for  $i < j$ ,

$$\int_{y_i}^{a_i} \frac{q^{x_i+b_j(j-i)}}{(j-i)!} \left( \frac{q^{(x_j-b_j)_+} - 1}{\log q} \right)^{j-i} dx_i$$

$$= \frac{q^{b_j(j-i+1)}}{(j-i+1)!} \left( \frac{q^{(a_i-b_j)_+} - 1}{\log q} \right)^{j-i+1} - \frac{q^{b_j(j-i+1)}}{(j-i+1)!} \left( \frac{q^{(x_{i-1}-b_j)_+} - 1}{\log q} \right)^{j-i+1}.$$

As an application of (16) we have the following result.

Let  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  be the order statistic of a random sample of size  $n$  from a distribution with pdf  $((\log q)/(q^b - q^a))q^x, a \leq x \leq b, 0 < q < 1$ . Then for nonnegative nondecreasing vectors  $(a_1, \dots, a_n)$  and  $(b_1, \dots, b_n)$  having components in the interval  $[a, b]$ , we have

$$(17) \quad P(b_i \leq X_{(i)} \leq a_i, i = 1, 2, \dots, n)$$

$$= n! \left( \frac{\log q}{q^b - q^a} \right)^n \det_{n \times n} \left\| \frac{q^{b_j(j-i+1)}}{(j-i+1)!} \left( \frac{q^{(a_i-b_j)_+} - 1}{\log q} \right)^{j-i+1} \right\|.$$

For the proof, we recall that the joint pdf of  $X_{(1)}, \dots, X_{(n)}$  is  $n! ((\log q)/(q^b - q^a))^n q^{x_1 + \dots + x_n}, a < x_1 < x_2 < \dots < x_n < b$ . Hence the expression for the required probability is the expression (16) multiplied by the factor  $n! ((\log q)/(q^b - q^a))^n$ .

If we put  $a = 0, b = 1$  and  $q \rightarrow 1$  in the above, the distribution tends to unit uniform distribution and the right-hand side (17) reduces to

$$n! \det_{n \times n} \left\| \frac{(a_i - b_j)_+^{j-i+1}}{(j-i+1)!} \right\|.$$

This result for the uniform distribution was obtained by Steck in [11].

**3. A  $q$ -analogue of the Vandermonde type convolution.** In this section, we present a  $q$ -analogue of the Vandermonde type convolution (see [7]) for a class of coefficients defined below.

Consider a sequence  $\{f_n(\xi, q)\}$  with the properties

$$f_n(\xi, q) = 1 \quad \text{and} \quad f_n(0, q) = \delta_n^0 = \begin{cases} 1 & \text{for } n = 0, \\ 0 & \text{for } n \neq 0. \end{cases}$$

For any sequence  $b_1, b_2, \dots, a$   $g(q)$ -coefficient denoted by  $g(0, b_k, b_{k+1}, \dots, b_n)_q, 1 \leq k \leq n, n \geq 1$ , is defined by

$$(18) \quad g(0)_q = 1,$$

$$g(0, b_k, \dots, b_n)_q = \det_{(n-k+1) \times (n-k+1)} \| f_{j-i+1}(b_{k+i-1}, q) \|.$$

We remark that apart from a slight change in the notation, the definition of  $g(q)$ -coefficient given above is exactly the same as that of  $g$ -coefficient in § 2 of [7], except that the sequence  $\{f_n\}$  now depends on an additional parameter  $q$ . In particular, if

$$(19) \quad f_n(\xi, q) = q^{n(n-1)/2} \binom{\xi}{n}_q, \quad n \geq 0,$$

we refer to  $g(q)$ -coefficient as  $g^*(q)$ -coefficient. Note that  $g^*(0, a_1, \dots, a_n)_q$  is the determinant of order  $n$ , which is the same as the determinant on the right-hand side of (8) with  $b_i = 0$  for all  $i$ . Thus the class of  $g(q)$ -coefficients is a wider class which includes those of (19).

Let the generating function of the sequence given by (19) be  $\phi_q(s, \xi)$ . Since

$$\phi_q(s, \xi) = \prod_{n=0}^{\infty} \{(1 + sq^n)/(1 + sq^{\xi+n})\} \quad \text{for } |q| < 1$$

(see [2, Th. 2.1], in which put  $t = -sq^\xi$  and  $a = q^{-\xi}$ ), it is immediate that  $\phi_q(s, \xi)$  must satisfy

$$\begin{aligned} \phi_q(s, x + y) &= \phi_q(q^y s, x) \phi_q(s, y) \\ &= \phi_q(s, x) \phi_q(q^x s, y). \end{aligned}$$

Using this functional equation as a model, we define a generalization of the concept of additivity used in [7]. Let  $G_q(s, \xi)$  be the generating function of the sequence  $\{f_n(\xi, q)\}$ . We say that  $G_q(s, \xi)$  is  $q$ -additive in  $\xi$  if

$$(20) \quad \begin{aligned} G_q(s, x + y) &= G_q(q^y s, x) G_q(s, y) \\ &= G_q(s, x) G_q(q^x s, y). \end{aligned}$$

Then additivity defined in [7] is in fact 1-additivity. Now it is possible to prove a  $q$ -analogue of the Vandermonde type convolution (33) in [7] valid for  $g(q)$ -coefficients.

LEMMA 1. *If the generating function  $G_q(s, \xi)$  of the sequence  $\{f_n(\xi, q)\}$  is  $q$ -additive in  $\xi$ , then*

$$(21) \quad g(0, \underbrace{\xi, \dots, \xi}_r)_q = (-1)^r q^{r\xi} f_r(-\xi, q).$$

Note that this corresponds to Lemma 1 in [7].

LEMMA 2. *Let  $G_m$  and  $M_m$  be square upper triangular matrices of order  $m + 1$  given by*

$$\begin{aligned} G_m &= [g(0, b_i, \dots, b_{i-1})_q], \\ M_m &= [(-1)^{j-i} f_{j-i}(b_i, q)]. \end{aligned}$$

Then

$$(22) \quad G_m M_m = I = M_m G_m \text{ for any } m \geq 0.$$

This is exactly the same as Lemma 2 of [7].

Next define the following infinite order upper triangular matrices:

$$\begin{aligned} G &= [g(0, b_i, b_{i+1}, \dots, b_{j-1})_q], \\ B &= [g(0, a_i - b_i, a_{i+1} - b_i, \dots, a_{j-1} - b_i)_q q^{(j-i)b_i}], \\ G^* &= G \quad \text{with } b_i \text{ replaced by } a_i \text{ for all } i. \end{aligned}$$

THEOREM 3. *A necessary and sufficient condition for the matrix equation*

$$(23) \quad GB = G^*$$

*to hold for any  $a_i$  and  $b_i$  is that  $G_q(s, \xi)$  is  $q$ -additive in  $\xi$ .*

Statement (23) is equivalent to the convolution identity,

$$(24) \quad \begin{aligned} &g(0, a_i - b_i, a_{i+1} - b_i, \dots, a_n - b_i)_q q^{(n-i+1)b_i} \\ &+ \sum_{j=i}^{n-1} \{g(0, a_{j+1} - b_{j+1}, \dots, a_n - b_{j+1})_q q^{(n-j)b_{j+1}} \times g(0, b_i, \dots, b_j)_q\} \\ &+ g(0, b_i, \dots, b_n)_q = g(0, a_i, \dots, a_n)_q, \end{aligned}$$

for  $i = 1, 2, \dots, n - 1, n = 2, 3, \dots$ , and

$$g(0, a_n - b_n)q^{b_n} + g(0, b_n)_q = g(0, a_n)_q, \quad n = 1, 2, \dots.$$

The proofs of all the above results follow essentially the same steps as in the corresponding proofs of [7]. We illustrate this by giving the proof of Theorem 3.

*Proof of Theorem 3.* We write  $G_\infty, M_\infty$  as  $G, M$  respectively and let  $S' = (1, s, s^2, \dots)$ . Then by Lemma 2  $GMS = S$ , which implies that

$$(25) \quad \sum_{n=i}^\infty g(0, b_{i+1}, \dots, b_n)_q s^n G_q(-s, b_{n+1}) = s^i$$

for  $i \geq 0$ . Replacing  $b_r$  by  $a_r - b_{i+1}$  for  $r = i + 1, i + 2, \dots$ , in (25) we obtain the relations

$$(26) \quad \sum_{n=i}^\infty \{g(0, a_{i+1} - b_{i+1}, \dots, a_n - b_{i+1})_q s^n G_q(-s, a_{n+1} - b_{i+1})\} = s^i \quad \text{for } i \geq 0.$$

*Necessity.* Let  $M^*$  denote  $M$  with  $b_j$  replaced by  $a_j$  for all  $j$ . Then

$$(27) \quad \begin{aligned} GB &= G^* \\ \Rightarrow MGBM^*S &= MG^*M^*S \\ \Rightarrow BM^*S &= MS, \quad \text{by (22)} \\ \Rightarrow \sum_{n=i}^\infty \{g(0, a_{i+1} - b_{i+1}, \dots, a_n - b_{i+1})_q q^{(n-i)b_{i+1}} G_q(-s, a_{n+1}) s^n\} \\ &= s^i G(-s, b_{i+1}) \quad \text{for } i \geq 0. \end{aligned}$$

If we put  $b_j = \beta$  for all  $j$  and replace  $s$  by  $q^{-\beta}s$  in (27) we have

$$(28) \quad \sum_{n=i}^\infty \left\{ g(0, q_{i+1} - \beta, \dots, a_n - \beta)_q \frac{G_q(-sq^{-\beta}, a_{n+1})}{G_q(-sq^{-\beta}, \beta)} s^n \right\} = s^i.$$

Comparing (28) with (25) where  $b_j$  is replaced by  $a_j - \beta$  for all  $j$ , we must have

$$G_q(-sq^{-\beta}, a_{n+1}) = G_q(-sq^{-\beta}, \beta) G_q(-s, a_{n+1} - \beta),$$

for  $n \geq 0$ ; i.e.,

$$G_q(s, a_{n+1}) = G_q(s, \beta) G_q(sq^\beta, a_{n+1} - \beta),$$

for  $n \geq 0$ .

*Sufficiency.* Replacing  $s$  by  $q^{b_{i+1}}s$  in (26), and using the fact that  $G_q(s, \xi)$  is  $q$ -additive, reduces (26) to (27). Then by reversing the steps which lead to (27) we get (23). This completes the proof.

We have already given one example of  $g(q)$ -coefficient, viz.,  $g^*(q)$  for which the convolution (24) holds. As another example, if we set

$$f_n(\xi, q) = \frac{\left(\frac{q^\xi - 1}{\log q}\right)^n}{n!}, \quad n \geq 0,$$

we see that the generating function of this sequence is

$$\exp\left\{\frac{q^\xi - 1}{\log q} s\right\},$$

which is clearly  $q$ -additive. Thus the corresponding  $g(q)$ -coefficient would also satisfy

convolution (24). Moreover, this specialized coefficient is the same as the determinant in the right-hand side of (16) with  $b_i = 0$  for all  $i$  and is equal to the  $g_3$ -coefficient in [7], when  $q = 1$ .

Finally we remark that by using Lemma 1, Lemma 2, and Theorem 3, one can extend all the results on inverse series relations described in [7] to  $g(q)$ -coefficients of which (19) is a special case. This is done by changing  $f(\cdot)$  and  $g(\cdot)$  in [7] to  $f(\cdot, q)$  and  $g(0, \cdot)_q$  respectively. For example, two  $q$ -analogue inverse relations are:

$$(i) \quad x_n = \sum_{r=n}^m g(0, b_{n+1}, \dots, b_r)_q y_r$$

if and only if

$$y_n = \sum_{r=n}^m (-1)^{r-n} f_{r-n}(b_{n+1}, q) x_r;$$

$$(ii) \quad x_n = \sum_{r=0}^n g(0, b_{r+1}, \dots, b_n)_q y_r$$

if and only if

$$y_n = \sum_{r=0}^n (-1)^{n-r} f_{n-r}(b_{r+1}, q) x_r.$$

Note that in both (i) and (ii),  $g(0, b_{n+1}, \dots, b_n) = g(0) = 1$ .

**4. Acknowledgment.** Both authors are grateful to the National Research Council, Canada for financial assistance. The authors are also grateful to the referees for their comments, in particular to the one who provided a direct proof to (13) or (14).

REFERENCES

[1] G. E. ANDREWS, *Applications of basic hypergeometric functions*, SIAM Rev. 16 (1974), pp. 441–484.  
 [2] ———, *The Theory of Partitions*, Encyclopedia of Mathematics and its Applications, Vol. 2, Addison Wesley, Reading, MA, 1976.  
 [3] H. W. GOULD, *The  $q$ -series generalization of a formula of Sparre Anderson*, Math. Scand., 9 (1961), pp. 90–94.  
 [4] ———, *The operator  $(a^x \Delta)^n$  and Stirling numbers of the first kind*, Amer. Math. Monthly, 71 (1964), pp. 850–858.  
 [5] E. HEINE, *Handbuch die Kugelfunctionen, Vol. 1: Theorie und Anwendung*, Reimer, Berlin, 1878.  
 [6] G. KREWERAS, *Sur une classe de problèmes de dénombrement liés au treillis des partitions des entiers*, Cahiers du BURO, 6 (1965), pp. 5–105.  
 [7] S. G. MOHANTY AND B. R. HANDA, *A generalized Vandermonde type convolution and associated inverse series relations*, Proc. Camb. Phil. Soc., 68 (1970), pp. 459–474.  
 [8] S. G. MOHANTY AND T. V. NARAYANA, *Some properties of compositions and their application to probability and statistics: I*, Biometrische Zeitschrift, 3 (1961), pp. 252–258.  
 [9] G. PÓLYA, *Gaussian binomial coefficients and enumeration of inversions*, in Proceedings of 2nd Chapel Hill Conference on Combinatorial Mathematics and its Applications, University of North Carolina, Cahiers du BURO, 6 (1965), pp. 5–105.  
 [10] G. C. ROTA AND J. GOLDMAN, *On the foundations of combinatorial theory, IV: Finite vector spaces and Eulerian generating functions*, Studies in Appl. Math., 49 (1970), pp. 239–258.  
 [11] G. P. STECK, *Rectangular probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distributions*, Ann. Math. Stat., 42 (1971), pp. 1–11.



## A PERTURBATION OF AN ABSTRACT VOLTERRA EQUATION\*

T. KIFFE†

**Abstract.** This paper discusses the existence of solutions to equations of the form  $u(t, x) + \int_0^t a(t-s)[Au(s, x) + g(u(s, x))] ds = f(t, x)$  where  $A$  is a differential operator on  $L^2(\Omega)$ ,  $\Omega$  a bounded open subset of  $R^n$ , and  $g$  is a discontinuous real-valued function which is not necessarily monotone increasing.

**1. Introduction.** In this paper we will discuss the existence of solutions to Volterra equations of the form

$$(1.1) \quad u(t, x) + \int_0^t a(t-s)[Au(s, x) + g(u(s, x))] ds = f(t, x), \quad (t, x) \in (0, T) \times \Omega,$$

where  $\Omega$  is a bounded open subset of  $R^n$  with smooth boundary  $\Gamma$ ,  $x = (x_1, \dots, x_n)$ ,  $a$  and  $g$  are real-valued functions and  $A$  can be a differential operator. In recent years equations like (1.1) have been viewed as special cases of abstract Volterra equations of the form

$$(1.2) \quad u(t) + \int_0^t a(t-s)Bu(s) ds \ni f(t), \quad 0 \leq t \leq T,$$

where  $B$  is a nonlinear operator defined on a subset of a Banach space  $X$ ,  $f: [0, T] \rightarrow X$  and  $a: [0, T] \rightarrow R^1$ .

General existence results for (1.2) have been obtained by Barbu [2], Gripenberg [7] and Londen [11] when  $B$  is a maximal monotone operator on a Hilbert space. These results were extended to  $m$ -accretive operators on a Banach space by Crandall and Nohel [6]. Existence results for (1.1) were obtained by taking  $L^2(\Omega)$  as the underlying Hilbert space and requiring  $A$  to be a maximal monotone operator. These results also required that  $g$  be a monotonically increasing function. Without this crucial assumption on  $g$  the above mentioned abstract results do not apply to (1.1). An existence result for (1.1) when  $A$  is a linear, self-adjoint differential operator and  $g$  is Lipschitz continuous was obtained in [8].

In this paper we will extend the known existence results for (1.1) by replacing the monotonicity condition on  $g$  required in [2], [7], [11] and the continuity assumption required in [8] by a growth condition on  $g$ .

The paper will proceed as follows: § 2 will contain the preliminaries for a precise formulation of the existence question, § 3 will contain the proofs of existence, and three examples which illustrate the applicability of our results will be presented in § 4.

**2. Statement and discussion of results.** Let  $\Omega$  be a bounded open subset of  $R^n$  with smooth boundary  $\Gamma$  and let  $L^2(\Omega)$  denote the usual Hilbert space of real-valued square integrable functions defined on  $\Omega$ . The standard norm and inner product on  $L^2(\Omega)$  will be denoted by  $|\cdot|$  and  $\langle \cdot, \cdot \rangle$ , respectively.

Let  $\phi$  be a proper, convex, lower semicontinuous function mapping  $L^2(\Omega)$  into  $(-\infty, \infty]$ , with  $D(\phi) = \{u \in L^2(\Omega) : \phi(u) < \infty\}$ . We shall assume that  $A = \partial\phi$ , where  $\partial\phi$  is the nonlinear, possibly multiple valued operator from  $L^2(\Omega)$  into  $L^2(\Omega)$  defined by  $u \in \partial\phi(v)$  if and only if  $\phi(w) \geq \phi(v) + \langle u, w - v \rangle$ , for every  $w \in L^2(\Omega)$ . Then  $A$  is a maximal monotone operator from its domain  $D(A) \subseteq L^2(\Omega)$  into  $L^2(\Omega)$ . For properties

\* Received by the editors May 17, 1979, and in revised form April 14, 1980.

† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

of such operators see [4]. In addition we shall assume that, for some  $p \geq 1$ ,

$$(A_1) \quad \{u: \phi(u) + |u| \leq K\} \text{ is precompact in } L^2(\Omega) \\ \text{and bounded in } L^{2p}(\Omega) \text{ for each } K > 0.$$

Concerning the function  $g$  we shall assume

$$(2.1) \quad g: R^1 \rightarrow R^1, \text{ } g \text{ is measurable, and} \\ |g(t)| \leq c_1 + c_2|t|^p, \quad -\infty < t < \infty,$$

for some positive constants  $c_1$  and  $c_2$  where  $p$  is given by  $(A_1)$ . If  $(A_1)$  is true with  $p = \infty$ , then (2.1) reduces to  $g \in L^\infty_{loc}(-\infty, \infty)$ . Furthermore we shall also assume that

$$(2.2) \quad G(t) = \int_0^t g(s) ds \geq -c_3 - c_4t^2, \quad -\infty < t < \infty,$$

for some positive constants  $c_3$  and  $c_4$ .

Multiple valued functions play an essential role in solving nonlinear equations like (1.1). If  $g$  is a monotone increasing function there is no difficulty in ‘‘filling in the gaps’’ in the graph of  $g$  to make  $g$  a maximal monotone operator. We will follow the idea of Rauch [12] in extending  $g$  when  $g$  is no longer monotone. To this end we first introduce two auxiliary functions. For each  $\epsilon > 0$  define

$$\bar{g}_\epsilon(t) = \text{ess sup}_{|s-t|<\epsilon} g(s), \quad g_\epsilon(t) = \text{ess inf}_{|s-t|<\epsilon} g(s).$$

For fixed  $t$ ,  $\bar{g}_\epsilon$  is a decreasing function and  $g_\epsilon$  is an increasing function of  $\epsilon$  for decreasing  $\epsilon$ . Let

$$\bar{g}(t) = \lim_{\epsilon \rightarrow 0} \bar{g}_\epsilon(t) \quad \text{and} \quad \underline{g}(t) = \lim_{\epsilon \rightarrow 0} g_\epsilon(t).$$

Then  $\bar{g}$  is upper semicontinuous and  $\underline{g}$  is lower semicontinuous. (Note that if  $g$  has a jump discontinuity at  $t$  then  $\bar{g}(t) = \max(g(t+), g(t-))$  and  $\underline{g}(t) = \min(g(t+), g(t-))$ ). We define a multiple valued function by

$$(2.3) \quad \hat{g}(t) = [g(t), \bar{g}(t)],$$

i.e.,  $s \in \hat{g}(t)$  if and only if  $g(t) \leq s \leq \bar{g}(t)$ . Let  $B$  denote the usual extension of  $g$  to  $L^2(\Omega)$ , i.e.,  $v \in B(u)$  if and only if  $v(x) \in \hat{g}(u(x))$  a.e. on  $\Omega$  and  $v, u \in L^2(\Omega)$ .

We first assume that the kernel function  $a(t)$  satisfies

$$(a_1) \quad a: R^1 \rightarrow R^1, \quad a \in AC[0, T] \quad \text{and} \quad a(0) > 0,$$

where  $' = d/dt$  and  $AC[0, T]$  denotes the set of absolutely continuous functions on  $[0, T]$ .

With these preliminaries the abstract equation we will consider can be written as

$$(2.4) \quad u(t) + \int_0^t a(t-s)Au(s) ds + \int_0^t a(t-s)B(u(s)) ds \ni f(t), \quad 0 \leq t \leq T,$$

where the given function  $f$  and the unknown function  $u$  lie in  $L^2[0, T; L^2(\Omega)]$ .

**THEOREM 1.** *Let (2.1), (2.2), (A<sub>1</sub>) and (a<sub>1</sub>) be satisfied. If  $f \in W^{1,2}[0, T; L^2(\Omega)]$  and  $f(0) \in D(\phi)$  then there exist functions  $u, w, v \in L^2[0, T; L^2(\Omega)]$  satisfying*

$$(2.5) \quad u \in W^{1,2}[0, T; L^2(\Omega)],$$

$$(2.6) \quad u(t) \in D(A) \quad \text{and} \quad w(t) \in Au(t), \quad \text{a.e. } 0 \leq t \leq T,$$

$$(2.7) \quad v(t) \in B(u(t)), \quad \text{a.e. } 0 \leq t \leq T,$$

$$(2.8) \quad u(t) + \int_0^t a(t-s)(w(s) + v(s)) \, ds = f(t), \quad 0 \leq t \leq T.$$

We remark that (2.7) can be rewritten as  $g(u(t, x)) \leq v(t, x) \leq \bar{g}(u(t, x))$ . It is this inequality which reveals the essential role multiple valued functions play in nonlinear problems. In fact if we require  $v(t, x) = g(u(t, x))$ , (2.4) need not have a solution even in the case of a scalar Volterra equation [9]. If  $p = \infty$  in (A<sub>1</sub>) then Theorem 1 generalizes Theorem 2 of [10].

The existence of solutions to (2.4) can be established under different hypotheses on the kernel  $a(t)$ . With a view toward applications we would like to allow  $a(t) \rightarrow \infty$  as  $t \rightarrow 0^+$ . We shall now assume that the kernel satisfies

$$(a_2) \quad \begin{aligned} &a \in L^2[0, T], a \in AC[\delta, T] \text{ for every } \delta > 0, \text{ and} \\ &a(t) \text{ is positive decreasing on } (0, t_0) \text{ for some } t_0 < T. \end{aligned}$$

**THEOREM 2.** *Let (2.1), (2.2), (A<sub>1</sub>) and (a<sub>2</sub>) be satisfied. If  $f \in W^{1,2}[0, T; L^2(\Omega)]$  and  $f(0) \in D(\phi)$ , then there exist a function  $u \in C[0, T; L^2(\Omega)]$  and functions  $w, v \in L^2[0, T; L^2(\Omega)]$  satisfying (2.6), (2.7) and (2.8).*

Since  $a'$  need not be in  $L^1[0, T]$  we are unable to obtain (2.5).

The condition  $a \in L^2[0, T]$  may be weakened to  $a \in L^1[0, T]$  if we strengthen the hypotheses on  $A$ . We shall now assume that  $A$  satisfies

$$(A_2) \quad \begin{aligned} &\{u: \phi(u) \leq K\} \text{ is precompact in } L^2(\Omega) \text{ and bounded in } L^{2p}(\Omega) \\ &\text{for each } K > 0, \quad \inf_{u \in L^2(\Omega)} \phi(u) > -\infty. \end{aligned}$$

We also modify (a<sub>2</sub>) to

$$(a_3) \quad \begin{aligned} &a \in L^1[0, T], a \in AC[\delta, T] \text{ for every } \delta > 0, \text{ and} \\ &a(t) \text{ is positive decreasing on } (0, T]. \end{aligned}$$

**THEOREM 3.** *Let (2.1), (2.2), (A<sub>2</sub>) and (a<sub>3</sub>) be satisfied. If  $f \in W^{1,2}[0, T; L^2(\Omega)]$  and  $f(0) \in D(\phi)$  then there exist functions  $u, w, v \in L^2[0, T; L^2(\Omega)]$  satisfying (2.6), (2.7) and (2.8) a.e. on  $[0, T]$ .*

It should be observed that under (a<sub>3</sub>) we cannot even claim that  $u(t)$  is continuous. Concerning the strengthening of the last part of (a<sub>2</sub>) see the remark at the end of the proof of Theorem 3.

### 3. Proofs.

*Proof of Theorem 1.* The first step in the proof of Theorem 1 is to write down an appropriate approximating equation for (2.4). We begin by approximating  $g$  by smooth functions. Let  $h(t)$  be a  $C^\infty$  function with support contained in the interval  $(-1, 1)$  satisfying  $h(t) \geq 0$  for all  $t$  and  $\int_{-\infty}^\infty h(t) \, dt = 1$ . Set  $h_n(t) = nh(nt)$  and define

$$(3.1) \quad g_n(t) = \int_{-\infty}^\infty h_n(t-s)\chi_n(s)g(s) \, ds,$$

where  $\chi_n(s) = 1$  if  $-n < s < n$  and is zero elsewhere. Each  $g_n$  is Lipschitz continuous on

$(-\infty, \infty)$  and it is easy to show that (2.1) implies that there are positive constants  $c_5$  and  $c_6$  so that

$$(3.2) \quad |g_n(t)| \leq c_5 + c_6|t|^p, \quad -\infty < t < \infty,$$

for all  $n$ . Next set  $G_n(t) = \int_0^t g_n(s) ds$ , for  $-\infty < t < \infty$ . It is not difficult to show that (2.2) implies the existence of constants  $c_7$  and  $c_8$  so that

$$(3.3) \quad G_n(t) \geq -c_7 - c_8t^2, \quad -\infty < t < \infty,$$

for all  $n$ .

Now define nonlinear operators  $B_n: L^2(\Omega) \rightarrow L^2(\Omega)$  by  $(B_n u)(x) = g_n(u(x))$ , for  $u \in L^2(\Omega)$ . Then each  $B_n$  is Lipschitz continuous on  $L^2(\Omega)$  and by (3.2) there are positive constants  $K_1$  and  $K_2$  so that

$$(3.4) \quad |B_n(u)| \leq K_1 + K_2 \int_{\Omega} |u(x)|^{2p} dx,$$

for all  $n$ . Define  $\psi_n: L^2(\Omega) \rightarrow (-\infty, \infty)$  by  $\psi_n(t) = \int_{\Omega} G_n(u(x)) dx$ . It follows from (3.3) that there are constants  $K_3$  and  $K_4$  independent of  $n$ , so that

$$(3.5) \quad \psi_n(u) \geq -K_3 - K_4|u|^2,$$

and it is not difficult to show that if  $u \in W^{1,1}[0, T; L^2(\Omega)]$  then  $\psi_n(u(t))$  is absolutely continuous on  $[0, T]$  and

$$(3.6) \quad \frac{d}{dt}(\psi_n u(t)) = \langle u'(t), B_n u(t) \rangle, \quad \text{a.e. on } [0, T].$$

The approximating equation we wish to consider is

$$(3.7) \quad u(t) + \int_0^t a(t-s)[Au(s) + B_n u(s)] ds = f(t).$$

To establish the existence of solutions to (3.7) we begin by approximating  $A$ . For each  $\lambda > 0$ , let  $J_\lambda = (I + \lambda A)^{-1}$  and  $A_\lambda = \lambda^{-1}(I - J_\lambda)$ , where  $I$  is the identity operator on  $L^2(\Omega)$ .  $A_\lambda$  is called the Yosida approximation of  $A$ . If we set  $\phi_\lambda(u) = (\lambda/2)|A_\lambda u|^2 + \phi(J_\lambda u)$ , then  $A_\lambda = \partial\phi_\lambda$  and we have

$$(3.8) \quad \phi(J_\lambda u) \leq \phi_\lambda(u) \leq \phi(u).$$

Now fix  $n$  and consider the equation

$$(3.9) \quad u_\lambda(t) + \int_0^t a(t-s)[A_\lambda u_\lambda(s) + B_n u_\lambda(s)] ds = f(t), \quad 0 < t < T.$$

Since  $A_\lambda$  and  $B_n$  are Lipschitz continuous on  $L^2(\Omega)$ , (3.9) has a unique solution  $u_\lambda(t) \in W^{1,2}[0, T; L^2(\Omega)]$ . We now want to obtain bounds on the functions  $u_\lambda, A_\lambda u_\lambda$  and  $B_n u_\lambda$ . First we have

$$(3.10) \quad |u_\lambda(t)| \leq \left( \int_0^t |a(s)|^2 ds \right)^{1/2} \left( \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \right)^{1/2} + M,$$

where  $M = \sup_{0 \leq t \leq T} |f(t)|$ . Differentiate (3.9) with respect to  $t$ , multiply by  $A_\lambda u_\lambda + B_n u_\lambda$

and integrate. With straightforward estimates and (3.6) we obtain

$$\begin{aligned}
 & \phi_\lambda(u_\lambda(t)) - \phi_\lambda(f(0)) + \psi_n(u_\lambda(t)) - \psi_n(f(0)) \\
 & \quad + a(0) \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \\
 (3.11) \quad & \equiv \left( \int_0^t |a'(s)| ds \right) \left( \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \right) \\
 & \quad + \left( \int_0^t |f'(s)|^2 ds \right)^{1/2} \left( \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \right)^{1/2}.
 \end{aligned}$$

Since  $\phi$  is convex and l.s.c., (3.8) implies that there are constants  $\alpha$  and  $\beta$  so that

$$(3.12) \quad \phi_\lambda(u_\lambda(t)) \geq -\alpha|u_\lambda(t)| - \beta,$$

for all  $\lambda > 0$ . Applying (3.5), (3.10) and (3.12) to (3.11) there are constants  $M_1$  and  $M_2$  so that

$$\begin{aligned}
 & a(0) \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \\
 (3.13) \quad & \leq \left[ \int_0^t |a'(s)| ds + 2K_4 \int_0^t |a(s)|^2 ds \right] \left[ \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \right] \\
 & \quad + M_1 \left( \int_0^t |A_\lambda u_\lambda(s) + B_n u_\lambda(s)|^2 ds \right)^{1/2} + M_2.
 \end{aligned}$$

Choose  $T_1$  so small that  $\int_0^{T_1} |a'(s)| ds + 2K_4 \int_0^{T_1} |a(s)|^2 ds \leq \frac{1}{2}a(0)$ . Then (3.10), (3.11) and (3.13) imply that

$$\begin{aligned}
 & \{A_\lambda u_\lambda + B_n u_\lambda\} \text{ is bounded in } L^2[0, T_1; L^2(\Omega)], \\
 (3.14) \quad & \{u_\lambda(t)\} \text{ is bounded in } L^\infty[0, T_1; L^2(\Omega)], \\
 & \{\phi_\lambda(u_\lambda(t))\} \text{ is bounded in } L^\infty[0, T_1], \\
 & \{u'_\lambda(t)\} \text{ is bounded in } L^2[0, T_1; L^2(\Omega)].
 \end{aligned}$$

Since  $B_n$  is Lipschitz continuous on  $L^2(\Omega)$  we also have

$$\begin{aligned}
 (3.15) \quad & \{B_n u_\lambda\} \text{ is bounded in } L^\infty[0, T_1; L^2(\Omega)], \\
 & \{A_\lambda u_\lambda\} \text{ is bounded in } L^2[0, T_1; L^2(\Omega)].
 \end{aligned}$$

We now wish to establish the uniform convergence of  $\{u_\lambda(t)\}$ . By (3.8) and (3.14)  $\{\phi(J_\lambda u_\lambda(t))\}$  is bounded on  $[0, T_1]$  independent of  $\lambda$  and since  $J_\lambda$  is a contraction (3.14) also implies that  $\{J_\lambda u_\lambda(t)\}$  is bounded on  $[0, T_1]$ . By  $(A_1)$  the set  $\{J_\lambda u_\lambda(t) : \lambda > 0, 0 \leq t \leq T_1\}$  is precompact in  $L^2(\Omega)$ . Since each  $u_\lambda \in W^{1,2}[0, T_1; L^2(\Omega)]$  and  $\{u'_\lambda\}$  is bounded in  $L^2[0, T_1; L^2(\Omega)]$  we have that the set  $\{J_\lambda u_\lambda\}$  is equicontinuous on  $[0, T_1]$ . Thus by the Ascoli theorem there is a sequence  $\{\lambda_m\}$ ,  $\lambda_m \rightarrow 0$  as  $m \rightarrow \infty$  and a function  $u_n$  so that  $J_{\lambda_m} u_{\lambda_m}(t) \rightarrow u_n(t)$  as  $m \rightarrow \infty$  uniformly on  $[0, T_1]$ . Hence  $J_{\lambda_m} u_{\lambda_m} \rightarrow u_n$  in  $L^2[0, T_1; L^2(\Omega)]$  and since  $u_{\lambda_m} - J_{\lambda_m} u_{\lambda_m} = \lambda_m A_{\lambda_m} u_{\lambda_m}$ , (3.15) implies that  $u_{\lambda_m} - J_{\lambda_m} u_{\lambda_m} \rightarrow 0$  in  $L^2[0, T_1; L^2(\Omega)]$ . Hence we have

$$(3.16) \quad u_{\lambda_m} \rightarrow u_n \text{ in } L^2[0, T_1; L^2(\Omega)] \text{ as } m \rightarrow \infty.$$

By (3.14), (3.15) and (3.16) there is a subsequence, which we denote again by  $\{\lambda_m\}$  and a function  $w_n \in L^2[0, T_1; L^2(\Omega)]$  satisfying

$$(3.17) \quad \left. \begin{aligned} A_{\lambda_m} u_{\lambda_m} &\rightharpoonup w_n \\ B_n u_{\lambda_m} &\rightarrow B_n u_n \\ u'_{\lambda_m} &\rightharpoonup u'_n \end{aligned} \right\} \text{ in } L^2[0, T_1; L^2(\Omega)],$$

where  $u'_n$  is the distributional derivative of  $u_n$  and  $\rightharpoonup$  denotes weak convergence. It follows from [4, Prop. A.7] that  $u_n \in W^{1,2}[0, T_1; L^2(\Omega)]$  and it is clear that  $u_n, w_n$  and  $B_n u_n$  satisfy

$$(3.18) \quad u_n(t) + \int_0^t a(t-s)[w_n(s) + B_n u_n(s)] ds = f(t), \quad 0 \leq t \leq T_1.$$

Also (3.16) and the first part of (3.17) imply  $u_n(t) \in D(A)$  a.e. and  $w_n(t) \in Au_n(t)$  a.e. on  $[0, T_1]$  by a well-known property of maximal monotone operators. Thus (3.7) has a solution on  $[0, T_1]$ .

We now wish to let  $n \rightarrow \infty$  in (3.18). Since  $f(0) \in D(\phi)$  it follows from  $(A_1)$  and (3.2) that  $|\psi_n(f(0))|$  is bounded and hence we may repeat (3.10)–(3.13) with  $u_\lambda, A_\lambda u_\lambda$  and  $\phi_\lambda$  replaced by  $u_n, w_n$  and  $\phi$ , respectively. We now obtain

$$(3.19) \quad \begin{aligned} \{w_n + B_n u_n\} &\text{ bounded in } L^2[0, T_1; L^2(\Omega)], \\ \{u_n\} &\text{ bounded in } L^\infty[0, T_1; L^2(\Omega)], \\ \{\phi(u_n(t))\} &\text{ bounded in } L^\infty[0, T_1], \\ \{u'_n\} &\text{ bounded in } L^2[0, T_1; L^2(\Omega)]. \end{aligned}$$

By  $(A_1)$ , (3.4) and the first part of (3.19) we have

$$(3.20) \quad \begin{aligned} \{B_n u_n\} &\text{ bounded in } L^\infty[0, T_1; L^2(\Omega)], \\ \{w_n\} &\text{ bounded in } L^2[0, T_1; L^2(\Omega)]. \end{aligned}$$

By  $(A_1)$  and (3.19) the set  $\{u_n(t): 0 \leq t \leq T_1, n = 1, 2, 3, \dots\}$  is precompact in  $L^2(\Omega)$  and  $\{u_n(t)\}$  is equicontinuous on  $[0, T_1]$ . Hence by the Ascoli theorem there is a subsequence, which we again denote by  $\{n\}$ , and a function  $u \in C[0, T_1; L^2(\Omega)]$  satisfying

$$(3.21) \quad u_n(t) \rightarrow u(t) \text{ uniformly on } [0, T_1] \text{ as } n \rightarrow \infty.$$

By (3.19) and (3.20) there is a further subsequence, which we again denote by  $\{n\}$ , and functions  $w, v \in L^2[0, T_1; L^2(\Omega)]$  satisfying

$$(3.22) \quad \begin{aligned} w_n &\rightarrow w, \\ B_n u_n &\rightarrow v, \\ u'_n &\rightharpoonup u', \end{aligned}$$

weakly in  $L^2[0, T_1; L^2(\Omega)]$  as  $n \rightarrow \infty$ , where  $u'$  is the distributional derivative of  $u$ . It is clear that  $u, w$  and  $v$  satisfy (2.8). Again (2.5) follows from [4, Prop. A.7]. By a well known property of maximal monotone operators (3.21) and (3.22) imply (2.6).

To complete the proof all that remains is the verification of (2.7). Following [12] we will establish (2.7) by showing that

$$(3.23) \quad \underline{g}(u(t, x)) \leq v(t, x) \leq \bar{g}(u(t, x)), \quad \text{a.e. on } (0, T_1) \times \Omega.$$

To this end let  $Q$  be the subset of  $R^{n+1}$  given by  $Q = (0, T_1) \times \Omega$ . Then (3.21) implies that

$$(3.24) \quad u_n(t, x) \rightarrow u(t, x) \text{ in } L^2(Q) \text{ as } n \rightarrow \infty.$$

Hence there is a subsequence which we again denote by  $\{u_n\}$  satisfying

$$(3.25) \quad u_n(t, x) \rightarrow u(t, x), \text{ a.e. on } Q.$$

Fix  $0 < \delta < 1$ . By Egoroff's theorem there is a further subsequence which we again denote by  $\{u_n\}$  and a set  $A \subseteq Q$  such that

$$(3.26) \quad m(Q - A) < \delta, \quad u_n(t, x) \rightarrow u(t, x) \text{ uniformly on } A,$$

where  $m$  denotes Lebesgue measure on  $R^{n+1}$ . Let  $D_m = \{(t, x) \in Q : |u(t, x)| \leq m - 1\}$ . Fixing  $m$  we have that for any  $\epsilon > 0$  there is an integer  $N > 2\epsilon^{-1}$  so that if  $n > N$  then  $|u_n(t, x) - u(t, x)| < \epsilon 2^{-1}$ , for all  $(t, x) \in D_m \cap A$ . Thus if  $n > N$  and  $(t, x) \in D_m \cap A$  we have that  $|s - u_n(t, x)| < \epsilon 2^{-1}$  implies  $|s - u(t, x)| < \epsilon$ . Hence we have

$$(3.27) \quad \underline{g}_\epsilon(u(t, x)) \leq g_n(u_n(t, x)) \leq \bar{g}_\epsilon(u(t, x)).$$

Consequently for any  $h \in L^2(Q)$ ,  $h \geq 0$  we have

$$(3.28) \quad \int_{D_m \cap A} \underline{g}_\epsilon(u)h \, dt \, dx \leq \int_{D_m \cap A} g_n(u_n)h \, dt \, dx \leq \int_{D_m \cap A} \bar{g}_\epsilon(u)h \, dt \, dx.$$

By (3.22)  $g_n(u_n(t, x))$  converges weakly in  $L^2(Q)$  to  $v(t, x)$  so (3.28) gives us

$$(3.29) \quad \int_{D_m \cap A} \underline{g}_\epsilon(u)h \, dt \, dx \leq \int_{D_m \cap A} vh \, dt \, dx \leq \int_{D_m \cap A} \bar{g}_\epsilon(u)h \, dt \, dx.$$

Since  $u$  is bounded on  $D_m \cap A$ , Lebesgue's theorem allows us to take the limit as  $\epsilon \rightarrow 0$  to obtain

$$(3.30) \quad \int_{D_m \cap A} \underline{g}(u)h \, dt \, dx \leq \int_{D_m \cap A} vh \, dt \, dx \leq \int_{D_m \cap A} \bar{g}(u)h \, dt \, dx.$$

Since  $h \geq 0$  was arbitrary we conclude that

$$(3.31) \quad \underline{g}(u(t, x)) \leq v(t, x) \leq \bar{g}(u(t, x)), \text{ a.e. on } D_m \cap A.$$

Since we may choose  $\delta$  as small as we like and  $m$  as large as we like, (3.26) follows. The standard translation argument for Volterra equations may now be used to extend the solution to the whole interval  $[0, T]$ , cf. [11, p. 962].

*Proof of Theorem 2.* Let  $a_n(t) = a(t + 1/n)$  and consider the approximating equation

$$(3.32) \quad u_n(t) + \int_0^t a_n(t-s)[Au_n(s) + B_nu_n(s)] \, dx \ni f(t).$$

Since  $a_n$  satisfies (a<sub>1</sub>), Theorem 1 guarantees the existence of functions  $u_n(t)$  and  $w_n(t)$  satisfying (2.5), (2.6) and

$$(3.33) \quad u_n(t) + \int_0^t a_n(t-s)[w_n(s) + B_nu_n(s)] \, ds = f(t), \quad 0 \leq t \leq T.$$

Differentiating (3.33), multiplying by  $w_u + B_n u_n$  and integrating we obtain

$$\begin{aligned}
 & \phi(u_n(t)) - \phi(f(0)) + \psi_n(u_n(t)) - \psi_n(f(0)) \\
 & \quad + a_n(0) \int_0^t |w_n(s) + B_n u_n(s)|^2 ds \\
 (3.34) \quad & \cong \left( \int_0^t |a'_n(s)| ds \right) \left( \int_0^t |w_n(s) + B_n u_n(s)|^2 ds \right) \\
 & \quad + \int_0^t |f'(s)| |w_n(s) + B_n u_n(s)| ds.
 \end{aligned}$$

Since  $a'$  need not be in  $L^1[0, T]$  we apply a technique of Barbu [3]. By  $(a_2)$  and Young's inequality applied to the last term of (3.34) we obtain

$$\begin{aligned}
 & \phi(u_n(t)) - \phi(f(0)) + \psi_n(u_n(t)) - \psi_n(f(0)) + \frac{1}{2} a_n(t) \int_0^t |w_n(s) + B_n u_n(s)|^2 ds \\
 (3.35) \quad & \cong \frac{1}{2 a_n(t)} \int_0^t |f'(s)| ds, \quad 0 \leq t \leq t_0.
 \end{aligned}$$

Now choose  $T_1$  so small that  $0 < T_1 < t_0$  and

$$(3.36) \quad 8K_4 \int_0^t |a_n(s)|^2 ds < a(t_0), \quad \text{for } 0 < t < T_1,$$

where  $K_4$  is given by (3.5). Then by (3.5), (3.10), (3.35) and  $(a_2)$  there are constants  $M_1$  and  $M_2$  independent of  $n$  for large  $n$  so that

$$(3.37) \quad \phi(u_n(t)) + \frac{1}{4} a_n(t) \int_0^t |w_n(s) + B_n u_n(s)|^2 ds \leq M_1 + \frac{1}{2 a_n(t)} M_2, \quad 0 \leq t \leq T_1.$$

Since  $\phi$  is bounded below by an affine function it follows from (3.4), (3.10), (3.37) and  $(A_1)$  that the first three statements of (3.19), (3.20) and the first two statements of (3.22) again hold. The remainder of the proof of existence on  $[0, T_1]$  follows exactly as in the proof of Theorem 1 except for showing that the family  $\{u_n(t)\}$  is equicontinuous. Since  $a_n \rightarrow a$  in  $L^2[0, T]$  the first statement of (3.19) implies immediately that the family  $\{a_n * (w_n + B_n u_n)(t)\}$  is equicontinuous and hence so is  $\{u_n(t)\}$ . Also it should be observed that since  $a_n \rightarrow a$  in  $L^2[0, T_1]$  (3.22) implies that  $a_n * (w_n + B_n u_n) \rightarrow a * (w + v)$  weakly in  $L^2[0, T_1; L^2(\Omega)]$ .

To complete the proof of Theorem 2 we must establish existence on the whole interval  $[0, T]$ . We proceed by induction. Suppose that there exist functions  $u, w, v \in L^2[0, nT_1; L^2(\Omega)]$  satisfying (2.6), (2.7) and (2.8) on  $[0, nT_1]$ . For  $0 < x \leq nT_1$  consider the equation

$$\begin{aligned}
 & y(t) + \int_0^t a(t-s)[Ay(s) + G(y(s))] ds \\
 (3.38) \quad & \ni f(x+t) - \int_0^x a(x+t-s)[w(s) + v(s)] ds, \quad 0 \leq t \leq T_1.
 \end{aligned}$$

If the right-hand side of (3.38) lies in  $W^{1,2}[0, T_1; L^2(\Omega)]$  the first part of the proof of Theorem 2 guarantees that (3.38) has a solution  $y(t)$  on  $[0, T_1]$  ( $T_1$  is restricted only by (3.36)). If we define  $u(t+x) = y(t)$  then  $u(t)$  is a solution of (2.4) on  $[0, x + T_1]$ . Thus set  $h(t, x) = \int_0^x a(x+t-s)[w(s) + v(s)] ds$  for  $0 \leq x \leq nT_1$  and  $0 \leq t \leq T_1$ . It is easy to show



that for each fixed  $x$ ,  $h(t, x)$  is absolutely continuous in  $t$  and  $(\partial/\partial t)h(t, x) = \int_0^x a'(x+t-s)[w(s)+v(s)] ds$ , for  $0 < t \leq T_1$ . Straightforward calculations yield that

$$\int_0^{nT_1} \int_0^{T_1} \left| \frac{\partial}{\partial t} h(t, x) \right|^2 dt dx < \infty,$$

and hence for almost every  $x \in [0, nT_1]$  we have

$$\int_0^x a'(x+t-s)[w(s)+v(s)] ds \in L^2[0, T_1; L^2(\Omega)].$$

Thus (2.4) has a solution on  $[0, x + T_1]$  for almost every  $x \in [0, nT_1]$  and hence has a solution on  $[0, (n + \frac{1}{2})T_1]$ . In this way we eventually get the existence of a solution on  $[0, T]$ .

*Proof of Theorem 3.* The proof of Theorem 3 involves a few minor and one major change in the proof of Theorem 2. First we now have (3.35) for  $0 \leq t \leq T$ . Since  $c_4 = 0$  in (2.2),  $\psi_n(u_n(t))$  is bounded from below and hence we obtain (3.37) for  $0 \leq t \leq T$ . Since  $a$  is only assumed to be in  $L^1[0, T]$  we do not have (3.10), but  $(A_2)$  now implies that the first three statements of (3.19), (3.20) and the first two statements of (3.22) hold. The major change in the proof of Theorem 2 enters in establishing

$$(3.39) \quad \{u_n\} \text{ is precompact in } L^2[0, T; L^2(\Omega)],$$

which will then imply the existence of a function  $u \in L^2[0, T; L^2(\Omega)]$  and a subsequence  $\{n_k\}$  so that (3.24) holds. Under  $(a_3)$  the family  $\{u_n(t)\}$  need not be equicontinuous, but it is not difficult to show that (3.19), (3.33) and the fact that  $a_n \rightarrow a$  in  $L^1[0, T]$  imply

$$(3.40) \quad \int_0^T |u_n(t+\delta) - u_n(t)|^2 dt \rightarrow 0 \quad \text{as } \delta \rightarrow 0 \text{ uniformly in } n.$$

For each  $\alpha > 0$  define  $M_\alpha u_n(t) = (1/2\alpha) \int_\alpha^t u_n(t+s) ds$ . Then (3.40) implies that  $M_\alpha u_n \rightarrow u_n$  in  $L^2(0, T; L^2(\Omega))$  as  $\alpha \rightarrow 0$  uniformly in  $n$  and that the family  $\{M_\alpha u_n(t) : n = 1, 2, 3, \dots\}$  is equicontinuous on  $[0, T]$  for each fixed  $\alpha$ . Also the set  $\{M_\alpha u_n(t) : 0 \leq t \leq T, n = 1, 2, 3, \dots\}$  is contained in the closed convex hull of  $\{u_n(t) : 0 \leq t \leq T, n = 1, 2, 3, \dots\}$  for each fixed  $\alpha$  and hence is precompact in  $L^2(\Omega)$ . Thus for each  $\alpha > 0$  the set  $\{M_\alpha u_n(t) : 0 \leq t \leq T, n = 1, 2, 3, \dots\}$  is precompact in  $C[0, T; L^2(\Omega)]$  by the Ascoli theorem. Now it is not difficult to show that  $\{u_n : n = 1, 2, 3, \dots\}$  is totally bounded in  $L^2[0, T; L^2(\Omega)]$  and hence precompact, cf. [13, p. 86]. This completes the proof of Theorem 3.

*Remark.* The only reason for assuming that  $a(t)$  is positive decreasing on the whole interval  $[0, T]$  arises from the fact that  $a \in L^1[0, T]$  is not sufficient to guarantee that  $\int_0^x a'(x+t-s)[w(s)+v(s)] ds$  lies in  $L^2[0, T; L^2(\Omega)]$  for any  $T > 0$ . In particular if  $a(t) = O(t^{-1/2})$  as  $t \rightarrow 0^+$  and  $w(s)+v(s)$  is bounded away from zero as  $s \rightarrow 0$  then  $\int_0^x a'(x+t-s)[w(s)+v(s)] ds$  is not an element of  $L^2[0, T; L^2(\Omega)]$  for any  $T > 0$  and any  $x > 0$ .

**4. Examples.**

*Example 1.* Let  $\Omega$  be a bounded open domain in  $R^n$  with smooth boundary  $\Gamma$ , and consider the equation

$$(4.1) \quad \begin{aligned} u(t, x) - \int_0^t a(t-s)(\Delta u(s, x)) ds + \int_0^t a(t-s)g(u(s, x)) ds &= f(t, x), \\ (t, x) \in (0, T) \times \Omega, \\ -\frac{\partial u}{\partial n} \in \beta(u), \quad \text{a.e. on } (0, T) \times \Gamma, \end{aligned}$$

where  $\beta$  is maximal monotone on  $R$ ,  $0 \in \beta(0)$  and  $\beta = \partial j$  where  $j: R \rightarrow [0, \infty)$  is convex and lower semicontinuous. It is well known [5] that if we set

$$\begin{aligned} \phi(u) &= \frac{1}{2} \int_{\Omega} |\text{grad } u|^2 dx + \int_{\Gamma} j(u) dx, \\ D(\phi) &= \{u \in H^1(\Omega): j(u) \in L^1(\Gamma)\}, \end{aligned} \tag{4.2}$$

then  $\partial\phi(u) = -\Delta u$  and  $D(\partial\phi) = \{u \in H^2(\Omega): -(\partial u/\partial n) \in \beta(u) \text{ a.e. on } \Gamma\}$ . By (4.2) and  $j(u) \geq 0$  we have that  $\{u \in L^2(\Omega): \phi(u) + |u| \leq K\}$  is bounded in  $H^1(\Omega)$  and hence by the Sobolev imbedding theorems is precompact in  $L^2(\Omega)$  and bounded in  $L^{2p}(\Omega)$  for

$$\begin{aligned} 1 \leq p \leq \infty & \text{ if } n = 1, \\ 1 \leq p < \infty & \text{ if } n = 2, \\ 1 \leq p \leq \frac{n}{n-2} & \text{ if } n \geq 3. \end{aligned} \tag{4.3}$$

Thus  $(A_1)$  is satisfied if  $p$  is restricted by (4.3).

*Example 2.* Consider the equation

$$\begin{aligned} u(t, x) - \int_0^t (t-s)^{-1/2} (\sigma(u_x(s, x)))_x ds + \int_0^t (t-s)^{-1/2} g(u(s, x)) ds &= f(t, x), \\ 0 \leq t \leq T, \quad 0 < x < 1, \\ u(t, 0) = u(t, 1) &= 0, \end{aligned} \tag{4.4}$$

where  $\sigma \in C^1(-\infty, \infty)$ ,  $0 \leq \sigma' \leq M < \infty$  and  $\Sigma(r) = \int_0^r \sigma(s) ds \geq c(r^2 - 1)$ , for some  $c > 0$ . If we define  $\phi: L^2(0, 1) \rightarrow (-\infty, \infty]$  by

$$\phi(u) = \begin{cases} \int_0^1 \Sigma\left(\frac{du}{dx}\right) dx & \text{if } u \in H_0^1(0, 1), \\ +\infty & \text{otherwise,} \end{cases} \tag{4.5}$$

then  $\phi$  is a well-defined, proper, convex, lower semicontinuous function and

$$\partial\phi(u) = -\frac{d}{dx} \left( \sigma \left( \frac{du}{dx} \right) \right) \quad \text{with} \quad D(\partial\phi) = \left\{ u \in H_0^1(0, 1): \frac{d}{dx} \left( \sigma \left( \frac{du}{dx} \right) \right) \in L^2(0, 1) \right\}.$$

By the growth restriction on  $\Sigma$  it is clear that  $\{u: \phi(u) \leq K\}$  is bounded in  $H_0^1(0, 1)$  and hence bounded in  $L^{2p}(0, 1)$ , for all  $p \geq 1$ , and precompact in  $L^2(0, 1)$ .

*Example 3.* For our last example consider

$$\begin{aligned} u(t, x) - \int_0^t (t-s)^{-1/2} (\Delta u(s, x)) ds + \int_0^t (t-s)^{-1/2} g(u(s, x)) ds &= f(t, x), \\ 0 \leq t \leq T, \quad (t, x) \in (0, T) \times \Omega, \\ u(t, x) &= 0, \quad (t, x) \in (0, T) \times \Gamma. \end{aligned} \tag{4.6}$$

As in Example 1 we set  $\phi(u) = \frac{1}{2} \int_{\Omega} |\text{grad } u|^2 dx$ . It is immediate that  $(a_3)$  and  $(A_2)$  are satisfied if  $p$  is restricted by (4.3).

**Acknowledgment.** I wish to thank the referee for several suggestions which significantly increased the generality of the results presented here.

## REFERENCES

- [1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.
- [2] ———, *Nonlinear Volterra equations in a Hilbert space*, this Journal, 6 (1975), pp. 728–741.
- [3] ———, *On a nonlinear Volterra integral equation on a Hilbert space*, this Journal, 8 (1977), pp. 346–355.
- [4] H. BREZIS, *Opérateurs maximaux monotones et semi-groupes de contraction dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.
- [5] M. G. CRANDALL, S.-O. LONDEN AND J. A. NOHEL, *An abstract nonlinear Volterra integrodifferential equation*, J. Math. Anal. Appl., 64 (1978), pp. 701–735.
- [6] M. G. CRANDALL AND J. A. NOHEL, *An abstract functional differential equation and a related nonlinear Volterra equation*, Israel J. of Math., 29 (1978), pp. 313–328.
- [7] G. GRIPENBERG, *An existence result for a nonlinear Volterra integral equation in Hilbert space*, this Journal, 9 (1978), pp. 793–805.
- [8] T. KIFFE AND M. STECHER, *A characterization of the range of a nonlinear Volterra integral operator*, Nonlinear Equations in Abstract Spaces, V. Lakshmikantham, ed., Academic Press, New York, 1978.
- [9] T. KIFFE, *A discontinuous Volterra integral equation*, J. Integral Equations, 1 (1979), pp. 193–200.
- [10] ———, *A Volterra integral equation and multiple valued functions*, J. Integral Equations, to appear.
- [11] S.-O. LONDEN, *On an integral equation in a Hilbert space*, this Journal, 8 (1977), pp. 950–970.
- [12] J. RAUCH, *Discontinuous semilinear differential equations and multiple valued maps*, Proc. Amer. Math. Soc., 64 (1977), pp. 277–282.
- [13] A. VOIGT AND J. WLOKA, *Hilberträume und elliptische Differential-operatoren*, Bibliographisches Institut, Zurich, 1975.

## CONDITIONAL POSITIVITY OF QUADRATIC FORMS IN HILBERT SPACE\*

D. H. MARTIN†

**Abstract.** If  $X$  and  $Y$  are real Hilbert spaces,  $A : X \rightarrow Y$  is a bounded linear operator, and  $\Gamma \subseteq Y$  is a closed convex cone, an immediate sufficient condition for a quadratic form  $Q$  on  $X$  to be positive subject to the constraint  $Ax \in \Gamma$ , is that  $Q$  be decomposable as a sum  $Q(x) = C(Ax) + S(x)$ , where  $C$  is a quadratic form on  $Y$  which is positive on  $\Gamma$ , and  $S$  is positive definite on  $X$ . The necessity of such a decomposition is not obvious, but is established here for a class of quadratic forms which commonly occur in variational problems—the Legendre forms. The proof furnishes formulas for  $C$  and  $S$  which are explicit apart from the occurrence of an unknown scalar. The usefulness of the result is illustrated by the determination of the focal (conjugate) time of a linear-quadratic control problem with inequality constraints on the final state.

**1. Introduction and statement of results.** Let  $X$  and  $Y$  be two real Hilbert spaces, with a bounded linear operator

$$A : X \rightarrow Y,$$

and a closed convex cone  $\Gamma \subseteq Y$  being given. We shall say that a continuous quadratic form  $Q$  on  $X$  is *conditionally positive* (relative to  $A$  and  $\Gamma$ ) if

$$Q(x) > 0 \quad \text{whenever } Ax \in \Gamma \text{ and } x \neq 0,$$

and, as usual, that  $Q$  is *positive definite* or *positive semidefinite* if  $Q(x) > 0$  for all  $x \neq 0$ , or  $Q(x) \geq 0$  for all  $x$ , respectively. If there is  $\gamma > 0$  such that the inequality

$$Q(x) \geq \gamma \|x\|^2$$

may replace  $Q(x) > 0$ , we shall speak of *strong* conditional positivity or *strong* positive definiteness, respectively. The problem of testing for conditional positivity arises, for example, in second-order optimality conditions for constrained minimization problems in Hilbert space.

Regarding quadratic forms  $C$  on the space  $Y$ , we shall say that  $C$  is

- (a)  $\Gamma$ -*copositive* if  $C(y) \geq 0$  for all  $y \in \Gamma$ ;
- (b) *strictly*  $\Gamma$ -*copositive* if  $C(y) > 0$  for all  $y \in \Gamma \setminus \{0\}$ ;
- (c) *strongly*  $\Gamma$ -*copositive* if there exists  $\gamma > 0$  such that

$$C(y) \geq \gamma \|y\|^2, \quad \text{for all } y \in \Gamma.$$

The following sort of immediate sufficient condition for conditional definiteness was first noted by D. H. Jacobson in [6], dealing with the finite-dimensional case

$$(1.1) \quad X = \mathbb{R}^n, \quad \Gamma = \mathbb{R}_+^m \subseteq \mathbb{R}^m = Y.$$

Let us say that  $Q$  admits a *strict decomposition* (relative to  $A$  and  $\Gamma$ ) if it can be written in the form

$$(1.2) \quad Q(x) = C(Ax) + S(x),$$

where  $C$  is a *strictly*  $\Gamma$ -copositive form on  $Y$  and  $S$  is positive semidefinite on  $X$  with the additional property that

$$(1.3) \quad S(x) > 0 \quad \text{whenever } Ax = 0, x \neq 0.$$

\* Received by the editors August 24, 1979, and in revised form April 24, 1980.

† National Research Institute for Mathematical Sciences, CSIR, Pretoria, South Africa. This work was partially supported by a grant from Control Data.

If  $C$  is strongly  $\Gamma$ -copositive and  $S$  is strongly positive definite, we shall say that  $Q$  admits a strong decomposition.

**THEOREM 1.1.** *A quadratic form which admits a strict decomposition is conditionally positive.*

This paper concerns the deeper converse question. For the finite-dimensional case (1.1), Jacobson and the writer [8] have recently proved the strong converse result that a conditionally positive form must admit a strong decomposition. Further interest is lent to this fact by the failure, even in finite dimensions, of the corresponding conjecture regarding conditional nonnegativity. A counterexample, with

$$X = Y = \mathbb{R}^2, \quad \Gamma = \mathbb{R}_+^2$$

is afforded by the form  $Q(x_1, x_2) = x_1x_2$ , with  $A: \mathbb{R}^2 \rightarrow \mathbb{R}$  given by the matrix

$$\begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}.$$

Clearly  $Ax \in \Gamma$  iff  $x$  is of the form  $x' = (0, x_2)$ , and then  $Q(x) = 0$ , so that  $Q$  is conditionally nonnegative. However, as is shown in [8],  $Q$  admits no decomposition of the form (1.2) with  $S$  positive semidefinite.

In this paper the methods and main result of [8] are extended to the general Hilbert space setting introduced above, consideration being restricted, however, to a class of quadratic forms  $Q$  on  $X$  known as Legendre forms. This term was used in the detailed study [4] by M. R. Hestenes, and is motivated by the occurrence of such forms as second variation functionals in the calculus of variations and optimal control theory, when the so-called strengthened Legendre condition (see, for example, [3, p. 116]) holds.

**DEFINITION 1.1.** A Legendre form  $Q$  on a Hilbert space is a continuous quadratic functional which can be written as a sum

$$Q = Q_+ + Q_0$$

of a strongly positive definite form  $Q_+$  and a completely continuous form  $Q_0$  (i.e., a form continuous in the weak topology).

Properties of Legendre forms that we shall require are given in the next section.

The results of this paper are stated in the following three theorems.

**THEOREM 1.2.** *A conditionally positive Legendre form must admit a strong decomposition.*

**THEOREM 1.3.** *A necessary and sufficient condition for a Legendre form*

$$Q(x) = (x, Lx)_X$$

on  $X$  to be conditionally positive is that for some  $\nu > 0$

- (a) the form  $Q(x) + \nu \|Ax\|_Y^2$  is strongly positive definite on  $X$ , and
- (b) the form

$$(1.4) \quad C_\nu(y) := \nu \|y\|_Y^2 - \nu^2 (y, A(L + \nu A^*A)^{-1} A^*y)_Y$$

on  $Y$  is strictly  $\Gamma$ -copositive. For such  $\nu$ , this form  $C_\nu$  on  $Y$ , together with the form

$$(1.5) \quad S_\nu(x) := (Lx, (L + \nu A^*A)^{-1} Lx)_X$$

on  $X$  provide a strict decomposition of  $Q$ . Furthermore, both  $C_\nu$  and  $S_\nu$  are Legendre forms.

Note that the self-adjoint operator  $L + \nu A^*A$  has a bounded inverse because of (a).

The proof of Theorem 1.3 depends upon our final result, which is a theorem in the spirit of a well-known theorem of P. Finsler [2] on quadratic forms on  $\mathbb{R}^n$ . For further related theorems and extensions see [6, p. 90], [8], and also [1, p. 75], [5, § 2.6] and [7].

**THEOREM 1.4.** *Let  $Q_1$  and  $Q_2$  be quadratic forms on a Hilbert space  $H$ , such that  $Q_2$  is positive semidefinite while  $Q_1 + Q_2$  is a Legendre form, and let  $\mathcal{C} \subseteq H$  be a closed convex cone. Then*

$$(1.6) \quad Q_1(h) > 0 \quad \text{whenever } h \in \mathcal{C} \setminus \{0\} \text{ and } Q_2(h) = 0$$

*iff there exists  $\nu > 0$  such that*

$$(1.7) \quad Q_1(h) + \nu Q_2(h) > 0 \quad \text{whenever } h \in \mathcal{C} \setminus \{0\}.$$

These three theorems are proved in reverse order in §§ 3, 4, and 5, followed in § 6 by a remark concerning the evaluation of the forms  $C_\nu$ . Section 7 presents an application of Theorem 1.3 to the determination of focal times for a quadratic functional on a linear control system, subject to inequality constraints on the final states.

**2. Properties of Legendre forms.** The basic reference on Legendre forms is the study [4] by Hestenes. Legendre forms have many nice properties, some of which we list here. Let  $Q$  be a Legendre form on a Hilbert space  $H$ .

P1. There exists  $r > 0$  such that for all positive  $\delta < r$ , the form

$$Q_\delta(h) = Q(h) - \delta \|h\|^2$$

is also a Legendre form on  $H$ .

P2. There exists  $r > 0$  such that whenever  $h_n \in H$  is a sequence such that<sup>1</sup>

$$h_n \rightarrow 0 \text{ and } \|h_n\| = 1,$$

then

$$\underline{\lim}_n Q(h_n) \geq r.$$

P3.  $Q$  is weakly lower semicontinuous (henceforth abbreviated to wls) on  $H$ ; i.e., whenever  $h_n \rightarrow h_0$  then

$$Q(h_0) \leq \underline{\lim}_n Q(h_n).$$

P4. If  $Q(h) > 0$  for all nonzero  $h$  in some closed convex cone  $\mathcal{C} \subseteq H$ , there exists  $\gamma > 0$  such that

$$Q(h) \geq \gamma \|h\|^2, \quad \text{for all } h \in \mathcal{C}.$$

P5. Whenever  $h_n \rightarrow h_0$  and  $Q(h_n) \rightarrow Q(h_0)$ , then  $h_n \rightarrow h_0$ .

P6. There is a closed subspace  $H_1 \subseteq H$  of finite codimension such that  $Q$  is strongly positive on  $H_1$ .

In the study by Hestenes which has been cited, he defined as Legendre forms those forms having the properties P3 and P5, and he showed [4, Theorem 11.6] that this is equivalent to the definition adopted here (Definition 1.1), and also [4, Theorem 11.4] to property P6. Property P4 is an easy extension to the case of closed convex cones of [4, Theorem 11.1] which proves P4 for the case  $\mathcal{C} = H$ .

Properties P1 and P2 are immediate consequences of the definition and the fact that the norm is wls. What is interesting, and also useful in the sequel, is that a property apparently weaker at first sight than P2, namely

P2'. Whenever  $h_n \rightarrow 0$  with  $\|h_n\| = 1$ , then  $\underline{\lim}_n Q(h_n) > 0$ ,

is sufficient to characterize Legendre forms. Routine convergence arguments can be given to show this, but it is more interesting to see how P2' implies P6.

---

<sup>1</sup> The symbol  $\rightarrow$  denotes weak convergence, while  $\rightarrow$  denotes strong convergence.

Given a quadratic form  $Q$  on  $H$ , consider the following inductive procedure. Starting at step 1 with an arbitrarily chosen unit vector  $h_1$ , the  $N$ th step supposes that we have an orthonormal set  $\{h_1, h_2, \dots, h_N\}$  such that

$$Q(h_k) < \frac{1}{(k-1)}, \quad k = 2, 3, \dots, N,$$

and the sequence is continued by selecting, if possible, a unit vector  $h_{N+1}$ , orthogonal to each of  $h_1, \dots, h_N$ , such that also

$$Q(h_{N+1}) < \frac{1}{N}.$$

If this is always possible, the procedure never terminates, and produces an orthonormal sequence for which

$$\lim_n Q(h_n) \leq 0.$$

Since every orthonormal sequence converges weakly to zero, this means that  $Q$  does not satisfy P2'. The other alternative is that the procedure fails at the  $N$ th step (for some  $N$ ), which means that for all  $h \in H$  orthogonal to  $\{h_1, \dots, h_N\}$ , we have

$$Q(h) \geq \frac{1}{N} \|h\|^2,$$

showing that  $Q$  satisfies P6. Thus P2' implies P6, as asserted.

Finally we remark that the typical quadratic functional

$$J(x_0, u(\cdot)) := \int_0^T [u'(t)R(t)u(t) + x'(t)Q(t)x(t)] dt + x'(T)Hx(T),$$

where

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad 0 \leq t \leq T, \quad x(0) = x_0,$$

which occurs in optimal control theory, is a Legendre functional iff there exists  $r > 0$  such that for almost all  $t \in [0, T]$ , and all  $u \in \mathbb{R}^m$ ,

$$u'R(t)u \geq r|u|^2.$$

This is known as the strengthened Legendre condition in optimal control theory. It is, of course, not assumed that the matrices  $Q(\cdot)$  and  $H$  are positive semidefinite.

**3. Proof of Theorem 1.4.** Let  $Q_1, Q_2, \mathcal{C} \subseteq H$  be as in the first sentence of Theorem 1.4. It is immediately obvious that if (1.7) holds for some constant  $\nu$ , then (1.6) holds. For the converse, suppose (1.7) does not hold for any constant  $\nu > 0$ . Then for each  $n = 1, 2, \dots$  we can find  $h_n \in \mathcal{C}$ , with  $\|h_n\| = 1$  such that

$$(3.1) \quad Q_1(h_n) + nQ_2(h_n) \leq 0, \quad n = 1, 2, \dots$$

Being a bounded sequence,  $\{h_n\}$  has a weakly convergent subsequence

$$h_{n_k} \xrightarrow[k]{\infty} h_0,$$

where  $h_0$  is necessarily in the closed convex cone  $\mathcal{C}$ . From (3.1) it follows that

$$Q_2(h_{n_k}) \leq -\frac{Q_1(h_{n_k})}{n_k} \leq \frac{\|Q_1\|}{n_k} \xrightarrow[k]{\infty} 0,$$

so that, since  $Q_2$  is positive semidefinite, and hence wpsc, we have

$$(3.2) \quad Q_2(h_0) = 0.$$

However it also follows from (3.1) that  $Q_1(h_n) + Q_2(h_n) \leq 0$ , and hence by the property P3 (wpsc) of the Legendre form  $Q_1 + Q_2$  that

$$(3.3) \quad Q_1(h_0) = Q_1(h_0) + Q_2(h_0) \leq \varliminf_k [Q_1(h_{n_k}) + Q_2(h_{n_k})] \leq 0.$$

Comparing (3.3) and property P2 (or even P2'), it is clear that  $h_0 \neq 0$ . But then (3.3), (3.2) and the fact that  $h_0 \in \mathcal{C}$  contradict condition (1.6), which completes the proof.

**4. Proof of Theorem 1.3.** Let  $Q(x) = (x, Lx)_X$  be a quadratic form on  $X$ . Then, with the forms  $C_\nu$  on  $Y$  and  $S_\nu$  on  $X$  defined by (1.4) and (1.5) for any  $\nu$  for which  $L + \nu A^*A$  is invertible, we have

$$\begin{aligned} C_\nu(Ax) &= \nu \|Ax\|_Y^2 - \nu^2 (Ax, A(L + \nu A^*A)^{-1} A^*Ax)_Y \\ &= (x, [\nu A^*A - (\nu A^*A)(L + \nu A^*A)^{-1}(\nu A^*A)]x)_X. \end{aligned}$$

Hence using the general identity

$$L - L(L + M)^{-1}L = M - M(L + M)^{-1}M,$$

it follows that

$$C_\nu(Ax) = (x, [L - L(L + \nu A^*A)^{-1}L]x)_X,$$

or

$$(4.1) \quad Q(x) = C_\nu(Ax) + S_\nu(x).$$

The sufficiency part of Theorem 1.3 follows easily from this, and is actually independent of the assumption that  $Q$  be a Legendre form. For if  $\nu$  is such that conditions (a) and (b) hold, then first, as already noted, the operator  $L + \nu A^*A$  is strongly positive definite, and hence has a bounded inverse, which is also strongly positive definite. It follows that the form  $S_\nu(x)$ , given by (1.5), is positive semidefinite, with

$$S_\nu(x) = 0, \quad \text{iff } Lx = 0.$$

But, using (a) again, it follows that if  $Ax = 0$  but  $x \neq 0$ , then  $Lx \neq 0$ , since in these circumstances

$$0 < Q(x) + \nu \|Ax\|_Y^2 = Q(x).$$

Thus  $S_\nu$  also satisfies (1.3), which, together with (b) and (4.1), shows that  $Q$  admits a strict decomposition, and hence, by Theorem 1.1, is conditionally positive. This proves not only the sufficiency, but also that if (a), (b) hold, then the forms (1.4) and (1.5) provide a strict decomposition of  $Q$ .

For the converse, let  $Q$  be a conditionally positive Legendre form on  $X$ . Let  $H = X \times Y$ , and let  $\mathcal{C} \subseteq H$  be the closed convex cone

$$\mathcal{C} = \{[x, y] \in H \mid y \in \Gamma\}.$$

We shall apply Theorem 1.4, with the role of  $Q_1$  taken by the form  $[x, y] \mapsto Q(x)$ , and that of  $Q_2$  by the form  $\|y - Ax\|_Y^2$ . To see that

$$(4.2) \quad Q(x) + \|y - Ax\|_Y^2$$



is a Legendre form on  $H$ , suppose  $[x_n, y_n] \rightarrow [0, 0]$  with

$$(4.3) \quad \|[x_n, y_n]\|_H^2 = \|x_n\|_X^2 + \|y_n\|_Y^2 = 1.$$

Since  $Q$  is wslc on  $X$  and  $x_n \rightarrow 0$ , we have

$$(4.4) \quad \underline{\lim}_n [Q(x_n) + \|y_n - Ax_n\|_Y^2] \geq 0,$$

with equality only if there is a subsequence  $[x_{n_k}, y_{n_k}]$  for which

$$Q(x_{n_k}) \rightarrow 0 \quad \text{and} \quad \|y_{n_k} - Ax_{n_k}\|_Y \rightarrow 0.$$

By P5, the first of these conditions would imply that  $x_{n_k} \rightarrow 0$ , and then the second would imply that  $y_{n_k} \rightarrow 0$ , contrary to (4.3). Thus strict inequality holds in (4.4), showing that the form (4.2) has property P2', and is thus a Legendre form. Finally, we note that the conditional positivity of  $Q$  may be stated as

$$Q(x) > 0 \quad \text{whenever} \quad [x, y] \in \mathcal{C} \setminus \{0\} \quad \text{and} \quad \|y - Ax\|_Y^2 = 0.$$

Thus Theorem 1.4 applies, and so there exists  $\nu > 0$  such that

$$(4.5) \quad Q(x) + \nu \|y - Ax\|_Y^2 > 0 \quad \text{whenever} \quad y \in \Gamma \quad \text{and} \quad [x, y] \neq [0, 0].$$

Henceforth in this proof  $\nu$  is held fixed at this value. Obviously any greater value would also ensure (4.5). For  $y = 0$ , (4.5) reduces to

$$Q(x) + \nu \|Ax\|_Y^2 > 0 \quad \text{whenever} \quad x \neq 0.$$

Since this form is manifestly also a Legendre form on  $X$ , it follows from P4 that it must be strongly positive definite—i.e., condition (a) holds. Furthermore, by standard theorems, the self-adjoint operator  $L + \nu A^*A$  associated with this form has a bounded, strongly positive definite inverse.

To prove that (b) holds (for this same value of  $\nu$ ), we introduce a bounded linear operator  $B_\nu : Y \rightarrow X$  by

$$(4.6) \quad B_\nu = \nu(L + \nu A^*A)^{-1}A^*.$$

Direct calculation verifies that for any  $y \in Y$ , if we substitute  $x = B_\nu y$  into

$$(4.7) \quad \tilde{Q}_\nu[x, y] := Q(x) + \nu \|y - Ax\|_Y^2,$$

we obtain precisely  $C_\nu(y)$ , as given by (1.4). Hence, by (4.5), we have

$$C_\nu(y) > 0 \quad \text{whenever} \quad y \in \Gamma \quad \text{and} \quad y \neq 0,$$

showing that  $C_\nu$  is strictly  $\Gamma$ -cpositive, as required.

It remains to show that the forms  $C_\nu$  and  $S_\nu$  are Legendre forms. The argument used above to show that (4.2) is a Legendre form also shows that for any  $\nu > 0$ , the form  $\tilde{Q}_\nu$  given by (4.7) is a Legendre form on  $H$ . Suppose now that  $y_n \rightarrow 0$  with  $\|y_n\| = 1$ . Then also  $B_\nu y_n \rightarrow 0$ , and so by P3 we have

$$\underline{\lim}_n C_\nu(y_n) = \underline{\lim}_n \tilde{Q}_\nu[B_\nu y_n, y_n] \geq 0,$$

with equality only if for some subsequence  $y'_n$ , we have

$$\tilde{Q}_\nu[B_\nu y'_n, y'_n] \rightarrow 0.$$

But then P5 would imply that  $y'_n \rightarrow 0$ , which is false. Thus the form  $C_\nu$  has property P2', and is a Legendre form on  $Y$ .

Finally, suppose  $x_n \rightarrow 0$  with  $\|x_n\|_X^2 = 1$ . Since  $(L + \nu A^*A)^{-1}$  is strongly positive definite, there exists  $\gamma > 0$  such that

$$S_\nu(x) \geq \gamma \|Lx\|_X^2.$$

Hence

$$(4.8) \quad \liminf_n S_\nu(x_n) \geq 0,$$

with equality only if for some subsequence  $x'_n$  we have

$$Lx'_n \rightarrow 0.$$

Since  $Q$  is a Legendre form, property P5 would imply that  $x'_n \rightarrow 0$ , which is false. Thus strict inequality holds in (4.8), proving that  $S_\nu$  is a Legendre form. This completes the proof of Theorem 1.3.

**5. Proof of Theorem 1.2.** This is an easy corollary of Theorem 1.3 and properties P1 and P4 of the Legendre form  $Q$ . These properties imply that if  $Q$  is a conditionally positive Legendre form, then there exists  $r > 0$  such that

$$Q_r(x) := Q(x) - r\|x\|_X^2$$

is also a conditionally positive Legendre form. Applying Theorem 1.3, we conclude that  $Q_r$  admits a strict decomposition. Hence there is a strictly  $\Gamma$ -copositive Legendre form  $C$  on  $Y$  and a positive semidefinite form  $S$  on  $X$  such that

$$Q(x) = C(Ax) + S(x) + r\|x\|_X^2.$$

By P4, the form  $C$  is actually *strongly*  $\Gamma$ -copositive, while the form  $S(x) + r\|x\|_X^2$  is manifestly strongly positive definite. Thus  $Q$  admits a strong decomposition, as claimed.

**6. Evaluation of the form  $C(y)$ .** The operator  $B_\nu$ , given by (4.6), was important in the proof of Theorem 1.3 because of the relation

$$(6.1) \quad C_\nu(y) = \tilde{Q}_\nu[B_\nu y, y].$$

However, one can say more: if condition (a) holds, i.e., if

$$Q(x) + \nu\|Ax\|^2 = \tilde{Q}_\nu[x, 0]$$

is strongly positive definite, then, as is easily verified,

$$(6.2) \quad C_\nu(y) = \min \{ \tilde{Q}_\nu[x, y] \mid x \in X \},$$

with the unique minimizer given by

$$x = B_\nu y.$$

Thus, for given  $\nu$ ,  $C_\nu(y)$  may be evaluated as the *unconstrained minimum* over  $x \in X$  of  $\tilde{Q}_\nu[x, y]$ .

**7. An application to constrained focal times.** For controllers  $u(t)$ ,  $t_0 \leq t \leq T$ , in  $X := L^2([t_0, T], \mathbb{R}^m)$ , let

$$Q(u(\cdot)) := \int_{t_0}^T (u'(t)Ru(t) + x'(t)Sx(t)) dt + x'(T)Hx(T),$$

be a quadratic functional defined on the linear control system in  $\mathbb{R}^n$ :

$$(7.1) \quad \dot{x} = Ax + Bu, \quad x(t_0) = 0,$$

where  $R, S, H, A, B$  are given matrices, with  $R, S$  and  $H$  symmetric. We suppose that  $R$  is *positive definite*—as remarked in § 2, this is necessary and sufficient for  $Q$  to be a Legendre form on  $X$ . Suppose however that  $S$  and  $H$  are not both positive semi-definite. Given a further  $r \times n$  matrix  $D$ , we could then ask: what is the infimum  $t_0^*$  of all initial times  $t_0$  such that

$$(7.2) \quad Q(u(\cdot)) > 0 \quad \text{whenever } Dx(T) \geq 0, u(\cdot) \neq 0?$$

This instant  $t_0^*$  is a “conditional focal time” for the functional  $Q$ . The final time  $T$  is, of course, held fixed.

For each  $t_0$ , (7.2) is a question of conditional positivity, where

$$X = L^2([t_0, T], \mathbb{R}^n), \quad \Gamma = \mathbb{R}_+^r \subseteq \mathbb{R}^r = Y,$$

and the operator  $A : X \rightarrow Y$  assigns to each controller  $u(\cdot) \in X$  the vector  $Dx(T) \in \mathbb{R}^r$ .

Before dealing with the form  $C_\nu$ , we consider the testing of condition (a) in Theorem 1.3—i.e., the question as to whether, for given  $\nu$  and  $t_0$ ,

$$(7.3) \quad Q(u(\cdot)) + \nu \|Au(\cdot)\|_Y^2 = \int_{t_0}^T (u'(t)Ru(t) + x'(t)Sx(t)) dt + x'(T)(H + \nu D'D)x(T)$$

is strongly positive definite on  $X$ . It is well known (see, for example, [6]), that this is so iff the Riccati problem

$$(7.4) \quad -\dot{P} = PA + A'P - PBR^{-1}B'P + S, \quad P(T) = H + \nu D'D,$$

has a solution which exists over the full interval  $[t_0, T]$ . Thus, for given  $\nu > 0$ , one may, by determining the “blow up” time of the Riccati problem (7.4), determine the infimum of times  $t_0$  for which that  $\nu$  satisfies condition (a).

Using the strategy (6.2) it turns out that a similar statement can be made regarding condition (b) of Theorem 1.3. The form  $\tilde{Q}_\nu$  on  $X \times Y$  is given by

$$\tilde{Q}_\nu[u(\cdot), y] := \int_{t_0}^T (u'(t)Ru(t) + x'(t)Sx(t)) dt + x'(T)Hx(T) + \nu |y - Dx(T)|^2.$$

If we introduce a further state vector  $z \in \mathbb{R}^r$  satisfying

$$\dot{z} = 0,$$

we may regard  $y$  as both the initial and final value of  $z$ . This leads to the representation

$$(7.5) \quad \begin{aligned} \tilde{Q}_\nu[u(\cdot), y] = & \int_{t_0}^T (u'(t)Ru(t) + x'(t)Sx(t)) dt \\ & + (x'(T), z'(T)) \begin{pmatrix} H + \nu D'D & -\nu D' \\ -\nu D & \nu I \end{pmatrix} \begin{pmatrix} x(T) \\ z(T) \end{pmatrix}, \end{aligned}$$

with

$$(7.6) \quad \begin{pmatrix} \dot{x} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} u, \quad \begin{pmatrix} x(t_0) \\ z(t_0) \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}.$$

As is well known, it follows that the minimum value of  $\tilde{Q}_\nu[u(\cdot), y]$ , for the fixed initial state given in (7.6), over all  $u(\cdot) \in X$ , is given by the expression

$$(7.7) \quad (0', y') \tilde{P}(t_0) \begin{pmatrix} 0 \\ y \end{pmatrix},$$

where  $\tilde{P}(\cdot)$  denotes the solution of the  $(n+r)$ -square Riccati problem associated with (7.5) and (7.6) in the same way that (7.4) is associated with (7.3) and (7.1), *provided* that the solution  $\tilde{P}(\cdot)$  extends over the whole interval  $[t_0, T]$ . Because of the zero blocks in (7.6), this new Riccati problem is found to decompose into the following coupled trio, in which we use the block notation

$$\tilde{P}(\cdot) = \begin{pmatrix} P_{11}(\cdot) & P_{12}(\cdot) \\ P'_{12}(\cdot) & P_{22}(\cdot) \end{pmatrix};$$

$$(7.8) \quad -\dot{P}_{11} = P_{11}A + A'P_{11} - P_{11}BR^{-1}B'P_{11} + S, \quad P_{11}(T) = H + \nu D'D$$

$$(7.9) \quad -\dot{P}_{12} = A'P_{12} - P_{11}BR^{-1}B'P_{12}, \quad P_{12}(T) = -\nu D'$$

$$(7.10) \quad -\dot{P}_{22} = -P'_{12}BR^{-1}B'P_{12}, \quad P_{22}(T) = \nu I.$$

One notices immediately that the only nonlinearity occurs in (7.8), which problem is identical with (7.4). Thus, for given  $\nu$  and  $t_0$ , if (7.4) does not “blow up”, neither does (7.8)–(7.10), and, by (6.2) and (7.7), for any  $y \in \mathbb{R}^r$  we have

$$(7.11) \quad C_\nu(y) = y'P_{22}(t_0)y.$$

Since  $\Gamma = \mathbb{R}_+^m$ , condition (b) in Theorem 1.3 therefore requires that

$$(7.12) \quad y'P_{22}(t_0)y > 0 \quad \text{whenever } y \geq 0, y \neq 0.$$

Symmetric matrices which have this property are called *strictly copositive*—they were first studied by T. S. Motzkin in 1952, who in [9] and [10] gave the following test for strict copositivity of a symmetric matrix  $M$ :  $M$  is strictly copositive iff its diagonal entries are positive, and for every principal submatrix  $\bar{M}$  of  $M$ , whenever the cofactors of the last row of  $\bar{M}$  are all positive, then so is the determinant of  $\bar{M}$ . (See also [6] for further discussion and references.) Thus provided  $r$  is not too large, (7.12) can be readily tested. It follows that for any given  $\nu > 0$ , we can determine the infimum of initial times  $t_0$  for which that  $\nu$  satisfies conditions (a) and (b) of Theorem 1.3 simply by solving the coupled system (7.8)–(7.10) backwards in time, and noting the time  $t_0(\nu)$  at which *either the solution blows up or  $P_{22}(\cdot)$  loses the property of strict copositivity*. Note that because of the remark following equation (4.5), if for some  $t_0$ ,  $\nu$  satisfies (a) and (b), then so does any larger  $\nu$ . Consequently  $t_0(\nu)$  is a nonincreasing function.

Theorem 1.3 may now be applied to conclude that the desired focal point  $t_0^*$  is given by

$$t_0^* = \inf_{\nu > 0} t_0(\nu) = \lim_k t_0(\nu_k),$$

for any sequence

$$\nu_k \rightarrow \infty.$$

This scheme was successfully applied to the following example. For the “double-integrator” control system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad \text{with } x(t_0) = 0,$$

find the infimum time  $t_0^*$  of times  $t_0 < 0$  for which

$$\int_{t_0}^0 (u^2 - x_1^2) dt > 0,$$

whenever  $u(\cdot) \neq 0$  and

$$x_1(0) \geq 0, \quad x_2(0) \leq 0.$$

Firstly, the Riccati system (7.8)–(7.10) can be solved analytically, and, writing

$$c = \cos t, \quad s = \sin t, \quad C = \cosh t, \quad S = \sinh t$$

for brevity, the result for  $P_{22}(t)$  is given by

$$(7.13) \quad \Delta \cdot P_{22} = \nu^2 \begin{pmatrix} -cS - sC & sS \\ sS & cS - sC \end{pmatrix} + \nu(1 + cC) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where

$$(7.14) \quad \Delta(t, \nu) = \nu^2(1 - cC) - 2\nu sC + 1 + cC.$$

Let

$$t_1 = -4.730040 \dots$$

denote the first negative zero of the coefficient

$$1 - \cos t \cosh t$$

of  $\nu^2$  in (7.14). For any  $t_0 > t_1$ , this coefficient is positive on  $[t_0, 0)$ , and so for all sufficiently large  $\nu$ , the Riccati system does not blow up and  $\Delta > 0$  on  $[t_0, 0]$ .

Let  $M(t)$  denote the  $2 \times 2$  coefficient matrix of  $\nu^2$  in (7.13). If  $M(t)$  is strictly copositive, and  $t > t_1$ , then, for all sufficiently large  $\nu > 0$ ,  $P_{22}(t)$  will also be copositive. Thus if  $t_2$  denotes the first negative time at which  $M(t)$  ceases to be strictly copositive, the desired infimum time  $t_0^*$  is given by

$$t_0^* = \max(t_1, t_2).$$

Since the off-diagonal entry is positive in  $(-\pi, 0)$ , it follows from Motzkin's test that  $M(t)$  loses strict copositivity in this range only when one of the diagonal entries ceases to be positive. This distinction falls first to the leading entry, and thus  $t_2$  is the first negative zero of

$$(7.15) \quad \cos t \sinh t + \sin t \cosh t.$$

We thus establish analytically that

$$(7.16) \quad t_0^* = t_2 = -2.365020 \dots$$

is the first negative zero of the function (7.15).

As a numerical test, (7.8)–(7.10) for this example were integrated using a Kutta-Merson fourth-order method, checking the strict copositivity of  $P_{22}$  every 0.02 time units, and using cubic polynomial interpolation to find the time at which strict copositivity was lost. With truncation error controlled by a mixed test of the form

$$|\delta P| < 10^{-5}(1 + |P|),$$

the results for  $\nu = 1, 2, 4, 8, 16, 32$  are given in the following table.

$\nu$	$t_0(\nu)$
1	-2.144572
2	-2.225842
4	-2.285386
8	-2.322211
16	-2.342798
32	-2.343695

Remarkably enough, these values of  $t_0(\nu)$  are all correct to the six decimal places shown, as may be verified by computing the first negative zero of the leading entry of  $P_{22}$ , which, from (7.13), is equal to the first negative zero of

$$-\nu(\cos t \sinh t + \sin t \cosh t) + 1 + \cos t \cosh t.$$

Finally, using cubic polynomials in  $\nu^{-1}$ , successive fours of these values were used to extrapolate for  $t_0^* = t_0(\infty)$ . These extrapolation results are given in the final table, and are satisfactorily close to  $t_0^*$  as given by (7.16).

Extrapolation from	$t_0^*$
$\nu = 1, 2, 4, 8$	-2.364624
$\nu = 2, 4, 8, 16$	-2.364991
$\nu = 4, 8, 16, 32$	-2.365018

**Acknowledgments.** In several respects this paper is a companion paper to [8], written in collaboration with D. H. Jacobson, and I should like to record my thanks to him for having interested me in the general question of conditional definiteness, and for many interesting discussions.

I am indebted also to T. Geveci, who pointed out an error in my first attempt to prove Theorem 1.3.

REFERENCES

[1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.  
 [2] P. FINSLER, *Über das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen*, Comment. Math. Helv., 9 (1937), pp. 188–192.  
 [3] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs NJ, 1963.  
 [4] M. R. HESTENES, *Applications of the theory of quadratic forms in Hilbert space to the calculus of variations*, Pacific J. Math., 1(4) (1951), pp. 525–581.  
 [5] ———, *Optimization Theory*, John Wiley, New York, 1975.  
 [6] D. H. JACOBSON, *Extensions of Linear Quadratic Control, Optimization and Matrix Theory*, Academic Press, New York, 1977.  
 [7] ———, *A generalization of Finsler’s theorem for quadratic inequalities and equalities*, Quaestiones Math., 1(1) (1976), pp. 19–28.  
 [8] D. H. MARTIN AND D. H. JACOBSON, *Copositive matrices and definiteness of quadratic forms subject to homogeneous linear inequality constraints*, Lin. Alg. Appl., to appear.  
 [9] T. S. MOTZKIN, *Quadratic forms positive for non-negative variables not all zero*, Notices Amer. Math. Soc., 12 (1965), p. 224.  
 [10] T. S. MOTZKIN, *Signs of minors*, Inequalities, O. Shisha, ed., Academic Press, New York, 1967.

## UNIFORMLY VALID EXPANSIONS FOR LAPLACE INTEGRALS\*

L. A. SKINNER†

**Abstract.** Uniformly valid asymptotic expansions for a class of parameter-dependent Laplace integrals involving coalescing saddle points are obtained by a direct method based on matched asymptotic expansion theory. Bessel functions of large order are treated as an example.

**1. Introduction.** The idea in this paper is to use singular perturbation techniques to obtain asymptotic expansions of functions defined by integrals of Laplace type. We shall, in particular, establish new uniformly valid expansions for Laplace integrals involving coalescing saddle points. A key feature of the method to be described is its freedom from dependence on nonlinear transformations. Computationally, it resembles the procedure developed by Erdelyi and Wyman [3] in which a certain fundamental factor of the integrand is identified and an infinite series is introduced for the remainder. The significant difference is that in our asymptotic expansions the gauge functions depend only on the large parameter in the problem.

Throughout this paper the conventional order symbols will be used exclusively to denote uniform estimates. Thus  $w(r, \varepsilon) = \phi(r, \varepsilon) + O(\varepsilon^N)$  as  $\varepsilon \rightarrow 0^+$ , for  $0 \leq r \leq 1$ , or all  $r \in [0, 1]$ , means there exist constants  $K, \delta > 0$  such that  $|w(r, \varepsilon) - \phi(r, \varepsilon)| \leq K\varepsilon^N$ , for all  $(r, \varepsilon) \in [0, 1] \times (0, \delta]$ . We shall write  $f(r) \in C^N[0, \infty]$  if  $f(r) \in C^N[0, 1]$  and  $g(r) \in C^N[0, 1]$ , where  $g(r) = f(1/r)$ . Also, for  $m \geq 0$ ,

$$(1.1) \quad f^{[m]}(r) = \frac{1}{m!} \left( \frac{d}{dr} \right)^m f(r),$$

and

$$(1.2) \quad f^{[-m]}(r) = \frac{1}{m!} \left( -r^2 \frac{d}{dr} \right)^m f(r).$$

Thus if  $f(r) \in C^\infty[0, \infty]$ ,  $f^{[m]}(0)$  is the coefficient of  $r^m$  in the (asymptotic) power series for  $f(r)$  as  $r \rightarrow 0^+$ , and  $f^{[-m]}(\infty)$  is the coefficient of  $r^{-m}$  in the corresponding series for  $r \rightarrow \infty$ . Analogous notation will be used for functions of more than one variable. In particular, for  $m, n \geq 0$ ,

$$(1.3) \quad f^{[m, -n]}(r, R) = \frac{1}{(m!)(n!)} \left( \frac{d}{dr} \right)^m \left( -R^2 \frac{d}{dR} \right)^n f(r, R).$$

For short,  $f^{[m]}(r, R, t) = f^{[m, 0, 0]}(r, R, t)$ . Finally,  $f(\varepsilon) = o(\varepsilon^\infty)$  as  $\varepsilon \rightarrow 0^+$  means  $f(\varepsilon) = o(\varepsilon^n)$  as  $\varepsilon \rightarrow 0^+$  for any fixed  $n$ .

Suppose for  $\varepsilon \neq 0$  that  $w(r, \varepsilon) = f(r, r/\varepsilon)$ , where  $f(r, R) \in C^{2N-1}([0, b] \times [0, \infty))$ . Then if  $\delta \in (0, b]$ ,

$$(1.4) \quad w(r, \varepsilon) = O_N w(r, \varepsilon) + O(\varepsilon^N), \quad \text{as } \varepsilon \rightarrow 0^+ \text{ for } \delta \leq r \leq b,$$

where

$$(1.5) \quad O_N w(r, \varepsilon) = \sum_{n=0}^{N-1} \left( \frac{\varepsilon}{r} \right)^n f^{[0, -n]}(r, \infty).$$

\* Received by the editors April 7, 1979, and in revised form April 8, 1980.

† Department of Mathematical Sciences, University of Wisconsin, Milwaukee, Wisconsin 53201.

Following Van Dyke [9] and Fraenkel [4], we call the function  $O_N w(r, \epsilon)$  the  $N$ -term outer expansion of  $w(r, \epsilon)$ . The matching  $M$ -term inner expansion is

$$(1.6) \quad I_M w(r, \epsilon) = \sum_{m=0}^{M-1} r^m f^{[m,0]} \left( 0, \frac{r}{\epsilon} \right).$$

Indeed, for  $M \leq N + 1$ ,

$$(1.7) \quad O_N I_M w(r, \epsilon) = \sum_{n=0}^{N-1} \left( \frac{\epsilon}{r} \right)^n \sum_{m=0}^{M-1} r^m f^{[m,-n]}(0, \infty) = I_M O_N w(r, \epsilon).$$

It follows that the composite expansion

$$(1.8) \quad C_N w(r, \epsilon) = [O_N + I_N - O_N I_N] w(r, \epsilon)$$

has the same inner and outer expansions, up to  $N$  terms, as  $w(r, \epsilon)$ . In fact we have the following theorem.

**THEOREM 1.** *If  $w(r, \epsilon) = f(r, r/\epsilon)$ , for  $0 \leq r \leq b$ ,  $0 < \epsilon \leq \epsilon_0$ , where  $f(r, R) \in C^\infty([0, b] \times [0, \infty])$ , then for  $N \geq 0$ ,*

$$(1.9) \quad w(r, \epsilon) = C_N w(r, \epsilon) + O(\epsilon^N), \quad \text{as } \epsilon \rightarrow 0^+ \text{ for } 0 \leq r \leq b.$$

*Proof.* Consider  $0 \leq r \leq \epsilon^{1/2}$  and  $\epsilon^{1/2} \leq r \leq b$ , separately, as suggested in Part II of [4]. If  $\epsilon \in (0, 1]$  and  $\epsilon^{1/2} \leq r \leq b$ , then  $r^m (\epsilon/r)^{2N} \leq \epsilon^N$  for  $0 \leq m \leq N$  and  $r^N (\epsilon/r)^n \leq \epsilon^N$  for  $n \geq N$ . Therefore

$$(1.10) \quad w(r, \epsilon) = O_{2N} w(r, \epsilon) + O(\epsilon^N),$$

and

$$(1.11) \quad I_N w(r, \epsilon) = O_{2N} I_N w(r, \epsilon) + O(\epsilon^N), \quad \text{as } \epsilon \rightarrow 0^+ \text{ for } \epsilon^{1/2} \leq r \leq b.$$

In other words,

$$(1.12) \quad w(r, \epsilon) = C_N w(r, \epsilon) + A_N w(r, \epsilon) + O(\epsilon^N),$$

where  $A_N = (O_{2N} - O_N) + I_N(O_{2N} - O_N)$ . A straightforward calculation reveals

$$(1.13) \quad A_N w(r, \epsilon) = r^N \sum_{n=N}^{2N-1} \left( \frac{\epsilon}{r} \right)^n \theta_n(r),$$

where

$$(1.14) \quad \theta_n(r) = r^{-N} \left[ f^{[0,-n]} \left( r, \frac{r}{\epsilon} \right) - \sum_{m=0}^{N-1} r^m f^{[m,-n]}(0, \infty) \right].$$

Clearly  $\theta_n(r)$  is a bounded function, indeed  $\theta_n(r) \in C^\infty[0, b]$ , and therefore  $A_N w(r, \epsilon) = O(\epsilon^N)$  as  $\epsilon \rightarrow 0^+$  for  $\epsilon^{1/2} \leq r \leq b$ . The validity of (1.9) for  $0 \leq r \leq \epsilon^{1/2}$  is similarly established by adding and subtracting  $I_{2N} w(r, \epsilon)$  and  $I_{2N} O_N w(r, \epsilon)$ .

Actually, (1.9) holds even if  $f(r, R)$  is only of class  $C^{2N-1}$  on  $[0, b] \times [0, \infty]$ . However, we shall not pursue this point here. For our present purposes we just need a straightforward corollary to Theorem 1.

**COROLLARY 1.** *Let  $\Delta$  denote a closed (not necessarily bounded) region in the complex plane. If  $f(r, R, t) \in C^\infty([0, b] \times [0, \infty] \times \Delta)$  and if  $f^{[0,-n,0]}(r, \infty, t) = 0$ , for  $0 \leq n \leq N$ , then*

$$(1.15) \quad f \left( r, \frac{r}{\epsilon}, t \right) = \sum_{n=0}^N \epsilon^n \left[ \left( \frac{r}{\epsilon} \right)^n f^{[n]} \left( 0, \frac{r}{\epsilon}, t \right) \right] + O(\epsilon^{N+1}),$$

as  $\epsilon \rightarrow 0^+$  for all  $(r, t) \in [0, b] \times \Delta$ .



**2. Laplace integrals.** We are interested in the asymptotic evaluation of integrals of the form

$$(2.1) \quad I_m(t, \nu) = \nu^{1/m} \int_0^b e^{-\nu h(r,t)} g(r, t) dr,$$

where  $g(r, t)$  and  $h(r, t)$  are  $C^\infty$  functions on  $[0, b] \times \Delta$  and  $\text{Re} [h(r, t)]$  has its minimum (with respect to  $r$ ) at  $r = 0$ . Let  $a_n(t) = h^{[n]}(0, t)$  and  $b_n(t) = g^{[n]}(0, t)$ . Without loss of generality we shall assume  $a_0(t) = a_1(t) = 0$ .<sup>1</sup>

By way of introduction, suppose  $\text{Re} [a_2(t)] > 0$ , for all  $t \in \Delta$ . If we put  $u(r, t) = r^{-2}h(r, t)$  then

$$(2.2) \quad I_2(t, \nu) = \nu^{-1/2} \int_0^b \phi(r, \nu^{1/2}r, t) dr,$$

where

$$(2.3) \quad \phi(r, R, t) = g(r, t) \exp [-R^2 u(r, t)].$$

Also,  $u(r, t) \in C^\infty([0, b] \times \Delta)$  and  $\text{Re} [u(r, t)] > 0$ , for all  $(r, t) \in [0, b] \times \Delta$ . Hence  $\phi(r, R, t) \in C^\infty([0, b] \times [0, \infty] \times \Delta)$ . Thus by Corollary 1 it follows almost immediately that

$$(2.4) \quad I_2(t, \nu) = \sum_{n=0}^{N-1} \nu^{-n/2} P_n(t) + O(\nu^{-N/2}), \quad \text{as } \nu \rightarrow \infty \text{ for all } t \in \Delta,$$

where

$$(2.5) \quad P_n(t) = \int_0^\infty R^n \phi^{[n]}(0, R, t) dR.$$

This, of course, is just what one gets by the classical Laplace method. From the power series expansions for  $g(r, t)$  and  $h(r, t)$ , it is readily verified that

$$(2.6) \quad \phi^{[n]}(0, R, t) = p_n(R, t) \exp [-a_2(t)R^2],$$

where

$$(2.7a, b) \quad p_0(R, t) = b_0(t), \quad p_1(R, t) = b_1(t) - b_0(t)a_3(t)R^2,$$

and

$$(2.7c) \quad p_2(R, t) = b_2(t) - [b_1(t)a_3(t) + b_0(t)a_4(t)]R^2 + \frac{1}{2}b_0(t)a_3^2(t)R^4.$$

In general  $p_n(R, t)$  is a polynomial of degree  $n$  in  $R^2$ .

The problem of coalescing saddle points arises in connection with (2.1) when  $a_2(t)$  has one or more zeros in  $\Delta$ . Suppose  $0 \in \Delta$  and  $a_k(0) = 0$ , for  $2 \leq k \leq m - 1$ , but  $\text{Re} [a_m(0)] > 0$ . Then, besides the saddle point of  $h(r, t)$  at  $r = 0$ , there will be additional saddle points (not necessarily  $m - 2$ ) which approach  $r = 0$  as  $t \rightarrow 0$ . We could proceed, at least in principle, according to the method of Chester, Friedman and Ursell [2], and transform the variable of integration in (2.1) from  $r$  to, say,  $s = (r, t)$  so that  $h(r, t)$  is replaced by a polynomial in  $s$  of degree  $m$ . Thus, as in [1] and [8],

$$(2.8) \quad h(r, t) = \frac{1}{m} s^m(r, t) + \sum_{k=0}^{m-2} \zeta_k(t) s^k(r, t),$$

<sup>1</sup> If at first  $a_1(t) \neq 0$ , let  $r' = r^{1/2}$ .

where the coefficients  $\zeta_k(t)$  are determined by the condition  $s^{[1]}(\zeta(t), t) \neq 0$ , for any coalescing saddle point solution  $r = \xi(t)$  of  $h^{[1]}(r, t) = 0$ . Unfortunately, explicit expressions for these functions generally do not exist. Furthermore as noted in [8], even computation of their power series expansions is not very practical when more than a few terms are required. We propose instead to generalize the analysis leading to (2.4).

Under the above assumptions on  $a_k(t)$  for  $0 \leq k \leq m$ , if we let

$$(2.9) \quad v_m(r, t) = r^{-m}[h(r, t) - \alpha_m(r, t)],$$

where

$$(2.10) \quad \alpha_m(r, t) = \sum_{k=2}^{m-1} a_k(t)r^k$$

then  $\text{Re}[v_m(r, 0)] > 0$  and therefore  $\text{Re}[v_m(r, t)] > 0$  in a neighborhood of  $t = 0$ . Hence we now want to rewrite the integrand for (2.1) in terms of

$$(2.12) \quad \psi_m^{[n]}(0, R, t) = q_{mn}(R, t) \exp[-a_m(t)R^m],$$

Corresponding to (2.6) and (2.7), we have

$$(2.12) \quad \psi_m^{[n]}(0, R, t) = q_{mn}(R, t) \exp[-a_m(t)R^m],$$

where

$$(2.13a, b) \quad q_{m0}(R, t) = 0, \quad q_{m1}(R, t) = b_1(t) - b_0(t)a_{m+1}(t)R^m,$$

$$(2.13c) \quad q_{m2}(R, t) = b_2(t) - [b_1(t)a_{m+1}(t) + b_0(t)a_{m+2}(t)]R^m + \frac{1}{2}b_0(t)a_{m+1}^2(t)R^{2m},$$

and in general  $q_{mn}(R, t)$  is a polynomial of degree  $n$  in  $R^m$ .

**THEOREM 2.** Let  $I_m(t, \nu)$  be the function defined by (2.1) where  $\text{Re}[h(r, t)] > 0$  on  $(0, b] \times \Delta$ ,  $g(r, t)$  and  $h(r, t)$  are of class  $C^\infty$  on  $[0, b] \times \Delta$ , and  $a_0(t) = a_1(t) = 0$ . Assume  $0 \in \Delta$  and  $a_k(0) = 0$ , for  $2 \leq k \leq m-1$ , but  $\text{Re}[a_m(0)] > 0$ , and let  $S[\rho] = \{t : t \in \Delta, |t| \leq \rho\}$ . If there exist  $b_0 \in (0, b)$  and  $c_0 > 0$  such that  $\text{Re}[\alpha_m(r, t)] \geq 0$  on  $[0, b_0] \times S[c_0]$ , where  $\alpha_m(r, t)$  is the function defined by (2.10), then there exists  $\rho_0 > 0$  such that

$$(2.14) \quad I_m(t, \nu) = \sum_{n=0}^{N-1} \nu^{-n/m} Q_{mn}(t, \nu^{1-2/m} a_2(t), \dots, \nu^{1/m} a_{m-1}(t)) + O(\nu^{-N/m}),$$

as  $\nu \rightarrow \infty$  for all  $t \in S[\rho_0]$ , where

$$(2.15) \quad Q_{mn}(t, T_2, \dots, T_{m-1}) = \int_0^\infty R^n q_{mn}(R, t) \exp\left[-a_m(t)R^m - \sum_{k=2}^{m-1} T_k R^k\right] dR.$$

*Proof.* Let  $\varepsilon = \nu^{-1/m}$ . Then

$$(2.16) \quad I_m(t, \nu) = \varepsilon^{-1} \int_0^{b_0} \psi_m\left(r, \frac{r}{\varepsilon}, t\right) \exp[-\nu \alpha_m(r, t)] dr + o(\varepsilon^\infty),$$

as  $\varepsilon \rightarrow 0^+$  for all  $t \in \Delta$ , where  $\psi_m(r, R, t)$  is given by (2.11), and, as already noted, there exists  $\rho_1 \in (0, c_0]$  such that  $\psi_m(r, R, t) \in C^\infty([0, b_0] \times [0, \infty] \times S[\rho_1])$ . Also,  $|\exp[-\nu \alpha_m(r, t)]| \leq 1$ , for all  $(r, t) \in [0, b_0] \times S[\rho_1]$  and therefore, by Corollary 1,

$$(2.17) \quad I_m(t, \nu) = \varepsilon^{-1} \sum_{n=0}^N \varepsilon^n \int_0^{b_0} \left(\frac{r}{\varepsilon}\right)^n \psi_m^{[n]}(0, \frac{r}{\varepsilon}, t) \exp[-\nu \alpha_m(r, t)] dr + O(\varepsilon^N),$$

as  $\varepsilon \rightarrow 0^+$  for all  $t \in S[\rho_1]$ . Let

$$(2.18) \quad \beta_m(r, t) = \sum_{k=2}^m a_k(t)r^{k-2} = r^{-2}[\alpha_m(r, t) + r^m a_m(t)].$$

Since  $\text{Re} [\beta_m(r, 0)] > 0$  for  $r > 0$ , there exists  $\rho_2 \in (0, c_0]$  such that  $\text{Re} [\beta_m(r, t)] > 0$  for all  $(r, t) \in [b_0, \infty) \times S[\rho_2]$ . Therefore, in view of (2.12),

$$(2.19) \quad \int_{b_0}^{\infty} \left(\frac{r}{\varepsilon}\right)^n \psi_m^{[n]} \left(0, \frac{r}{\varepsilon}, t\right) \exp[-\nu \alpha_m(r, t)] dr = o(\varepsilon^\infty),$$

as  $\varepsilon \rightarrow 0^+$  for all  $t \in S[\rho_2]$ . Thus we can replace  $b_0$  in (2.17) by  $\infty$ . It remains to show that

$$(2.20) \quad Q_{mn}(t, \varepsilon^{2-m} a_2(t), \dots, \varepsilon^{-1} a_{m-1}(t)) = O(1),$$

as  $\varepsilon \rightarrow 0^+$  for all  $t \in S[\rho_0]$ , where  $\rho_0$  is the minimum of  $\rho_1$  and  $\rho_2$ . But this follows directly from (2.19) and the fact that  $\text{Re} [a_m(t)] = \text{Re} [v_m(0, t)] > 0$ , for all  $t \in S[\rho_0]$ .

The assumption on  $\alpha_m(r, t)$  in Theorem 2 can be met in several ways. It obviously is satisfied if  $\text{Re} [a_k(t)] \geq 0$ , for  $2 \leq k \leq m-1$ . On the other hand, since  $h(r, t) = r^2[a_2(t) + O(r)]$  and  $\text{Re} [h(r, t)] > 0$ , for  $r \neq 0$ , we must have  $\text{Re} [a_2(t)] \geq 0$  anyway. Thus for  $m=3$  the assumption is redundant. As another possibility, suppose  $\text{Re} [a_2^{[1]}(0) e^{i\theta}] > 0$ , for  $\alpha \leq \theta \leq \beta$ . Then, since

$$(2.21) \quad \alpha_m(r, t) = r^2 t [a_2^{[1]}(0) + O((r^2 + t^2)^{1/2})],$$

there exists  $b_0, c_0 > 0$  such that  $(r^2|t|)^{-1} \text{Re} [\alpha_m(r, t)] > 0$  and therefore  $\text{Re} [\alpha_m(r, t)] \geq 0$ , for all  $(r, t) \in [0, b_0] \times V[c_0]$ , where  $V[\rho] = \{t: |t| \leq \rho, \alpha \leq \arg t \leq \beta\}$ . These considerations also suggest the following result.

**THEOREM 3.** *If the conditions of Theorem 2 are satisfied and if  $\text{Re} [\beta_m(r, t)] > 0$ , for all  $(r, t) \in [0, \infty) \times \Delta'$  where  $\Delta' = \Delta - S[\rho_0]$ , then (2.14) holds for all  $t \in \Delta$ .*

*Proof.* Note that

$$(2.22) \quad Q_{mn}(t, \nu^{1-2/m} a_2(t), \dots, \nu^{1/m} a_{m-1}(t)) = \nu^{(1+n)/m} \int_0^\infty r^n F_{mn}(r, \nu^{1/2} r, t) dr,$$

where

$$(2.23) \quad F_{mn}(r, R, t) = q_{mn}([R^2 r^{m-2}]^{1/m}, t) \exp[-R^2 \beta_m(r, t)].$$

Since  $\text{Re} [\beta_m(r, t)] > 0$  on  $[0, \infty) \times \Delta'$ , we have

$$(2.24) \quad \int_{b_0}^\infty r^n F_{mn}(r, \nu^{1/2} r, t) dr = o(\nu^{-\infty}),$$

as  $\nu \rightarrow \infty$  for all  $t \in \Delta'$ , and  $F_{mn}(r, R, t) \in C^\infty([0, b_0] \times [0, \infty) \times \Delta')$ . The condition on  $\beta_m(r, t)$  means also that  $\text{Re} [a_m(t)] > 0$  on  $\Delta'$ , and therefore

$$(2.25) \quad \psi_m(r, R, t) = \sum_{n=0}^{N-1} r^n q_{mn}(R, t) \exp[-a_m(t) R^m] + O(r^N),$$

as  $r \rightarrow 0^+$  for all  $(R, t) \in [0, \infty) \times \Delta$ . Now note that

$$(2.26) \quad \phi(r, R, t) = \psi_m(r, [R^2 r^{m-2}]^{1/m}, t) \exp[-R^2 \gamma_m(r, t)],$$

where  $\phi(r, R, t)$  is defined by (2.2) and

$$(2.27) \quad \gamma_m(r, t) = \sum_{k=2}^{m-1} a_k(t)r^{k-2} = r^{-2} \alpha_m(r, t).$$

Since  $\text{Re} [\gamma_m(0, t)] > 0$  on  $\Delta'$ , it follows from (2.25) and (2.26) that

$$(2.28) \quad \phi(r, R, t) = \Phi_{mN}(r, R, t) + O(r^N),$$

as  $r \rightarrow 0^+$  for all  $(R, t) \in [0, \infty] \times \Delta'$ , where

$$(2.29) \quad \Phi_{mN}(r, R, t) = \sum_{n=0}^{N-1} r^n F_{mn}(r, R, t).$$

This means that  $\Phi_{mN}^{[n]}(0, R, t) = \phi^{[n]}(0, R, t)$ , for  $0 \leq n \leq N-1$ . Therefore, by Corollary 1,

$$(2.30) \quad \nu^{1/2} \int_0^{b_0} \Phi_{mN}(r, \nu^{1/2}r, t) dr = \sum_{n=0}^{N-1} \nu^{-n/2} P_n(t) + O(\nu^{-N/2}),$$

as  $\nu \rightarrow \infty$  for all  $t \in \Delta'$ , where  $P_n(t)$  is given by (2.5), which is what we needed to show.

A comparable extension of the Chester, Friedman and Ursell theory for the special case of two coalescing saddle points has been given by Ursell [7]. In this case we require  $\text{Re} [a_3(t)] > 0$  on  $\Delta$  which is analogous (but not equivalent) to Ursell's boundedness condition on  $\zeta(t)$ .

As an illustration of these results, let  $h(r, t) = tr^2 + \frac{1}{2}r^3 \log(1+r)$  and consider

$$(2.31) \quad E(t, \nu) = \int_0^1 e^{-\nu h(r,t)} dr.$$

If we choose  $\delta > 0$  and let  $V[\rho] = \{t: |t| \leq \rho, |\arg t| \leq \pi/2 - \delta\}$ , then for any finite value of  $c > 0$ ,  $h(r, t) \in C^\infty([0, 1] \times V[c])$  and  $\text{Re} [h(r, t)]$  has its minimum at  $r = 0$ . From

$$(2.32) \quad h(r, t) = tr^2 + \frac{1}{2}r^4 - \frac{1}{4}r^5 + O(r^6),$$

we see that  $m = 4$  in this example and, referring to (2.13),

$$(2.33) \quad q_{40}(R, t) = 1, \quad q_{41}(R, t) = \frac{1}{4}R^4.$$

Also we have

$$(2.34) \quad \alpha_4(r, t) = tr^2, \quad \beta_4(r, t) = t + \frac{1}{2}r^2,$$

and the conditions on these functions in Theorems 2 and 3 are obviously satisfied. Therefore, substituting into (2.14) and (2.15),

$$(2.35) \quad E(t, \nu) = \nu^{-1/4} E_0(\nu^{1/2}t) + \frac{1}{4}\nu^{-1/2} E_5(\nu^{1/2}t) + O(\nu^{-3/4}),$$

as  $\nu \rightarrow \infty$  for all  $t \in V[c]$ , where

$$(2.36) \quad E_n(T) = \int_0^\infty R^n \exp[-\frac{1}{2}R^4 - TR^2] dR.$$

Alternatively,

$$(2.37) \quad E_0(T) = \left(\frac{\pi}{4}\right)^{1/2} D_{-1/2}(T) \exp[\frac{1}{4}T^2],$$

where  $D_{-1/2}$  is the parabolic cylinder function of order  $-\frac{1}{2}$ , and

$$(2.38) \quad E_5(T) = \left(\frac{\pi}{8}\right)^{1/2} (1 + T^2) \text{erfc}(2^{-1/2}T) \exp[\frac{1}{2}T^2] - \frac{1}{2}T.$$

Treatment of (2.31) by the method of Chester, Friedman and Ursell would involve the

other two coalescing saddle point positions, say  $r = \xi(t)$  and  $r = \eta(t)$ , which satisfy

$$(2.39) \quad 2t + \frac{3}{2}r^2 \log(1+r) + \frac{1}{2}r(1+r)^{-1} = 0,$$

plus the determination of  $s(0, t)$ ,  $s(\xi(t), t)$ ,  $s(\eta(t), t)$ , and  $\zeta_k(t)$  for  $k = 0, 1$  and  $2$  from the requirement that (2.8) and the derivative equation,

$$s^3(r, t) + \zeta_1(t) + 2\zeta_2(t)s(r, t) = 0,$$

must be satisfied for  $r = 0, \xi(t)$  and  $\eta(t)$ .

**3. Bessel function expansions.** Theorems 2 and 3 lead to some interesting results for the Bessel function

$$(3.1) \quad J_\nu(\nu \operatorname{sech} t) = \frac{1}{2\pi i} \int_{\infty-i\pi}^{\infty+i\pi} e^{-\nu f(z,t)} dz, \quad |\arg(\operatorname{sech} t)| < \frac{\pi}{2},$$

where  $f(z, t) = z - \sinh z \operatorname{sech} t$ . The results are different from those established in [2], [5] and [6], which involve  $\zeta(t) = [\frac{3}{2}(t - \tanh t)]^{2/3}$ . They include, for example, the first order result [10, § 8.43]

$$(3.2) \quad J_\nu(\nu \operatorname{sech} t) = \nu^{-1/3} e^{-\nu f(t,t)} [G_0(\nu^{1/3} \tanh t) + O(\nu^{-2/3})],$$

as  $\nu \rightarrow \infty$  for  $0 \leq t \leq \infty$ , where  $G_0(T)$  is defined below.

Choose  $\delta > 0$  and again let  $V[\rho] = \{t: |t| \leq \rho, |\arg t| < \pi/2 - \delta\}$ . Also let  $\lambda = \pi/2 - \frac{1}{4}\delta$ . Then

$$(3.3) \quad \operatorname{Re} [f^{[3]}(t, t) e^{\pm 3i\lambda}] = \frac{1}{6} \sin(\frac{3}{4}\delta) > 0,$$

and, since  $\tanh t = t + O(t^3)$  as  $t \rightarrow 0$ , there exists  $c > 0$  such that

$$(3.4) \quad \operatorname{Re} [f^{[2]}(t, t) e^{\pm 2i\lambda}] = \frac{1}{2}|t| \cos(\arg t \pm \frac{1}{2}\delta) + O(t^3) \geq 0,$$

for all  $t \in V[c]$ . This means we can choose  $b > 0$  such that  $\operatorname{Re} [h^{(\pm)}(r, t)] > 0$  for all  $(r, t) \in (0, b] \times V[c]$ , where

$$(3.5) \quad h^{(\pm)}(r, t) = f(t + r e^{\pm i\lambda}, t) - f(t, t).$$

Therefore

$$(3.6) \quad J_\nu(\nu \operatorname{sech} t) = \frac{\nu^{-1/3}}{2\pi i} e^{-\nu f(t,t)} [e^{i\lambda} I^{(+)}(t, \nu) - e^{-i\lambda} I^{(-)}(t, \nu) + R(t, \nu)],$$

where the integrals

$$(3.7) \quad I^{(\pm)}(t, \nu) = \nu^{1/3} \int_0^b \exp[-\nu h^{(\pm)}(r, t)] dr,$$

both satisfy the hypotheses of Theorem 2, with  $m = 3$ , and

$$(3.8) \quad R(t, \nu) = \int_{\infty-i\pi}^{t+be^{-i\lambda}} + \int_{t+be^{i\lambda}}^{\infty+i\pi} e^{-\nu[f(z,t)-f(t,t)]} dz.$$

It is a straightforward matter to show that  $R(t, \nu) = o(\nu^{-\infty})$  as  $\nu \rightarrow \infty$ , for all  $t \in V[c]$ , provided  $c > 0$  is sufficiently small. Hence, introducing

$$(3.9) \quad F_n(T) = \frac{1}{2\pi i} \int_{\infty e^{-i\lambda}}^{\infty e^{i\lambda}} z^n \exp[\frac{1}{2}Tz^2 + \frac{1}{6}z^3] dz,$$

with a little computation we see by Theorem 2 that there exists  $\rho_0 > 0$  such that

$$(3.10) \quad J_\nu(\nu \operatorname{sech} t) = \nu^{-1/3} e^{-\nu f(t,t)} \left( \sum_{n=0}^{N-1} \nu^{-n/3} Q_n(t, \nu^{1/3} \tanh t) + O(\nu^{-N/3}) \right),$$

as  $\nu \rightarrow \infty$  for all  $t \in V[\rho_0]$ , where

$$(3.11a, b) \quad Q_0(t, T) = F_0(T), \quad Q_1(t, T) = \frac{1}{4!} (\tanh t) F_4(T),$$

and

$$(3.11c) \quad Q_2(t, T) = \frac{1}{5!} F_5(T) + \frac{1}{2} \left( \frac{1}{4!} \right)^2 (\tanh t)^2 F_8(T).$$

The functions  $F_n(T)$  are closely related to the Airy function

$$(3.12) \quad \operatorname{Ai}(T) = \frac{1}{2\pi i} \int_{\infty e^{-i\pi/3}}^{\infty e^{i\pi/3}} \exp \left[ \frac{1}{3} z^3 - Tz \right] dz.$$

In particular,

$$(3.13) \quad F_0(T) = 2^{1/3} \operatorname{Ai} \left( 2^{-2/3} T^2 \right) \exp \left[ \frac{1}{2} T^3 \right],$$

and

$$(3.14) \quad F_1(T) = -TF_0(T) - 2^{2/3} \operatorname{Ai}^{[1]} \left( 2^{-2/3} T^2 \right) \exp \left[ \frac{1}{3} T^3 \right].$$

Also,  $F_2(T) = -2TF_1(T)$  and for  $n \geq 1$ ,

$$(3.15) \quad F_{n+2}(T) + 2TF_{n+1}(T) + 2nF_{n-1}(T) = 0.$$

As  $T \rightarrow \infty$ ,  $F_{2n}(T) = O(T^{-n-1/2})$  and  $F_{2n+1}(T) = O(T^{-n-5/2})$  provided  $|\arg T| \leq \pi/2 - \delta$ . Therefore

$$(3.16) \quad Q_1(t, \nu^{1/3} \tanh t) = O(\nu^{-1/3}),$$

and

$$(3.17) \quad Q_2(t, \nu^{1/3} \tanh t) = \frac{1}{5!} F_5(\nu^{1/3} \tanh t) + O(\nu^{-2/3}),$$

as  $\nu \rightarrow \infty$  for all  $t \in V[\rho_0]$ . In fact it is true in general that

$$(3.18) \quad Q_{2n}(t, \nu^{1/3} \tanh t) = \sum_{k=0}^n \nu^{-2k/3} A_{nk}(\nu^{1/3} \tanh t),$$

and

$$(3.19) \quad Q_{2n+1}(t, \nu^{1/3} \tanh t) = \nu^{-1/3} \sum_{k=0}^n \nu^{-2k/3} B_{nk}(\nu^{1/3} \tanh t),$$

where  $A_{nk}(T) = O(1)$  and  $B_{nk}(T) = O(1)$  for all  $T$  such that  $|\arg T| \leq \pi/2 - \delta$ . We can therefore rewrite (3.10) in the more attractive form

$$(3.20) \quad J_\nu(\nu \operatorname{sech} t) = \nu^{-1/3} e^{-\nu f(t,t)} \left( \sum_{n=0}^{N-1} \nu^{-2n/3} G_n(\nu^{1/3} \tanh t) + O(\nu^{-2N/3}) \right),$$

as  $\nu \rightarrow \infty$  for all  $r \in S[\rho_0]$ , where  $G_0(T) = F_0(T)$  and

$$(3.21) \quad G_1(T) = \frac{1}{4!} [TF_4(T) + \frac{1}{3}F_5(T)],$$

or by (3.15),

$$(3.22) \quad G_1(T) = \frac{1}{10}T^2F_0(T) - \frac{1}{5}T^4F_1(T),$$

and, in general,

$$(3.23) \quad G_n(T) = p_n(T)F_0(T) + q_n(T)F_1(T),$$

where  $p_n(T)$  and  $q_n(T)$  are polynomials.

Comparable results are obtained for the Hankel functions

$$(3.24) \quad H_\nu^{(1)}(\nu \operatorname{sech} t) = \frac{1}{i\pi} \int_{\infty}^{\infty+i\pi} e^{-\nu f(z,t)} dz, \quad |\arg(\operatorname{sech} t)| < \frac{\pi}{2},$$

and

$$(3.25) \quad H_\nu^{(2)}(\nu \operatorname{sech} t) = \frac{-1}{i\pi} \int_{-\infty+i\pi}^{\infty} e^{\nu f(z,t)} dz, \quad |\arg(\operatorname{sech} t)| < \frac{\pi}{2}.$$

These results may be stated as follows. Let  $V_k[\rho] = \{t: |t| \leq \rho, |\arg t - \pi(1 - \frac{1}{3}k)| \leq \pi/2 - \delta\}$ . Then there exists  $\rho_0 > 0$  such that

$$(3.26) \quad H_\nu^{(1)}(\nu \operatorname{sech} t) = \nu^{-1/3} e^{-\nu f(t,t)} \left[ \sum_{n=0}^{N-1} \nu^{-2n/3} G_n^{(1)}(\nu^{1/3} e^{-2\pi i/3} \tanh t) + O(\nu^{-2N/3}) \right],$$

as  $\nu \rightarrow \infty$  for all  $t \in V_1[\rho_0]$  and

$$(3.27) \quad H_\nu^{(2)}(\nu \operatorname{sech} t) = \nu^{-1/3} e^{+\nu f(t,t)} \left[ \sum_{n=0}^{N-1} \nu^{-2n/3} G_n^{(2)}(\nu^{1/3} e^{i\pi/3} \tanh t) + O(\nu^{-2N/3}) \right],$$

as  $\nu \rightarrow \infty$  for all  $t \in V_2[\rho_0]$ , where for certain polynomials  $p_{nk}(T)$  and  $q_{nk}(T)$ ,

$$(3.28) \quad G_n^{(k)}(T) = p_{nk}(T)F_0(T) + q_{nk}(T)F_1(T).$$

In particular,

$$(3.29) \quad p_{0k}(T) = -2(-1)^k e^{-k\pi i/3}, \quad q_{0k}(T) = 0,$$

and

$$(3.30) \quad p_{1k}(T) = \frac{1}{5}e^{k\pi i/3}T, \quad q_{1k}(T) = \frac{2}{5}T^4.$$

In view of the connection formula

$$(3.31) \quad J_\nu(z) = \frac{1}{2}[H_\nu^{(1)}(z) + H_\nu^{(2)}(z)],$$

(3.26) and (3.27) together determine  $J_\nu(\nu \operatorname{sech} t)$  uniformly on  $|\arg t - \pi/2| \leq \pi/3 - \delta$ . This result along with (3.20), and the fact that  $\operatorname{sech}(-t) = \operatorname{sech} t$ , establishes the asymptotic behavior of  $J_\nu(\nu \operatorname{sech} t)$  in a full neighborhood of  $t = 0$ . In contrast to Chester, Friedman and Ursell we do not, however, obtain a single expansion for the whole neighborhood. On the other hand,  $\operatorname{Re} [f^{[2]}(t, t) e^{\pm 2i\lambda}] > 0$ , for all  $t > 0$ , and it is readily verified that  $R(t, \nu) = o(\nu^{-\infty})$ , for all  $t \in [0, \infty]$ . Therefore, by Theorem 3, (3.20) is uniformly valid for all  $t \in [0, \infty]$ . The Chester, Friedman and Ursell expansion for  $J_\nu(\nu \operatorname{sech} t)$  is known to be valid for unbounded complex  $t$ . But this is an independent result of Olver's [6] derived from the Bessel differential equation. From the asymptotic expansion of  $\operatorname{Ai}(z)$  and its derivative as  $z \rightarrow \infty$  in  $|\arg z| < \pi$  ([5]), it is clear that for  $t \neq 0$  expansions (3.20), (3.26) and (3.27) reduce to the classical Debye expansions, and thus are uniformly valid in certain unbounded regions of the complex  $\operatorname{sech} t$  plane determined by Watson [10, §§ 8.6, 8.61].

**Acknowledgment.** I am grateful to the referees for their constructive comments on an earlier version of this paper.

## REFERENCES

- [1] N. BLEISTEIN, *Uniform asymptotic expansions of integrals with many nearby stationary points and algebraic singularities*, J. Math. Mech., 17 (1967), pp. 533–559.
- [2] C. CHESTER, B. FRIEDMAN, AND F. URSELL, *An extension of the method of steepest descents*, Proc. Camb. Phil. Soc., 53 (1957), pp. 599–611.
- [3] A. ERDELYI AND M. WYMAN, *Asymptotic evaluation of certain integrals*, Arch. Rational Mech. Anal., 14 (1963), pp. 217–260.
- [4] L. E. FRAENKEL, *On the method of matched asymptotic expansions, Parts I–III*, Proc. Camb. Phil. Soc., 65 (1969), pp. 209–231, pp. 233–251, pp. 263–284.
- [5] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
- [6] ———, *The asymptotic expansion of Bessel functions of large order*, Phil Trans. Roy. Soc. A, 247 (1954), pp. 328–368.
- [7] F. URSELL, *Integrals with a large parameter, The continuation of uniformly asymptotic expansions*, Proc. Camb. Phil. Soc., 61 (1965), pp. 113–128.
- [8] ———, *Integrals with a large parameter, Several nearly coincident saddle points*, Proc. Camb. Phil. Soc., 72 (1972), pp. 49–65.
- [9] M. D. VAN DYKE, *Perturbation Methods in Fluid Mechanics*, Academic Press, New York, 1964.
- [10] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd Ed., Cambridge University Press, Cambridge, 1944.



## HARMONIC ANALYSIS ON BILATERAL CLASSES\*

KURT BERNARDO WOLF† AND THOMAS H. SELIGMAN‡

**Abstract.** The theory of harmonic analysis over coset and conjugation class spaces in groups is generalized to functions over the space of *bilateral classes*. The latter are novel equivalence sets which include the above as particular cases. The relevant orthogonal function bases are *partial traces*. The standard harmonic functions and characters are recovered as special examples.

**1. Introduction.** Equivalence classes of group elements are among the main objects of study not only of group theory per se, but of any branch of mathematical physics which requires homogeneous spaces for group action. Closely related to these is the theory of group representations and the associated harmonic analysis. All textbooks on this matter introduce the concepts of cosets and of conjugation classes, and the ensuing developments of harmonic functions and characters are ubiquitous throughout the literature. It is thus perhaps surprising that a generalization of these two cases of equivalence classes can be defined rather naturally, and certain consequences drawn which somehow seem to have escaped notice by several generations of thorough workers in this field.

The need for a more general classification of group elements in equivalence classes exhibiting a certain correlation between the right and left group action arose originally in applied studies in quantum chemistry [3]. They concerned the classification of transition amplitudes between certain molecules called permutational isomers, which differ only in the way in which the ligands are distributed on the skeleton. The mathematical meaning and subsequent construction of what are now called *bilateral classes* were explored shortly thereafter, and appeared in condensed form in [4]. A more complete discussion is given in [5]; some of the results of this paper were briefly summarized in [7].

As the construction of bilateral classes is not yet widely known, we shall restate the relevant points in § 2, stressing certain particular cases. Section 3 sets up the notation and the needed subgroup reduction adapted to the most general bilateral class partition such that functions of this space can be subject to a reduced harmonic analysis. The complete and orthogonal basis function set, which we call *partial traces*, is then constructed in § 4. In the two traditional cases of cosets and conjugation classes they reduce to the well-known spherical functions and characters. In § 5 we offer some concluding remarks.

**2. Short survey of bilateral classes.** The elements of a group  $G$  can be partitioned into a complete family of disjoint sets through an equivalence relation. Equivalence relations with group theoretical significance which have been fruitfully exploited are those which lead to left, right or double cosets, or conjugation subclasses: if  $H$  and  $K$  are subgroups of  $G$ ,  $g \in G$ ,  $h \in H$ ,  $k \in K$ , then the above relations are  $g' \sim g$  iff there exist  $h, k$ , such that, respectively,  $g' = hg$ ,  $g' = gk$ ,  $g' = hkg$  or  $g' = ghk^{-1}$ .

A generalization of the above relations with group-theoretical definition makes use of the following construction.

(a) Let  $(g_1, g_2) \in G \times G$ , and consider the action of this group on the elements  $g \in G$

\* Received by the editors November 2, 1979, and in final revised form December 21, 1979.

† Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México 20 D.F., Mexico.

‡ Instituto de Física, Universidad Nacional Autónoma de México, México 20 D.F., Mexico.

given by

$$(1) \quad g \xrightarrow{(g_1, g_2)} g' = g_1 g g_2^{-1}.$$

(b) Select a subgroup  $P \subset G \times G$  and introduce the following equivalence relation between the elements of  $G$ :

$$(2) \quad g' \overset{P}{\sim} g \Leftrightarrow \exists (g_1, g_2) \in P | g' = g_1 g g_2^{-1}.$$

The properties of symmetry, reflexivity, and transitivity hold for  $\overset{P}{\sim}$ , since  $P$  is a group.

The equivalence relation  $\overset{P}{\sim}$  thus partitions  $G$  into a complete family of disjoint sets which we call bilateral classes or, more specifically,  $P$ -bilateral classes. The bilateral class containing an element  $g_i \in G$  is the set

$$(3) \quad B_i^P = \{g_1 g_i g_2^{-1}, (g_1, g_2) \in P \subset G \times G\}.$$

The exploration of all possible subgroups  $P$  of  $G \times G$  (sometimes called *subdirect* products of  $G$  with  $G$ ) was undertaken by Goursat [1], who showed that the relevant structure is described by the quintuplet

$$P\{\hat{H}, H; \varphi; K, \hat{K}\},$$

where  $\hat{H} \triangleleft H \subset G \supset K \triangleright \hat{K}$ , ( $H$  and  $K$  are subgroups of  $G$ , and  $\hat{H}$  and  $\hat{K}$  are normal subgroups, respectively, of  $H$  and  $K$ ), and where  $\varphi: H/\hat{H} \rightarrow K/\hat{K}$  is an isomorphism which correlates the factor groups.

The elements of  $H/\hat{H}$  and  $K/\hat{K}$  are sets of elements in  $G$ . Out of each one of these we can choose a representative  $\hat{h} \in \mathcal{H}$  and  $\hat{k} \in \mathcal{K}$ , where  $\mathcal{H}$  and  $\mathcal{K}$  are sets of elements in  $G$  which need not form a group. Different representatives  $\hat{h}, \hat{k}$  of the same element of the factor groups may be obtained through multiplication by elements of the normal groups. We subduce from the isomorphism  $\varphi$  a one-to-one mapping  $\varphi: \mathcal{H} \rightarrow \mathcal{K}$  which by abuse we denote through the same symbol as  $\varphi(\hat{h}) = \hat{k}$ . The equivalence relations (2) may then be presented in more detail as,

$$(4) \quad g' \overset{P}{\sim} g \Leftrightarrow \exists \hat{h} \in \hat{H}, \hat{h} \in \mathcal{H}, \hat{k} \in \hat{K} | g' = \hat{h} g \varphi(\hat{h})^{-1} \hat{k}.$$

The bilateral class (3) of an element  $g_i$  is then characterized as the set

$$(5) \quad B_i^P = \hat{H} \hat{h}_i \varphi(\hat{h})^{-1} \hat{K}, \quad \hat{h} \in \mathcal{H}.$$

The number of elements of  $P$  is  $|P| = |\hat{H}| |\hat{K}| |H/\hat{H}|$ . If  $P_i$  is the stability group  $P$  of a given element  $g_i$ , then the number of elements of  $B_i^P$  is  $|B_i^P| = |P|/|P_i|$ . Of course,  $|P_i|$  divides  $|P|$ , and  $|P|$  divides  $|G|^2$ .

As particular cases of the  $P$ -bilateral classes we have the following classical ones. Left cosets of  $G$  by  $H$  are determined by  $P = H \times e \simeq H$  given through  $P\{H, H; -, e, e\}$ , where  $e$  is the group identity in  $G$  and the dash indicates that the correlation function is trivial. Right cosets by  $K$  are characterized by  $P = e \times K \simeq K$  through  $P\{e, e; -, K, K\}$ , and double cosets by  $P = H \times K$  through  $P\{H, H; -, K, K\}$ ; i.e., the left and right factors in (1) are uncorrelated. Finally, conjugation subclasses of  $G$  by  $H$  correspond to  $P = (H \times H)_D \simeq H$  given though  $P\{e, H; \varphi_e; H, e\}$ , where  $\varphi_e$  is the identity isomorphism in  $H$ ; i.e., the right and left factors in (1) are totally correlated.

A particular case which is of practical importance occurs when the group  $H$  splits, i.e., when it is a semidirect product  $H = \hat{H} \wedge \check{H}$ ,  $\hat{H} \triangleleft H \supset \check{H}$ . In that case  $\mathcal{H}$  may be identified with  $\hat{H}$ , because a set of representatives exists whose elements form by

themselves a subgroup of  $G$ . When  $H = \hat{H} \wedge \dot{H}$  and  $K = \hat{K} \wedge \dot{K}$ , every element of these groups can be decomposed uniquely as  $h = \hat{h}\dot{h} = \dot{h}\hat{h}'$  and  $k = \hat{k}\dot{k} = \dot{k}\hat{k}'$ . In this case  $\varphi : \dot{H} \rightarrow \dot{K}$  is an isomorphism between subgroups of  $G$ .

Bilateral classes defined by  $P\{\hat{H}, H; \varphi; K, \hat{K}\}$ , where  $H$  and  $K$  split, will be related now to bilateral classes defined through  $P\{e, \dot{H}; \varphi; \dot{K}, e\}$ , where  $\dot{K} = \varphi(\dot{H})$ , and to double cosets defined through  $P\{\hat{H}, \dot{H}; -, \hat{K}, \dot{K}\}$ . For the former,  $P \approx \dot{H}$  and

$$(6) \quad g' \stackrel{P}{\sim} g \Leftrightarrow \exists \dot{h} \in \dot{H} | g' = \dot{h}g\varphi(\dot{h})^{-1}.$$

The set containing  $g_i$ ,

$$(7) \quad C_i^\varphi = \dot{H}g_i\varphi(\dot{H})^{-1},$$

will be called a  $\varphi$ -twisted subclass. A bilateral class  $P\{\hat{H}, H; \varphi; K, \hat{K}\}$  in this case is the union of an entire number of  $\varphi$ -twisted subclasses  $P\{e, \dot{H}; \varphi; \dot{K}, e\}$  (as in (7)) whose representatives  $g_i$  are subject to the double-coset equivalence relation defined by  $P\{\hat{H}, \dot{H}; -, \hat{K}, \dot{K}\}$ . Conversely, the same bilateral class consists of an entire number of the latter double cosets whose representatives are subject to the  $\varphi$ -twisted subclass equivalence relation (6). Note that whereas in general a bilateral class always consists of a direct union of double cosets  $P\{\hat{H}, \dot{H}; -, \hat{K}, \dot{K}\}$ , the decomposition into  $\varphi$ -twisted subclasses occurs only when  $H$  and  $K$  split.

There are two main types of  $\varphi$ -twisted subclasses  $P\{e, \dot{H}; \varphi; \dot{K}, e\}$ : those which are defined by conjugation automorphisms  $\varphi_l(\dot{h}) = l\dot{h}l^{-1}$ , and those which are not. The former may be extended from  $\dot{H}$  to the whole of  $G$ , and further classified into those which are inner to  $\dot{H}$  (i.e.,  $l \in \dot{H}$ ), and those which are outer to  $\dot{H}$  but inner to  $G$  (i.e.,  $l \notin \dot{H}$  but  $l \in G$ ). The following result allows us to recognize  $\varphi$ -twisted subclasses defined by conjugation automorphisms, and to determine all the possible extensions to  $G$ :

A  $\varphi$ -twisted subclass partition of  $G$  by a subgroup  $\dot{H}$  stems from a conjugation automorphism if and only if there exist classes consisting of a single group element. The  $\varphi$ -twist is then induced by an element  $l \in G$ , where  $l^{-1}$  is any of the one-element classes.

In order to prove this statement, consider the centralizer  $Z$  of  $\dot{H}$  in  $G$ , i.e.,  $z\dot{h} = \dot{h}z$  for all  $z \in Z, \dot{h} \in \dot{H}$ . The set  $Z$  contains at least  $e$ . Then, every element  $z \in Z$  is an (untwisted  $\varphi_e$ ) conjugation subclass of  $G$  by  $\dot{H}$ , and every  $zl^{-1}$  (for fixed  $l \in G$ ) will be a  $\varphi_l$ -twisted conjugation subclass. Conversely, let  $l^{-1} \in G$  be a single-element  $\varphi$ -twisted subclass of  $G$  by  $\dot{H}$ , i.e.,  $l^{-1} = \dot{h}l^{-1}\varphi(\dot{h})^{-1}$  for all  $\dot{h} \in \dot{H}$ ; it follows that the automorphism  $\varphi$  is given by  $\varphi(\dot{h}) = l\dot{h}l^{-1}$ . If we replace  $l$  by any  $lz, z \in Z$ , the new automorphism will coincide on  $\dot{H}$  with the original one, but its extension to  $G$  will be different for elements of  $Z$  which are not in the center of the latter.

It is also evident that if  $C_i^{\varphi_e}$  is an ordinary (i.e., untwisted) subclass, then for any conjugation automorphism  $\varphi_l$  the corresponding  $\varphi_l$ -twisted subclass will be  $C_i^{\varphi_l} = C_i^{\varphi_e}l^{-1}$ ; i.e.,  $g \stackrel{P(\varphi_l)}{\sim} g' \Leftrightarrow gl \stackrel{P(\varphi_e)}{\sim} g'l$ . The  $\varphi_l$ -twisted partition of  $G$  is thus simply a right translate by  $l^{-1}$  of the untwisted conjugation class partition.

If the automorphism  $\varphi$  defining  $P \subset G \times G$  is outer to  $G$ , the above arguments no longer hold.

**3. Representations and subgroup adaptation.** In describing harmonic analysis in this article, we shall consider only the case when  $G$  is a discrete, finite group of  $|G|$  elements. This is done in order to keep our considerations as simple as possible, without involving ourselves with the definitions of Haar and Plancherel measures. The structure of the results in this and the following section, however, will point to a straightforward generalization to compact Lie groups and, provided sufficient knowledge is available about the Plancherel measures, to locally compact Lie groups as well.

Let the Unitary Irreducible Representations (UIR's) of  $G$  be labelled by  $\gamma, \gamma \in \tilde{G}$ , and let  $D_{\rho\rho'}^\gamma(g)$  be the UIR matrix elements with row and column labels  $\rho$  and  $\rho'$ . Let the dimension of  $\mathbf{D}^\gamma$  be  $d(\gamma)$ . Then, it is known [2] that the UIR matrix elements form an orthogonal and complete set of functions over  $G$ .

We can also define skew UIR's [6] by choosing two unitary matrices  $\mathbf{U}$  and  $\mathbf{V}$ , and writing

$$(8) \quad \mathbf{\Delta}^\gamma(g) = \mathbf{U}^\dagger \mathbf{D}^\gamma(g) \mathbf{V}.$$

The representation properties then require the use of a metric  $\mathbf{V}^\dagger \mathbf{U}$  as

$$(9) \quad \mathbf{\Delta}^\gamma(g'g) = \mathbf{\Delta}^\gamma(g') \mathbf{V}^\dagger \mathbf{U} \mathbf{\Delta}^\gamma(g).$$

This set of skew UIR matrix elements has the same orthogonality and completeness relations as the ordinary UIR's:

$$(10a) \quad \sum_{g \in G} \Delta_{\rho'\mu'}^{\gamma'}(g)^* \Delta_{\rho\mu}^\gamma(g) = \delta_{\gamma',\gamma} \delta_{\rho',\rho} \delta_{\mu',\mu} |G|/d(\gamma),$$

$$(10b) \quad \sum_{\gamma \in \tilde{G}} \frac{d(\gamma)}{|G|} \sum_{\rho,\mu} \Delta_{\rho\mu}^\gamma(g')^* \Delta_{\rho\mu}^\gamma(g) = \delta_{g',g},$$

where the  $\delta$ 's are Kronecker symbols over  $\tilde{G}$ ,  $G$  and the  $d(\gamma)$ -dimensional space of rows and columns, as implied by the context. Any complex-valued function  $A(g)$  with domain on  $G$  can be thus expanded in the basis afforded by the skew UIR matrix elements as

$$(11a) \quad A(g) = \sum_{\gamma \in \tilde{G}} \frac{d(\gamma)}{|G|} \sum_{\rho,\mu} A_{\rho\mu}^\gamma \Delta_{\rho\mu}^\gamma(g)^*.$$

The generalized Fourier coefficients  $\mathbf{A}^\gamma$  are matrix-valued functions on  $\tilde{G}$  which can be determined through

$$(11b) \quad A_{\rho\mu}^\gamma = \sum_{g \in G} A(g) \Delta_{\rho\mu}^\gamma(g).$$

The unitary transformation matrices  $\mathbf{U}$  and  $\mathbf{V}$  may be chosen to symmetry-adapt the row and column labels to different chains of subgroups: we may choose  $\rho = (p, \eta, r)$  where  $\eta$  labels the UIR's of  $H \subset G$ ,  $p$  resolves the multiplicities in the subduction from  $\gamma$  to  $\eta$ , and  $r$  is some column label for  $\mathbf{D}^\eta(h)$ , the UIR's of  $H$ . Similarly, we choose  $\mu = (q, \kappa, s)$ , where  $\kappa$  labels the UIR's of  $K \subset G$ .

This sequence adaptation allows for the relation

$$(12) \quad \Delta_{p\eta r, q\kappa s}^\gamma(hgk^{-1}) = \sum_{r's'} D_{rr'}^\eta(h) \Delta_{p\eta r', q\kappa s'}^\gamma(g) D_{s's}^\kappa(k^{-1}).$$

We shall now consider the sequence adaptation to the chains of subgroups which are relevant for the decomposition into  $P$ -bilateral classes. In descending along the chains  $\hat{H} \triangleleft H \subset G$  and  $G \supset K \triangleright \hat{K}$ , we are interested in those representations of  $H$  and  $K$  which contain the trivial (unit) representation of the normal subgroups. We denote these by  $\eta_0$  and  $\kappa_0$ . These are the most general UIR's of the factor groups  $H/\hat{H}$  and  $K/\hat{K}$ . (This fact may be most familiar to the reader in the case of the Poincaré group,

where the representations containing the null momentum are labelled by the UIR's of Lorentz group.) Finally, because the elements of the factor groups  $H/\hat{H}$  and  $K/\hat{K}$  are related through the isomorphism  $\varphi$ , we may choose the row-and-column labels of their UIR's such that

$$(13) \quad D_{\bar{r}\bar{r}'}^{\eta_0(\tau)}(h_f) = D_{\bar{r}\bar{r}'}^{\kappa_0(\tau)}(\varphi(h_f)) = D_{\bar{r}\bar{r}'}^{\tau}(h_f),$$

where  $h_f \in H/\hat{H}$  and  $\tau$  labels the UIR's of the factor group.

Having made the above considerations on the UIR row and column indices following the structure of  $p\{\hat{H}, H; \varphi; K, \hat{K}\}$ , we shall now relate the space of bilateral classes to a remainder of these indices.

**4. Functions over the space of bilateral classes and partial traces.** We shall consider functions  $A$  on  $G$  which are constant over bilateral classes  $B_i$ , i.e.,  $g' \stackrel{P}{\sim} g \Rightarrow A(g') = A(g)$ , and which thus may depend only on the bilateral class to which  $g$  belongs. We will express them as  $A(g_i) = A(B_i)$ , where  $B_i \in \Gamma$  and  $\Gamma$  is the space of  $P$ -bilateral classes. The Fourier coefficients of such functions will have corresponding restrictions and independences, as we shall now see. The sum over the group  $G$  in the Fourier analysis formula (11b) can be split into a sum over the  $|B_i|$  elements  $g \in B_i$ , times a sum over  $B_i, B_i \in \Gamma$ . The former, in turn, will be expressed as sums (due to (4)) over the group elements of  $\hat{H}, \hat{K}$  and some of representatives  $\hat{h}$  in  $\mathcal{H}$ , that is,

$$(14) \quad A_{p\eta r, q\kappa s}^\gamma = \sum_{B_i \in \Gamma} A(B_i) \sum_{g \in B_i} \Delta_{p\eta r, q\kappa s}^\gamma(g).$$

We shall now calculate the last sum using (a) the decomposition (12)–(13) of the last section; (b) the orthogonality relation (10a) for each of the subgroups in question, noting that the one-dimensional trivial representation appears for  $\hat{H}$  and  $\hat{K}$ ; and (c) the fact that the stability group  $P_i$  of any one element  $g_i$  in  $B_i$ , has  $|P_i| = |P|/|B_i| = |\hat{H}||\hat{K}||H/\hat{H}|/|B_i|$  elements. As a matter of notation, we shall indicate by a bar (as  $\bar{r}$  and  $\bar{s}$ ) the row indices of the representations (as  $r$  and  $s$ ) of  $H/\hat{H} \simeq K/\hat{K}$ .

We can thus write:

$$\begin{aligned} & \sum_{g \in B_i} \Delta_{p\eta r, q\kappa s}^\gamma(g) \\ &= \frac{1}{|P_i|} \sum_{r's'} \sum_{\hat{h} \in \mathcal{H}} \sum_{\hat{h} \in \hat{H}} \sum_{\hat{k} \in \hat{K}} D_{r'r'}^\eta(\hat{h}\hat{h}) \Delta_{p\eta r', q\kappa s'}^\gamma(g_i) D_{s's}^{\kappa_0}(\varphi(\hat{h})^{-1}\hat{k}) \\ (15) \quad &= \delta_{\eta, \eta_0(\tau)} \delta_{\kappa, \kappa_0(\tau)} \frac{|\hat{H}||\hat{K}|}{|P_i|} \sum_{\bar{r}'\bar{s}'} \sum_{h_f \in H/\hat{H}} D_{\bar{r}'\bar{s}'}^\tau(h_f) \\ & \quad \cdot \Delta_{p\eta_0(\tau)\bar{r}', q\kappa_0(\tau)\bar{s}'}^\gamma(g_i) D_{\bar{s}'\bar{s}}^{\tau'}(h_f^{-1}) \\ &= \delta_{\eta, \eta_0(\tau)} \delta_{\kappa, \kappa_0(\tau)} \delta_{\tau, \tau'} \delta_{\bar{r}, \bar{s}} \frac{|B_i|}{d(\tau)} \sum_{\bar{r}} \Delta_{p\eta_0(\tau)\bar{r}, q\kappa_0(\tau)\bar{r}}^\gamma(\tau) \bar{r}(g_i). \end{aligned}$$

The first expression is thus diagonal in, and independent of, the row and column labels of the  $H$  and  $K$  UIR's, namely  $r$  and  $s$ . It depends only on the  $G$  UIR index  $\gamma$ , the  $H/\hat{H}$  UIR index  $\tau$  and the possible multiplicity indices  $p$  and  $q$ . This implies that the index dependence of the Fourier coefficients (14) will be restricted likewise to

$$(16) \quad A_{p\eta r, q\kappa s}^\gamma = \delta_{\eta, \eta_0(\tau)} \delta_{\kappa, \kappa_0(\tau)} \delta_{\tau, \tau'} \delta_{rs} A_{p\tau q}^\gamma.$$

We define the *partial traces* associated to the bilateral class partition  $P\{\hat{H}, H; \varphi; K, \hat{K}\}$  as

$$(17) \quad \begin{aligned} \chi_{p\tau q}^\gamma(B_i) &= \sum_{\bar{i}} \Delta_{p\eta_0(\tau)\bar{i}, q\kappa_0(\tau)\bar{i}}^\gamma(g_i) \\ &= \frac{d(\tau)}{|B_i|} \sum_{g \in B_i} \Delta_{p\eta_0(\tau)\bar{i}, q\kappa_0(\tau)\bar{i}}^\gamma(g). \end{aligned}$$

These will be an orthogonal and complete set of functions on the space  $\Gamma$  of  $P$ -bilateral classes since, from (10) and through steps analogous to (15) and tracing, we obtain

$$(18a) \quad \sum_{B_i \in \Gamma} \frac{|B_i|}{d(\tau)} \chi_{p'\tau'q'}^{\gamma'}(B_i)^* \chi_{p\tau q}^\gamma(B_i) = \delta_{\gamma', \gamma} \delta_{\tau', \tau} \delta_{p', p} \delta_{q', q} |G| / d(\gamma),$$

$$(18b) \quad \sum_{\gamma \in \hat{G}} \frac{d(\gamma)}{|G|} \sum_{p, \tau, q} \frac{|B_i|}{d(\tau)} \chi_{p\tau q}^\gamma(B_i)^* \chi_{p\tau q}^\gamma(B_j) = \delta_{i, j}.$$

Thus any function on  $\Gamma$  may be expanded as

$$(19a) \quad A(B_i) = \sum_{\gamma \in \hat{G}} \frac{d(\gamma)}{|G|} \sum_{p\tau q} A_{p\tau q}^\gamma \chi_{p\tau q}^\gamma(B_i)^*,$$

$$(19b) \quad A_{p\tau q}^\gamma = \sum_{B_i \in \Gamma} \frac{|B_i|}{d(\tau)} A(B_i) \chi_{p\tau q}^\gamma(B_i).$$

Since the bilateral class partition generalizes the coset and conjugation subclass partitions the partial traces (17) will generalize harmonic functions and characters over the group. Thus, for left cosets  $P\{H, H; -, e, e\}$ , the sum over the row indices  $\bar{i}$  of  $H/H = \{e\}$  disappears. The partial traces become the harmonic functions  $D_{p0,q}^\gamma(B_i)$  over the manifold of left cosets, where the row index is specified by an appropriate multiplicity label for the subduction  $G \supset H$  which contains the trivial representation of  $H$ . The column index  $q$  is fully determined by a complete subgroup chain of  $G$ . Similarly, right cosets  $P\{e, e; -, K, K\}$  lead to “spherical harmonics”  $D_{p,q0}^\gamma(B_i)$ , and double cosets  $P\{H, H; -, K, K\}$  to “diamond Wigner  $d$ -functions”  $D_{p0,q0}^\gamma(B_i)$ . For ordinary conjugation classes  $P\{e, G; \varphi_e; G, e\}$  we have the usual characters  $\chi^g(B_i) = \sum_p D_{pp}^\gamma(B_i)$ , while for conjugation subclasses  $P\{e, H; \varphi_e; H, e\}$ , the  $\chi_{p\tau q}^\gamma(B_i)$  are as given by (17) for the row and column labels referring to the same subgroup chain.

For the case of split subgroups and  $\varphi$ -twisted subclasses defined through conjugation automorphisms  $\varphi_l$  as  $P\{e, \hat{H}; \varphi_l; \hat{K}, e\}$ ,  $\hat{K} = l\hat{H}l^{-1}$ . The group subgroup chain for the column indices of the UIR matrices is thus obtained from the row indices through a transformation by  $l$  as

$$(20) \quad \Delta_{\rho\sigma}^\gamma(g) = \sum_\tau D_{\rho\tau}^\gamma(g) D_{\tau\sigma}^\gamma(l),$$

where by  $\mathbf{D}^\gamma(\cdot)$  we indicate UIR matrices whose rows and columns are classified by the same subgroup chain. The values of the partial traces for a  $\varphi$ -twist will thus be equal to the ordinary partial traces mentioned above, but valued at the class  $C_i^\varphi = C_i^\varphi l^{-1}$ . The Fourier coefficients of functions constant on  $C_i^\varphi$  can be referred to the same subgroup chain, in place of (16)–(19b), as

$$(21a) \quad {}^e A_{p\tau\tau', q\tau's}^\gamma = \sum_{p'} {}^e A_{p'\tau p'}^\gamma D_{p'\tau\tau', q\tau's}^\gamma(l^{-1}),$$

$$(21b) \quad {}^e A_{p\tau p'}^\gamma = \sum_{C_i^\varphi \in \Gamma} \frac{|C_i^\varphi|}{d(\tau)} A(C_i^\varphi) {}^e \chi_{p\tau p'}^\gamma(C_i^\varphi l).$$

In the above,  $\tau$  refers to the UIR label of  $H$ . If we now let  $A(C_i^\varphi)$  be constant over double cosets by  $\hat{H}$  and  $\hat{K}$ , i.e.,  $A(C_i^\varphi) = A(\hat{h}C_i^\varphi\hat{k})$  the multiplicity indices  $p$  and  $q$  must be replaced by  $(p, \eta)$  and  $(q, \kappa)$ , while constancy over these sets now imposes  $\delta_{\eta, \eta_0}$  and  $\delta_{\kappa, \kappa_0}$  factors, reducing the number of independent Fourier coefficients in (21) to those containing the trivial representations of  $\hat{H}$  and  $\hat{K}$ .

**5. Conclusion.** We have extended the classical results of harmonic analysis on cosets and conjugation classes to bilateral classes. This unifies their treatment as well as that of a number of other special cases. From a practical point of view this result may also be quite useful, e.g., if we keep in mind the original applications [3]. Assume we wish to describe functions valued over the different transitions; these can certainly be expanded and analyzed with respect to their harmonic components. Such a procedure has proved useful in many applications and the possibility of performing it in this new situation seems relevant.

#### REFERENCES

- [1] E. GOURSAT, *Annales Scientifiques de l'Ecole Normale Supérieure*, Paris (3) 6 (1889), p. 9; see also W. B. Scott, *Group Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1964, p. 71.
- [2] M. HAMERMESH, *Group Theory*, Addison Wesley, Reading, MA, 1962.
- [3] W. HÄSSELBARTH AND E. RUCH, *Classifications of rearrangement mechanisms by means of double cosets and counting formulas for the numbers of classes*, *Theor. Chim. Acta* 29 (1973), pp. 259–267.
- [4] W. HÄSSELBARTH, E. RUCH, D. J. KLEIN AND T. H. SELIGMAN, *Double classes: a new classification scheme for group elements*, *Proceedings of the Colloquium on Group Theory and its Applications*, Montreal, 1976, Academic Press, New York, 1977, pp. 617–622.
- [5] W. HÄSSELBARTH, E. RUCH, D. J. KLEIN AND T. H. SELIGMAN, *Bilateral classes*, *J. Math. Phys.*, 21 (1980), pp. 951–953.
- [6] D. J. KLEIN, *Finite groups and semisimple algebras in quantum mechanics*, in *Group Theory and its Applications*, Vol. III, E. M. Loebl, ed., Academic Press, New York, 1975.
- [7] T. H. SELIGMAN AND K. B. WOLF, *Harmonic analysis on double classes*, *Proceedings of the Colloquium on Group Theory and its Applications*, Montreal, 1976, Academic Press, New York, 1977, pp. 591–598.

## INVARIANT SETS FOR NONLINEAR ELLIPTIC AND PARABOLIC SYSTEMS\*

HENDRIK J. KUIPER†

**Abstract.** In this paper we consider systems of weakly coupled nonlinear second order elliptic and parabolic equations with nonlinear, possibly coupled, boundary conditions. The aim is to find invariant sets of the form

$$S = \{(u_1, u_2, \dots, u_m) \mid \varphi_i(x) \leq u_i(x) \leq \psi_i(x) \text{ a.e.}\}$$

for certain nonlinear reaction-diffusion equations,

$$U_t + LU = F(U) \quad \text{in } \Omega \times [0, \infty),$$

$$BU = G(U) \quad \text{on } \partial\Omega \times [0, \infty),$$

where  $L = (L_1, L_2, \dots, L_m)$ , ( $L_i$  a linear second order elliptic operator),  $B = (B_1, B_2, \dots, B_m)$ , ( $B_i$  a linear boundary operator of a general type), and  $U = (u_1, u_2, \dots, u_m)$ . One of the main results says in essence that  $S = \{U \mid \Phi \leq U \leq \Psi\}$  is an invariant set if

$$L\Phi \leq F(\Phi) \text{ and } L\Psi \geq F(\Psi) \text{ in } \Omega \times [0, \infty),$$

and

$$B\Phi \leq G(\Phi) \text{ and } B\Psi \geq G(\Psi) \text{ on } \partial\Omega \times [0, \infty).$$

The work also includes some existence results for the parabolic problem and the associated nonlinear elliptic problem.

### 1. Introduction. Consider the reaction-diffusion equations

$$(DE) \quad \frac{\partial u_k}{\partial t} + L_k u_k = f_k(x, t, U), \quad 1 \leq k \leq m,$$

where  $L_1, L_2, \dots, L_m$  are second order elliptic partial differential operators on a bounded open set  $\Omega \subset R^n$ , together with the conditions

$$(BC) \quad B_k u_k = g_k(x, U), \quad 1 \leq k \leq m,$$

imposed on  $U(t) = (u_1(\cdot, t), u_2(\cdot, t), \dots, u_m(\cdot, t))$  at the boundary. For  $t \in [0, T)$  we may think of  $U(t)$  as belonging to some Banach space  $\mathcal{X}$  of real valued functions from  $\Omega$  into  $R^m$ . Let  $K \subset [0, T) \times \mathcal{X}$  be a set whose sections  $K(t)$  are closed convex sets in  $\mathcal{X}$ .  $K$  is then called an invariant set for the problem (DE)–(BC) if  $U(t_0) \in K(t_0)$  implies that the solution  $U(t) \in K(t)$  for all  $t \in (t_0, T)$ . Reaction-diffusion equations have lately received a great deal of attention. Their interest lies partially in the fact that they occur in the mathematical models for a wide range of natural processes (see e.g. [4], [5], [6], [24], and the references given in those papers). In particular there has been interest in the existence of invariant sets. Usually some restrictions are put on the form of  $K$ . For example, Weinberger [23] considered the case where  $K(t)$  is independent of  $t$  and consists of functions which take on their values in some closed convex subset  $C \subset R^m$ . Unless the elliptic operators  $L_i$  are the same for all  $i$ , more restrictions have to be placed on  $C$  ([1], [4]), such as requiring that  $C = [\alpha_1, \beta_1] \times [\alpha_2, \beta_2] \times \dots \times [\alpha_m, \beta_m]$ . The present author [13] obtained results for invariant sets of the form

$$K(t) = \{(u_1, u_2, \dots, u_m) \mid \varphi_i(x, t) \leq u_i(x, t) \leq \psi_i(x, t) \forall x \in \Omega\}.$$

\* Received by the editors April 26, 1979, and in final revised form February 8, 1980. This work was sponsored by the United States Army under Contract No. DAAG29-75-C-0024, and supported in part by an Arizona State University Faculty Grant-in-Aid.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85281.



Such results might be more properly referred to as ‘‘comparison theorems’’. In 1964 Walter obtained such comparison theorems for classical solutions to general nonlinear boundary value problems, thus extending some work of Mlak and Szarski dating back to the 1950s [23; § 32].

In this paper we consider generalized rather than classical solutions. For the semilinear equation we obtain results which are analogous to some of those obtained by Walter. We also obtain some existence results for elliptic as well as parabolic problems.

In order to handle the nonlinear boundary conditions we use the nonlinear semi-group theory of Crandall, Liggett and Pazy, which seems to be particularly well-suited. This approach reduces the problem to one of studying invariant sets for associated elliptic problems. The solutions of the elliptic problem which we consider will be distributional solutions in  $(H^1(\Omega))^m$ . The solution of the reaction-diffusion equations which we will look at will also be of a weak type. This then means that the results on invariant sets which we obtain will also be valid for solutions of stronger type such as classical solutions or solutions  $U \in C^1((0, T), (L^2(\Omega))^m) \cap C^0([0, T], D)$  where  $D \subset (H^1(\Omega))^m$  is the domain of  $(L_1, L_2, \dots, L_m)$ .

Although the main theorems (9), (10), (11), (16) can be read with only the aid of a few well marked definitions (the hypotheses being explicitly stated), we feel it might be helpful to the reader if we state a somewhat simplified version of the invariant set theorem for the parabolic problem, and give one simple application.

Let  $L_k (k = 1, 2, \dots, m)$  be uniformly strongly elliptic with coefficients in  $C^1(\bar{\Omega})$ . Let the functions  $f_i$  and  $g_i$  be of class  $C^1(\bar{\Omega} \times R^1 \times R^m)$ , and assume the boundary conditions are of the form

$$B_k u_k \equiv \beta_k \cdot \nabla u_k + \gamma_k u_k = g_k(x, u_1, \dots, u_m) \quad \text{on } \partial\Omega,$$

where  $\beta_k$  is a nowhere vanishing  $C^1$  vector field on  $\partial\Omega$  (which is assumed to be of class  $C^2$ ), and  $0 \leq \gamma_k \in C(\bar{\Omega})$ . Alternatively, the boundary condition may be of the Dirichlet type,

$$B_k u_k \equiv u_k(x) = g_k(x, u_1, \dots, u_m) \equiv \theta_k(x) \quad \text{on } \partial\Omega.$$

Let  $\varphi_i$  and  $\psi_i (1 \leq i \leq m)$  be  $C^1(\bar{\Omega}) \cap C^2(\Omega)$  functions which satisfy, for all  $u_i \in C^1(\bar{\Omega})$  with  $\varphi_i \leq u_i \leq \psi_i$ ,

$$\begin{aligned} L_i \varphi_i &\leq f_i(x, t, u_1, u_2, \dots, u_{i-1}, \varphi_i, u_{i+1}, \dots, u_m) \quad \text{in } \Omega, \\ L_i \psi_i &\geq f_i(x, t, u_1, u_2, \dots, u_{i-1}, \psi_i, u_{i+1}, \dots, u_m) \quad \text{in } \Omega, \\ B_i \varphi_i &\leq g_i(x, t, u_1, u_2, \dots, u_{i-1}, \varphi_i, u_{i+1}, \dots, u_m) \quad \text{on } \partial\Omega, \\ B_i \psi_i &\geq g_i(x, t, u_1, u_2, \dots, u_{i-1}, \psi_i, u_{i+1}, \dots, u_m) \quad \text{on } \partial\Omega. \end{aligned} \tag{A_0}$$

Then  $\{(u_1, u_2, \dots, u_m) \mid \varphi_i(x) \leq u_i(x) \leq \psi_i(x) \ \forall x \in \Omega, 1 \leq i \leq m\}$  is an invariant set for the problem (DE)–(BC).

*Application.* Let us consider a system of equations which arise in the theory of combustion (cf. [4], [11]). For simplicity we restrict ourselves to the one-dimensional case,

$$\begin{aligned} n_t - k_1 n_{xx} &= -ne^{-E/RT}, \\ T_t - k_2 T_{xx} &= Qne^{-E/RT}, \end{aligned}$$

where  $T$  and  $n$  denote the temperature and concentration of the fuel, and where  $E, R, Q, k_1$ , and  $k_2$  are constants (the calculations below can still be carried out, however

if, for example,  $k_1$  and  $k_2$  depend on  $x$ ). We assume that the region of interest is  $x \in [0, L]$ , and that fuel is fed in at the right end and heat is lost at the left end: we impose the boundary conditions

$$\begin{aligned} n_x(0) &= 0, & n_x(L) &= g(n), \\ T_x(0) &= \beta T^\gamma, & T_x(L) &= 0, \end{aligned}$$

where  $g(0) = 0$  and  $g(z) \geq 0$  whenever  $z \geq \alpha_0$ , and where  $\alpha_0 > 0$  is some constant. Suppose that we are given some initial conditions and that  $T_0 = \max T(x, 0)$ ,  $n_0 = \max n(x, 0)$ , and  $\alpha = \max(\alpha_0, n_0, k_2\beta T_0^\gamma/Q L)$ . We then claim that the set

$$\mathcal{F} = \{(n, T) \mid 0 \leq n \leq n_0, 0 \leq T \leq \psi\}$$

is invariant if we choose

$$\psi(x) = (\tau/\beta)^{1/\gamma} + \tau x - \tau x^2/2L,$$

where  $\tau = Q\alpha L/k_2$ . Let us verify that  $\psi$  satisfies the right inequalities (see  $(A_0)$ ), the inequalities for the other functions used in the definition of  $\mathcal{F}$  being trivially satisfied.

$$\begin{aligned} -k_2\psi''(x) &= k_2\tau/L = Q\alpha \geq Qn_0 \geq Qne^{-EIR\psi}, \\ \frac{\partial\psi(0)}{\partial\nu} &= -\psi'(0) = -\tau = -\beta\psi(0)^\gamma, \\ \frac{\partial\psi(L)}{\partial\nu} &= \psi'(L) = \tau - \tau \geq 0. \end{aligned}$$

Also we note that  $\psi(x) \geq \psi(0) = (\tau/\beta)^{1/\gamma} \geq T_0$ , and therefore we know that

$$(n(\cdot, t), T(\cdot, t)) \in \mathcal{F} \quad \text{for all } t \geq 0.$$

**2. The linear elliptic problem.** Let  $L_k$ ,  $1 \leq k \leq m$ , be linear second order uniformly elliptic operators with real coefficients acting on real valued functions of  $x = (x_1, x_2, \dots, x_n)$  in a bounded open set  $\Omega \subset R^n$ ,

$$L_k u \equiv -D_i[a_k^{ij}(x)D_j u + d_k^i(x)u] + b_k^i(x)D_i u + c_k(x)u,$$

where summation is, and subsequently will be, carried out over any index which occurs both as a subscript and as a superscript within the same term. Next let  $B_k$ ,  $1 \leq k \leq m$ , be first order boundary operators, of transversal order 1, acting on real valued functions defined on some subset  $\Delta_k$  of the boundary  $\partial\Omega$ . In this section we shall look at the weakly coupled linear system

$$(1) \quad (L_k + \lambda)u_k(x) - h_k^j(x)u_j(x) = f_k(x), \quad (x \in \Omega),$$

with boundary conditions

$$(2) \quad B_k u_k(x) - e_k^j(x)u_j(x) = g_k(x), \quad (x \in \Delta_k),$$

$$(3) \quad u_k(x) = \theta_k(x), \quad (x \in \Gamma_k \equiv \partial\Omega \setminus \Delta_k),$$

for all  $1 \leq k \leq m$ . We will look at this problem from a variational point of view, and hence it will be necessary for us to write the operator  $B_k$  in the form

$$B_k u = \nu_i[a_k^{ij}(x)D_j u + d_k^i u] + \sigma_k(x)u + t_k^i(x)D_i u,$$

where  $\nu = (\nu_1, \nu_2, \dots, \nu_m)$  is the unit outward normal on  $\partial\Omega$ , and  $t_k = (t_k^1, t_k^2, \dots, t_k^n)$  is a tangential vector field on  $\partial\Omega : \nu_i(x)t_k^i(x) \equiv 0$  on  $\partial\Omega$  for all  $1 \leq k \leq m$ .

We will use  $(\cdot, \cdot)_Y$  to denote the usual  $L^2(Y)$  inner product. When we take a

direct sum of  $m$  copies of  $L^2(Y)$  we shall still use the same symbol for the inner product on this direct sum; i.e., if  $F = (f_1, f_2, \dots, f_m)$  and  $G = (g_1, g_2, \dots, g_m)$  are members of  $\oplus_{i=1}^m L^2(Y)$ , then  $(F, G)_Y = \sum_{i=1}^m (f_i, g_i)_Y$ . The norm is denoted by  $\|\cdot\|_{0,Y}$ .

If  $Y = \Omega$  we shall delete the subscript  $\Omega$ . Hence  $\|\cdot\|_0$  denotes the  $L^2(\Omega)$ -norm,  $(\cdot, \cdot)$  the  $L^2(\Omega)$  inner product (or the  $(L^2(\Omega))^m$ -norm and inner product respectively). The norm on the Sobolev space  $W^{m,p}(\Omega)$  (derivatives of order  $\leq m$  are in  $L^p(\Omega)$ ) is denoted by  $\|\cdot\|_{m,p}$ . If  $p = 2$  we also use  $\|\cdot\|_m$  to denote the norm on  $H^m(\Omega) \equiv W^{m,2}(\Omega)$ . Corresponding script letters will be used to denote  $m$ -fold direct sums of function spaces e.g.,  $\mathcal{H}^1(\Omega) = \oplus_{i=1}^m H^1(\Omega)$ ,  $\mathcal{C}^1(\bar{\Omega}) = \oplus_{i=1}^m C^1(\bar{\Omega})$ , etc.

A formal integration by parts of the expression

$$\int_{\Omega} \sum_{k=1}^{\infty} [(L_k + \gamma)u_k]v_k dx,$$

with  $u_k$ 's which satisfy (1)–(3) and  $v_k$ 's which vanish on  $\Gamma_k$ , leads to the equation

$$\begin{aligned} A_{\gamma}(U, V) &\equiv \sum_{k=1}^m \{(a_k^{ij}D_j u_k, D_i v_k) + (d_k^i u_k, D_i v_k) \\ &\quad + (b_k^i D_i u_k, v_k) + ((c_k + \lambda)u_k, v_k) - (h_k^i u_j, v_k) \\ &\quad + (\sigma_k u_k, v_k)_{\Delta_k} - (e_k^i u_j, v_k)_{\Delta_k} + (t_k^i D_i u_k, v_k)_{\Delta_k}\} \\ &= \sum_{k=1}^m \{(f_k, v_k) + (g_k, v_k)_{\Delta_k}\}, \end{aligned}$$

where  $U = (u_1, u_2, \dots, u_m)$  and  $V = (v_1, v_2, \dots, v_m)$ . If the coefficients of  $L_k$  and  $B_k$  are sufficiently well-behaved, then the bilinear form  $A_{\lambda}$  is certainly well defined on  $\mathcal{C}^1(\bar{\Omega}) \times \mathcal{C}^1(\bar{\Omega})$ . We will impose conditions which will allow  $A_{\lambda}$  to be extended to a continuous  $\mathcal{U}$ -coercive form on  $\mathcal{U} \times \mathcal{U}$  for some subspace  $\mathcal{U}$  of  $\mathcal{H}^1(\Omega)$ .

(I).  $\Omega$  is a bounded open set in  $R^n$  whose boundary is of class  $C^2$ .

This condition can be weakened to requiring that  $\partial\Omega$  be Lipschitz continuous in a sense defined for example by Nečas [20]. However it seems this would require us to handle the tangential derivatives in  $B_k$  rather than, as we shall be able to do, remove them from consideration by treating another but equivalent problem. The nature of the work involved is then such that one might as well consider very general boundary operators, namely those which map  $H^{1/2}(\partial\Omega)$  into  $H^{-1/2}(\partial\Omega)$  (see e.g., [2]).

Let  $\mathcal{D}(\Omega)$  be the  $C^{\infty}(\Omega)$  functions of compact support, and  $\mathcal{D}(\Omega)'$  the Schwartz distributions. The nonnegatively valued functions in  $\mathcal{D}(\Omega)$ , denoted by  $\mathcal{D}_+(\Omega)$ , form a cone in  $\mathcal{D}(\Omega)$ . Let  $\mathcal{D}_+(\Omega)'$  be the dual cone:  $f \in \mathcal{D}_+(\Omega)'$  iff  $f(\phi) \geq 0$  for all  $\phi \in \mathcal{D}_+(\Omega)$ . Consequently, we have a partial order  $\geq$  on  $\mathcal{D}(\Omega)'$ :  $f \geq g$  iff  $f - g \in \mathcal{D}_+(\Omega)'$ . This partial order extends the usual partial order on  $L^1(\Omega)$ -functions:  $f \geq g$  iff  $f(x) \geq g(x)$  a.e. Furthermore we can extend such partial orders to vectors and matrices by saying  $F \geq G$  if the relationship is satisfied componentwise.

We use  $\gamma_0$  to denote the 0th order trace map, i.e., the extension of the map  $u \rightarrow u|_{\partial\Omega}$  from  $C^1(\bar{\Omega})$  into  $C^1(\partial\Omega)$  to a continuous map from  $W^{1,p}(\Omega)$  onto  $W^{1-1/p,p}(\partial\Omega) \subset L^p(\partial\Omega)$  for  $p > 1$  ([1], [16], [17] or [18]).

We shall, on occasion, refer to the various Sobolev-Kondrasov embedding results. We mention the following [15, p. 43]:

$$W^{r,p}(\Omega) \subset L^s(\Omega) \quad \text{if } \frac{1}{s} \geq \frac{1}{p} - \frac{r}{n}, \quad pr < n, \quad s \geq 1.$$

$$W^{r,p}(\Omega) \subset C^\alpha(\bar{\Omega}) \quad \text{if } pr > n, \alpha < 1, a \leq \frac{pr - n}{p}.$$

The second embedding is a compact linear map, as will be the first embedding provided the first inequality is strict.

We also need

$$\begin{aligned} \text{(II). } & a_k^{ij} \in L^\infty(\Omega), \quad d_k^i \in L^q(\Omega), \quad b_k^i \in L^q(\Omega), \quad c_k \in L^{q/2}(\Omega), \\ & D_i d_k^i \in L^{q/2}(\Omega), \quad 0 \leq h_k^i \in L^{q/2}(\Omega), \quad 0 \leq e_k^i \in L^p(\partial\Omega), \\ & 0 \leq \sigma_k \in L^p(\partial\Omega), \quad t_k^i \in W^{1,q}(\Omega), \end{aligned}$$

with  $\text{supp } \gamma_0 t_k^i \subset \Delta_k$ , where  $p > n - 1, p > 1, q > n$ , and  $q \geq 2$ . Also  $v_i d_k^i \geq 0$  on  $\partial\Omega^1$ , and there exists a constant  $\mu_1$  such that  $c_k - D_i d_k^i \geq -\mu_1$ .

The above hypotheses are directly related to the Sobolev embedding theorems. We also require the operators  $L_k$  to be uniformly elliptic:

(III). There exists a positive constant  $\nu_0$  such that for every  $1 \leq k \leq m$  and all  $\xi \in R^n$  we have

$$a_k^{ij}(x) \xi_i \xi_j \geq \nu_0 \sum_{i=1}^m \xi_i^2.$$

This condition can be weakened in order to treat certain degenerate-elliptic problems by methods described in [19].

(IV).  $\Delta_k$  is an open subset of  $\partial\Omega$  such that the  $(n - 1)$ -dimensional Lebesgue measure of its boundary in  $\partial\Omega$  is zero.

Let  $W_0^{k,p}(\Omega)$  (respectively,  $H_0^k(\Omega)$ ) be the subspace of  $W^{k,p}(\Omega)$  (respectively,  $H^k(\Omega)$ ) obtained by taking the closure of  $\mathcal{D}(\Omega)$ . The dual space of  $W_0^{k,p}(\Omega)$  may be represented by  $W^{-k,p^*}(\Omega)$ , the collection of all Schwartz distributions of the form  $D_i \psi^i + \psi$  with  $\psi_i$  and  $\psi$  in  $L^{p^*}(\Omega)$ , where  $1/p^* + 1/p = 1$ .

Before we proceed it should be noted that the usual Green's formula

$$(v, D_k w) = -(D_k v, w) + (v_k \gamma_0 v, \gamma_0 w)_{\partial\Omega},$$

which holds for  $v, w \in H^1(\Omega)$ , should be interpreted in the appropriate sense when  $n = 1$ . Although the results in this paper apply as well to the one-dimensional case, we shall not take the trouble here to point out the various obvious notational modifications which need to be made.

For  $S \subset \partial G$ , let  $H_s^1(G)$  be the closure in  $\bar{H}^1(G)$  of

$$\{u \in H^1(G) \mid u(x) = 0 \text{ a.e. on an open neighborhood of } \partial G \setminus S\}.$$

With this notation  $H_{\partial G}^1(G) = H_0^1(G)$ . If  $\partial G$  is sufficiently regular, it can be shown (e.g., [10]) that this space also is equal to  $\{u \in H^1(\Omega) \mid \gamma_0 u \equiv 0\}$ . We shall use  $\mathcal{H}_\Delta^1(\Omega)$  to denote  $\bigoplus_{k=1}^m H_{\Delta_k}^1(\Omega)$ .

Our first objective will be to simplify our problem somewhat. Consider the bilinear functional  $A_\lambda$ . Using the Sobolev inequalities one easily shows that the first 5 terms are continuous on  $\mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega)$  (see e.g., [15]). Using the fact that if  $u \in H^1(\Omega)$ , then  $\gamma_0 u \in H^{1/2}(\partial\Omega) \subset L^{(2n-2)/n-2}(\partial\Omega)$  (see e.g., [16] for the embedding theorem for fractional Sobolev spaces), we also easily verify that the next two terms in the expression for  $A_\lambda(U, V)$  are continuous bilinear functionals on  $\mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega)$ . The last term is also continuous. However we can use the following result of Fiorenza [9] to remove it from consideration.

<sup>1</sup> i.e.  $\int_\Omega (d_k^i D_i \phi + \phi D_i d_k^i) dx \geq 0$  whenever  $0 \leq \phi \in H^1(\Omega)$ .

**THEOREM 1.** *Suppose  $\Omega$  is a bounded open set in  $R^n$ ,  $n \geq 3$ , whose boundary is of class  $C^2$ , and suppose  $t_k^i \in W^{1,q}(\Omega)$ . Then there exist functions  $\alpha_k^{ij} (= -\alpha_k^{ji}) \in L^\infty(\Omega)$  and functions  $\gamma_k^i \in L^q(\Omega)$ , such that for all  $u, v \in H^1(\Omega)$  we have*

$$(t_k^i D_i u, v)_{\partial\Omega} = (\alpha_k^{ij} D_j u, D_i v) + (\gamma_k^i D_j u, v).$$

Although in the proof Fiorenza assumes  $t_k^i \in L^\infty(\Omega) \cap W^{1,n}(\Omega) (\supset W^{1,q}(\Omega))$ , and hence gets the  $\gamma_k^i$ 's in  $L^n(\Omega)$ , an examination of the proof easily reveals that assuming our slightly more restrictive condition  $t_k^i \in W^{1,q}(\Omega)$  does yield  $\gamma_k^i \in L^q(\Omega)$ . The proof for the case  $n = 2$  is especially easy: let  $s$  denote the distance along  $\partial\Omega$ , measured in such a way that when moving along the boundary in the direction of increasing  $s$ ,  $\Omega$  lies to the left of  $\partial\Omega$ . Let  $\vec{t}$  be the unit tangent vector in the direction of increasing  $s$ , and let  $\beta_k = \vec{t}_k \cdot \vec{t}$  ( $\vec{t}_k = (t_k^1, t_k^2)$ ). The last term in the expression for  $A_\lambda(U, V)$  takes the form

$$\begin{aligned} \sum_{k=1}^m \int_{\partial\Omega} \beta_k v_k (\nabla u_k \cdot \vec{t}) \, ds &= \sum_{k=1}^m \int \int_{\Omega} \nabla(\beta_k v_k) \times \nabla u_k \, dx \, dy \\ &= \sum_{k=1}^m \int \int_{\Omega} [\beta_k (\nabla v_k \times \nabla u_k) + v_k (\nabla \beta_k \times \nabla u_k)] \, dx \, dy \\ &= \sum_{k=1}^m \int \int_{\Omega} [\alpha_k^{ij} (D_i u_k)(D_j v_k) + v_k \gamma_k^i D_i u_k] \, dx \, dy. \end{aligned}$$

Of course this proof requires that we extend  $t$  to a  $W^{1,q}(\Omega)$  vector field. (Note that the product of two members of  $W^{1,q}(\Omega)$  is again in  $W^{1,q}(\Omega)$ ). We know however that  $t_k^i \in C^1(\partial\Omega) \subset W^{1-1/q,q}(\partial\Omega)$ , and hence the extension is possible by the trace theorem [16].

Using this theorem we can remove the last term for  $A_\lambda(U, V)$ , and replace  $a_k^{ij}$  by  $a_k^{ij} + \alpha_k^{ij}$  and  $b_k^i$  by  $b_k^i + \gamma_k^i$ . These new coefficients satisfy exactly the same hypotheses as the unaltered ones. Even the ellipticity constant  $\nu_0$  is preserved. Without loss of generality we shall from now on assume  $t_k^i \equiv 0$  for all  $k$  and  $i$ .

**LEMMA 2.** *For any  $\varepsilon > 0$  there exists a constant  $C(\varepsilon)$  such that for all  $u \in H^1(\Omega)$*

$$(i) \quad \int_{\partial\Omega} u^2 dS \leq \varepsilon \int_{\Omega} |Du|^2 dx + C(\varepsilon) \int_{\Omega} u^2 dx.$$

Moreover there exist constants  $\varepsilon_0 > 0$  and  $\mu_0 > 0$ , independent of  $\lambda$ , such that

$$(ii) \quad A_\lambda(U, U) \geq \varepsilon_0 \|U\|_1^2 + (\lambda - \mu_0) \|U\|_0^2,$$

i.e.,  $A_\lambda$  is  $\mathcal{H}_\Delta(\Omega)$ -coercive.

*Proof.* We shall use the following result due to Lions [20]: If  $X_a \subset X_b \subset X_c$  are Banach spaces with norms  $\|\cdot\|_a, \|\cdot\|_b, \|\cdot\|_c$  respectively, and if the first inclusion is compact linear and the second continuous linear, then for each  $\varepsilon > 0$  there exists a constant  $C(\varepsilon) > 0$  such that

$$\|x\|_b \leq \varepsilon \|x\|_a + C(\varepsilon) \|x\|_c \quad \forall x \in X_a.$$

We now close  $H^1(\Omega) \subset L^2(\Omega)$  with respect to the norm

$$\|u\|^2 = \int_{\partial\Omega} u^2 dS + \int_{\Omega} u^2 dx,$$

and call this space  $H$ . Now we merely apply Lions' result to  $H^1(\Omega) \subset H \subset L^2(\Omega)$ . Of

course this proof can also be accomplished by the standard partition of unity argument. For the proof of (ii) we note that the first term of  $A_\lambda(U, U)$  satisfies

$$\sum_{k=1}^m \int_{\Omega} a_k^{ij}(D_j u_k)(D_i u_k) dx \cong \nu_0 \sum_{k=1}^m \sum_{i=1}^m \|D_i u_k\|_0^2.$$

Hence it suffices to show that each of the other terms is dominated, in absolute value, by a quantity of the form

$$\varepsilon \|U\|_1^2 + C(\varepsilon) \|U\|_0^2,$$

where the  $\varepsilon > 0$  can be chosen arbitrarily small. This is easily seen to be the case. For example

$$\begin{aligned} |(d_k^i u_k, D_i u_k)| &\leq \|\varepsilon D_i u_k\|_0^2 + \left\| \frac{1}{\varepsilon} d_k^i u_k \right\|_0^2 \\ &\leq \varepsilon^2 \|\nabla U\|_0^2 + \varepsilon^{-2} \|d_k^i\|_{0,q}^2 \|u_k\|_{0,2q/(q-2)}^2. \end{aligned}$$

By the Sobolev-Kondrasov embedding theorem the embedding  $H^1(\Omega) \subset L^{2q/(q-2)}(\Omega)$  is compact continuous. Hence we can again use Lions' results to deduce that the above quantity is

$$\leq \varepsilon^2 \|\nabla U\|_0^2 + \varepsilon^{-2} \|d_k^i\|_{0,q} \{ \varepsilon^4 \|\nabla u_k\|^2 + \tilde{C}(\varepsilon) \|u_k\|^2 \} < \varepsilon \|\nabla U\|_0^2 + C(\varepsilon) \|U\|_0^2,$$

provided  $\varepsilon$  is chosen sufficiently small. As another example let us take one of the integrals over  $\Delta_k$ ,

$$\left| \int_{\Delta_k} e_k^i u_j u_k dS \right| \leq \|e_k^i\|_{0,p,\Delta_k} \|u_j\|_{0,r,\partial\Omega} \|u_k\|_{0,r,\partial\Omega},$$

where  $p > n - 1, r = 2p/(p - 1)$ . This in turn is

$$\leq \text{cst} \|U\|_{0,r,\partial\Omega}^2.$$

We again apply Lions' result to  $\mathcal{H}^1(\Omega) \subset \mathcal{X} \subset \mathcal{L}^2(\Omega)$  where  $\mathcal{X}$  is the closure of  $\mathcal{H}^1(\Omega)$  with respect to the norm

$$\|U\| = \sum_{i=1}^m \|\gamma_0 u_i\|_{0,r,\partial\Omega} + \|U\|_0.$$

Since  $\mathcal{H}^1(\Omega) \subset \mathcal{L}^2(\Omega)$  is a compact embedding, and since  $H^1(\Omega) \rightarrow H^{1/2}(\partial\Omega) \subset L^r(\partial\Omega)$  is a composition of a continuous linear map  $\gamma_0$  and a compact embedding (since  $1/r > \frac{1}{2} - \frac{1}{2}(n - 1)$ , we can apply the Sobolev-Kondrasov results) we may conclude that

$$\left| \int_{\Delta_k} c_k^i u_j u_k dS \right| \leq \text{cst} \|U\| \leq \varepsilon \|U\|_1 + C(\varepsilon) \|U\|_0. \quad \square$$

At this point it will be convenient to introduce some abbreviated notation. If  $U = (u_1, u_2, \dots, u_m) \in \mathcal{H}^1(\Omega)$ , then  $HU = (h_1^i u_i, h_2^i u_i, \dots, h_m^i u_i)$ , and  $EU = (e_1^i u_i, e_2^i u_i, \dots, e_m^i u_i)$ . We also set  $F = (f_1, f_2, \dots, f_m), G = (g_1, g_2, \dots, g_m), \Theta = (\theta_1, \theta_2, \dots, \theta_m), \Delta = \Delta_1 \times \Delta_2 \times \dots \times \Delta_m$ , and  $\Gamma = \Gamma_1 \times \Gamma_2 \times \dots \times \Gamma_m, L$  and  $B$  will respectively denote the operators  $(L_1, L_2, \dots, L_m)$  and  $(B_1, B_2, \dots, B_m)$ . With this notation, (1)–(3) can be written as

$$(4) \quad (L + \lambda - H)U = F \quad \text{in } \Omega,$$

$$(5) \quad (B - E)U = G \quad \text{on } \Delta,$$

$$(6) \quad U = \Theta \quad \text{on } \Gamma.$$

DEFINITION. For  $F \in \mathcal{H}^1(\Omega)'$  (the dual space of  $\mathcal{H}^1(\Omega)$ ) and  $G \in \mathcal{H}^{-1/2}(\partial\Omega)$  (the dual space of  $\mathcal{H}^{1/2}(\Omega) = \gamma_0\mathcal{H}^1(\Omega)$ ), and  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$  we will define  $U$  to be a *generalized solution* of (1)–(3), if  $U - \Theta \in \mathcal{H}_\Delta^1(\Omega)$  and  $A_\lambda(U, V) = (F, V) + (G, V)_\Delta$  for all  $V \in \mathcal{H}_\Delta^1(\Omega)$  (or equivalently for all  $V \in \mathcal{C}_\Delta^\infty(\Omega) = \{(v_1, v_2, \dots, v_m) \in \mathcal{C}^\infty(\Omega) | v_i = 0 \text{ on an open neighborhood of } \Gamma_i, 1 \leq i \leq m\}$ ). Of course every classical solution is a generalized solution and, for sufficiently large  $\lambda$ , there exists at most one generalized solution.

We shall need the following theorem of Stampacchia.

THEOREM 3. Let  $A$  be a continuous bilinear functional on a real Hilbert space  $\mathcal{Y}$  with inner product  $\langle \cdot, \cdot \rangle$  and let  $\mathcal{U} \subset \mathcal{Y}$  be a closed convex subset. Suppose  $A$  is strongly coercive on  $\mathcal{U} - \mathcal{U}$ ; i.e., there is a positive constant  $c$  such that  $A(y, y) \geq c\langle y, y \rangle$  for all  $y \in \mathcal{U} - \mathcal{U}$ . Let

$$\mathcal{U}_z = \{y \in \mathcal{Y} | z + \varepsilon y \in \mathcal{U} \text{ for some } \varepsilon > 0\}.$$

Then for each  $f \in \mathcal{Y}$  there exists a unique element  $z \in \mathcal{U}_z$  such that

$$A(z, y) \geq \langle f, y \rangle \text{ for all } y \in \mathcal{U}_z.$$

The proof of this theorem can be found in [22] for the case where  $A$  is strongly coercive on all of  $\mathcal{Y}$ . However, an examination of the proof shows that strong coerciveness on  $\mathcal{U} - \mathcal{U}$  suffices. The minor modifications needed in the proof were pointed out in [12].

We use  $K$  to denote the cone of nonnegatively valued functions in  $H^1(\Omega)$ . Consistent with our earlier notation,  $\mathcal{K}$  will denote the Cartesian product of  $m$  copies of  $K$ .

We remark here that the following two lemmas, 4 and 5, are true even if we impose no regularity conditions on  $\partial\Omega$  or  $\Delta$ . These two lemmas correspond to similar results obtained by Stampacchia [22]. First we need another definition.

DEFINITION. Let  $\mathcal{U}$  be a subspace of  $\mathcal{H}^1(\Omega)$ . Then  $U \in \mathcal{H}^1(\Omega)$  is called a  $\mathcal{U}$ -subsolution for (1)–(3) if  $A_\lambda(U, V) \leq 0$  for all  $V \in \mathcal{U} \cap \mathcal{K}$ .

LEMMA 4. If  $U_1$  and  $U_2$  are  $\mathcal{H}_\Delta^1(\Omega)$ -subsolutions,  $\lambda > \mu_0, \mu_0$  as in Lemma 2, and  $W = \max(U_1, U_2)$ , the component-wise maximum, then  $W$  is also a  $\mathcal{H}_\Delta^1(\Omega)$ -subsolution.

Before we prove this lemma we need to make several observations whose proofs can be found in [15, pp. 50–54]. If  $k$  is a constant, then the function  $(u \vee k)(x) \equiv \max(u(x), k)$  is a member of  $H^1(\Omega)$  whenever  $u \in H^1(\Omega)$ . Also, if  $u_n \rightarrow u$  in  $H^1(\Omega)$ , then  $u_n \vee k \rightarrow u \vee k$  in  $H^1(\Omega)$ . Moreover the distributional derivatives of  $u \vee k$  satisfy

$$D_i(u \vee k)(x) = \begin{cases} 0 & \text{if } u(x) \leq k, \\ D_i u(x) & \text{if } u(x) > k. \end{cases}$$

But since  $u \vee v = u + (v - u) \vee 0 \in H^1(\Omega)$  if  $u, v \in H^1(\Omega)$ , we see that

$$D_i(u \vee v)(x) = \begin{cases} D_i u(x) & \text{if } u(x) \geq v(x), \\ D_i v(x) & \text{if } u(x) < v(x). \end{cases}$$

Of course everything is modulo sets of measure zero; in particular  $D_i u = D_i v$  a.e. on the set where  $u = v$ . Analogous results also hold if we replace  $u \vee v$  by  $u \wedge v \equiv \min(u, v)$ .

Proof. Let  $\mathcal{U} = \{U \in \mathcal{H}^1(\Omega) | U \leq W \text{ and } U - W \in \mathcal{H}_\Delta^1(\Omega)\}$ , where  $\leq$  should

be interpreted as componentwise a.e. Clearly  $\mathcal{U} - \mathcal{U} \subset \mathcal{H}_\Delta^1(\Omega)$ . For each  $\Psi \in \mathcal{U}$ , we define

$$\mathcal{U}_\Psi = \{V \in \mathcal{H}_\Delta^1(\Omega) \mid \Psi + \varepsilon V \in \mathcal{U} \text{ for some } \varepsilon > 0\}.$$

We have the inclusions  $\mathcal{U}_\Psi \subset \mathcal{H}_\Delta^1(\Omega)$  and  $-\mathcal{H} \cap \mathcal{H}_\Delta^1(\Omega) \subset \mathcal{U}_\Psi$ . Now let  $\Psi$  be the unique element in  $\mathcal{U}$  such that  $A_\lambda(\Psi, Z) \geq 0$  for all  $Z \in \mathcal{U}_\Psi$ . This means that  $\Psi$  must be an  $\mathcal{H}_\Delta^1(\Omega)$ -subsolution. Let  $\Phi = \max(U_1, \Psi)$ . We note that there exists an element  $V \in \mathcal{H}_\Delta^1(\Omega)$  such that  $\Psi = V + W$ . There exists a sequence  $\{V_n\} \subset \mathcal{H}^1(\Omega)$  such that for each  $i$  and  $n$  there exists an open neighborhood  $N_{n,i}$  of  $\Gamma_i$  such that the  $i$ th component of  $V_n$  vanishes on  $N_{n,i}$  and such that  $V_n \rightarrow V$  in  $\mathcal{H}^1(\Omega)$ . From the above remarks we see that  $\max(V_n + W, U_1) - (V_n + W)$  converges to  $\Phi - \Psi$  in  $\mathcal{H}^1(\Omega)$ , but also the  $i$ th component of  $\max(V_n + W, U_1) - (V_n + W)$  vanishes on  $N_{n,i}$ . Therefore  $\Phi - \Psi \in \mathcal{H}_\Delta^1(\Omega)$ , and we have  $\Phi - \Psi \in \mathcal{U}_\Psi$ , so that

$$(7) \quad A_\lambda(\Psi, \Phi - \Psi) \geq 0.$$

We also claim that

$$(8) \quad A_\lambda(\Phi, \Phi - \Psi) \leq A_\lambda(U_1, \Phi - \Psi).$$

To see this we write

$$A_\lambda(\Phi - U_1, \Phi - \Psi) = \sum_{k=1}^m \{-(h_k^j(\varphi_j - u_{1j}), \varphi_k - \psi_k) - (e_k^j(\varphi_j - u_{1j}), \varphi_k - \psi_k)_{\partial\Omega}\}$$

which is indeed  $\leq 0$  since  $\varphi_j \geq u_{1j}$  and  $\varphi_k \geq \psi_k$  while  $h_k^j$  and  $e_k^j$  are  $\geq 0$ . Combining (7) and (8) we obtain

$$A_\lambda(\Phi - \Psi, \Phi - \Psi) \leq A_\lambda(U_1, \Phi - \Psi) \leq 0,$$

since  $U_1$  is an  $\mathcal{H}_\Delta^1(\Omega)$ -subsolution and  $\Phi - \Psi \in \mathcal{H} \cap \mathcal{H}_\Delta^1(\Omega)$ . Because  $\lambda > \mu_0$ , we see that  $\Phi = \Psi$  and hence  $U_1 \leq \Psi$ . Similarly it follows that  $U_2 \leq \Psi$  and consequently  $W = \Psi$ , an  $\mathcal{H}_\Delta^1(\Omega)$ -subsolution.  $\square$

LEMMA 5. *If  $U$  is a generalized solution of (1)–(3) with  $\lambda > \mu_0$ ,  $F \geq 0$ ,  $G \geq 0$ , and  $\Theta \geq 0$ , then  $U \geq 0$ .*

*Proof.* Both  $-U$  and  $0$  are  $\mathcal{H}_\Delta^1(\Omega)$ -subsolutions and hence so is  $W = \max(0, -U)$ . But since  $W$  is also in  $\mathcal{H} \cap \mathcal{H}_\Delta^1(\Omega)$  we see that  $A_\lambda(W, W) \leq 0$  and therefore  $W = 0$ .  $\square$

*Remark.* If  $f \in L^1(\Omega)$  and  $f \geq 0$  a.e., then,  $f \in \mathcal{D}_+(\Omega)'$ , but the converse is not generally true. However since  $\Theta \in \mathcal{L}^\infty(\Omega)$  one can easily show that  $\theta_k \in \mathcal{D}_+(\Omega)'$  implies  $\theta_k \geq 0$  a.e. by merely taking a sequence in  $\mathcal{D}_+(\Omega)'$  which converges in  $L^1(\Omega)$  to the characteristic function  $\chi$  of the set  $\{x \mid \theta_k(x) < 0\}$ . Therefore  $(\theta_k, \chi) = 0$ .

THEOREM 6. *Problem (1)–(3) has a unique generalized solution for each  $F \in \mathcal{H}^1(\Omega)'$ ,  $G \in \mathcal{H}^{-1/2}(\partial\Omega)$  and  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$  provided  $\lambda \geq \mu_0$ .*

*Proof.* This is a simple application of Theorem 3. Let  $\mathcal{U}$  be the affine space  $\Theta + \mathcal{H}_\Delta^1(\Omega)$  and  $\mathcal{Y}$  the Hilbert space  $\mathcal{H}^1(\Omega)$ . By the Riesz representation theorem there exists a  $T \in \mathcal{H}^1(\Omega)$  such that  $\langle T, U \rangle = (F, U) + (G, U)_{\partial\Omega}$  for all  $U \in \mathcal{H}^1(\Omega)$ , where  $\langle \cdot, \cdot \rangle$  is the usual inner product on  $H^1(\Omega)$  extended in the obvious manner to the direct sum of such spaces. Hence, since  $\mathcal{U} - \mathcal{U} = \mathcal{H}_\Delta^1(\Omega)$ , there exists a unique  $U \in \mathcal{U}$  such that

$$A_\lambda(U, V) \geq \langle T, V \rangle \quad \forall V \in \mathcal{H}_\Delta^1(\Omega).$$

But since  $-V \in \mathcal{H}_\Delta^1(\Omega)$  we have in fact equality.  $\square$



Using the Sobolev embedding theorem one finds that  $L^{q/2}(\Omega) \subset H^1(\Omega)'$  and  $L^p(\partial\Omega) \subset (\gamma_0 H^1(\Omega))'$ . This justifies the following definition.

DEFINITION. Let  $\mathcal{H}$  denote the space  $\mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$  with norm  $\|U\| = \|U\|_{1,2} + \|U\|_{0,\infty}$ , and let  $G_\lambda$  be the map from  $\mathcal{L}^{q/2}(\Omega) \times \mathcal{L}^p(\Delta) \times \mathcal{H}$  into  $\mathcal{H}^1(\Omega)$  which associates with each triple  $(F, G, \Theta)$  the unique solution  $U$  to (1)–(3) ( $\lambda > \mu_0$ ).

THEOREM 7. Suppose  $\lambda > \mu_0$  and  $U$  is a generalized solution to (1)–(3) with  $F \in \mathcal{L}^{q/2}(\Omega)$ ,  $G \in \mathcal{L}^p(\partial\Omega)$  and  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ . Then  $U \in \mathcal{L}^\infty(\Omega)$ .

Proof. In order to apply known results for single component equations we first consider the case where  $H = 0$  and  $E = 0$ . It will suffice to show that the solution is bounded from above. Let  $\mu > \|\Theta\|_{0,\infty}$  and  $M = (\mu, \mu, \dots, \mu)$ ; then  $U' = U - M$  satisfies

$$\begin{aligned} (L + \lambda)U' &= F_1 \in \mathcal{L}^{q/2}(\Omega) \quad \text{in } \Omega, \\ BU' &\leq G_1 \in \mathcal{L}^p(\partial\Omega) \quad \text{on } \Delta, \\ U' &= \Theta - M \leq 0 \quad \text{on } \Gamma, \end{aligned}$$

where  $F_1 = (f_{11}, f_{12}, \dots, f_{1m})$  with  $f_{1i} = f_i + \mu D_j d_i^j - \mu c_i - \mu \lambda$  and  $G_1 = (g_{11}, g_{12}, \dots, g_{1m})$  with  $g_{1i} = g_i - \mu \sigma_i$ . Applying Lemma 5 we see that  $U' \leq V$  where,

$$\begin{aligned} (A) \quad (L + \lambda)V &= F_2 \quad \text{in } \Omega, \\ BV &= G_2 \quad \text{on } \partial\Omega, \end{aligned}$$

where  $F_2$  (resp.  $G_2$ ) consists of the absolute values of the components of  $F_1$  (resp.  $G_1$ ). Let  $V_1$  be the generalized solution of

$$\begin{aligned} (B) \quad (L + \lambda)V_1 &= F_2 \in \mathcal{L}^{q/2}(\Omega), \\ \partial V_1 / \partial N &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where  $\partial / \partial N = \nu_i (a_k^{ij} D_j + d_k^i)$ . We can apply a result of Stampacchia [21] which states that the solution  $u$  of

$$\begin{aligned} (L_k + \lambda)u &= f \in W^{-1,\sigma}(\Omega), \\ \partial u / \partial N &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

will be in  $L^p(\Omega)$  if  $\rho^{-1} > \sigma^{-1} - n^{-1}$  ( $\rho = \infty$  is allowed, setting  $1/\infty = 0$ ). By the Sobolev embedding theorem  $W_0^{1,\alpha^*} \subset L^{(q/2)^*}$  where  $\alpha^* \equiv \alpha(\alpha - 1)^{-1}$ . Therefore  $L^{q/2} \subset W^{-1,q}(\Omega)$  and, since  $q^{-1} - n^{-1} < 0$  we have  $V_1 \in \mathcal{L}^\infty(\Omega)$ . Recalling that  $\sigma_k \geq 0$  and noting that  $V \geq 0$ , we see that  $V - V_1 \leq V_2$ , where

$$\begin{aligned} (C) \quad (L + \lambda)V_2 &= 0, \\ \frac{\partial V_2}{\partial N} &= G_2 \in \mathcal{L}^p(\partial\Omega). \end{aligned}$$

Another regularity result, due to Murthy and Stampacchia [19], tells us that  $V_2 \in \mathcal{L}^\infty(\Omega)$  since  $p > n - 1$ . (The work of Murthy and Stampacchia deals with a more complicated problem, namely certain degenerate elliptic problems. Also the theorem we need, [19, p. 61], contains some minor, but confusing, errors. For these reasons we have included a proof of this result in an Appendix). The regularity results which we used were proven for single component problems. When one includes coupling terms  $-HU$  and  $-EU$  on the right-hand side of (A), they are no longer a priori in  $\mathcal{L}^{q/2}(\Omega)$  and  $\mathcal{L}^p(\partial\Omega)$  respectively (unless  $n \leq 3$ ). Although we will not take this route, we do note that one can treat (B), with coupling  $-HU$ , by bootstrapping,

showing that if  $U \in \mathcal{L}^\rho(\Omega)$ ,  $\rho < q/2$ , then  $U \in \mathcal{L}^{\rho+\varepsilon}(\Omega)$  for some  $0 < \varepsilon \leq q/2 - \rho$ , etc. The proof of the regularity result for (C) is relatively simple, and can be easily extended to the case where we introduce a coupling term  $-EU$ . However to use this approach to deal with the case where we have both coupling terms present is rather lengthy (unless  $n \leq 3$ ). Therefore, we will use a different approach. Let  $n - 1 < p' < p$ ,  $n < q' < q$ ,  $r$  so large that  $r^{-1} + p^{-1} < (p')^{-1}$ , and  $r^{-1} + (q/2)^{-1} < (q'/2)^{-1}$ . Let  $\mathcal{G}$  be the map from  $\mathcal{L}^{q'/2}(\Omega) \times \mathcal{L}^{p'}(\partial\Omega)$  into  $\mathcal{H}$  defined by

$$\begin{aligned} (L + \lambda)\mathcal{G}(F, G) &= F \in \mathcal{L}^{q'/2}(\Omega) \quad \text{in } \Omega, \\ B\mathcal{G}(F, G) &= G \in \mathcal{L}^{p'}(\Omega) \quad \text{on } \partial\Omega. \end{aligned}$$

One easily sees that  $\mathcal{G}$  is a closed linear operator and hence continuous by the closed graph theorem. We claim that the map  $\mathcal{G} \circ \mathcal{P} \circ \mathcal{I}$  defined by the diagram below is a compact continuous linear map from  $\mathcal{H}$  into itself:

$$\begin{aligned} U \in \mathcal{H} \xrightarrow{\mathcal{I}} (U, \gamma_0 U) \in \mathcal{L}^r(\Omega) \times \mathcal{L}^r(\partial\Omega) \xrightarrow{\mathcal{P}} (HU, EU) \in \mathcal{L}^{q'/2}(\Omega) \times \\ \mathcal{L}^{p'}(\partial\Omega) \xrightarrow{\mathcal{G}} V \in \mathcal{H}. \end{aligned}$$

To see this we note that by Hölder's inequality  $\mathcal{P}$  is bounded linear, while  $\mathcal{I}$  is obviously continuous. But  $\mathcal{I}$  is also compact, for if  $\{U_n\}$  is bounded there must exist a subsequence  $\{U_{n_i}\}$  such that both  $U_{n_i}$  and  $\gamma_0 U_{n_i}$  converge in  $\mathcal{L}^2(\Omega)$  and  $\mathcal{L}^2(\partial\Omega)$ . If  $U_n \rightarrow U$  then a fortiori,  $\gamma_0 U_{n_i} \rightarrow \gamma_0 U$ . It may furthermore be assumed, without loss of generality, that  $U_{n_i} \rightarrow U$  a.e. in  $\Omega$ , and  $\gamma_0 U_{n_i} \rightarrow \gamma_0 U$  a.e. in  $\partial\Omega$ . But we also have bounded conditions a.e.; therefore, applying the dominated convergence theorem, we have  $U_{n_i} \rightarrow U$  in  $\mathcal{L}^r(\Omega)$ , and  $\gamma_0 U_{n_i} \rightarrow \gamma_0 U$  in  $\mathcal{L}^r(\partial\Omega)$ . We next consider, in  $\mathcal{H}$ , the equation

$$(D) \quad U - \mathcal{G} \circ \mathcal{P} \circ \mathcal{I}U = \mathcal{G}(F_1, G_1).$$

According to the Fredholm theory this equation has a solution in  $\mathcal{H}$  if  $\ker(\text{id.} + \mathcal{G} \circ \mathcal{P} \circ \mathcal{I})$  is trivial. But if  $U_0$  were in the kernel, then one easily sees that  $(L + \lambda - H)U_0 = 0$  and  $(B - E)U_0 = 0$  with  $\lambda > \mu_0$ . By Lemma 5,  $U_0 = 0$ . Therefore (D) has a solution  $U \in \mathcal{H}$  which is also a solution of (1)–(3). By uniqueness the proof is complete.  $\square$

**THEOREM 8.**  $\mathbf{G}_\lambda: \mathcal{H} \rightarrow \mathcal{H}$  ( $\lambda > \mu_0$ ) is a continuous map which is monotone in the sense that  $F_1 \leq F_2$  a.e.,  $G_1 \leq G_2$  a.e., and  $\Theta_1 \leq \Theta_2$  a.e. imply  $\mathbf{G}_\lambda(F_1, G_1, \Theta_1) \leq \mathbf{G}_\lambda(F_2, G_2, \Theta_2)$  a.e.

*Proof.* The monotonicity is of course a direct consequence of linearity and Lemma 5. Let  $U_i = \mathbf{G}_\lambda(F_i, G_i, \Theta)$ ,  $i = 1, 2$ ; then  $U_1 - U_2 \in \mathcal{H}_\Delta^1(\Omega)$  and hence letting  $\varepsilon = \min(\varepsilon_0, \lambda - \mu_0)$ ,

$$\begin{aligned} \varepsilon \|U_1 - U_2\|_1^2 &\leq A_\lambda(U_1 - U_2, U_1 - U_2) \\ &= (F_1 - F_2, U_1 - U_2)_\Omega + (G_1 - G_2, \gamma_0 U_1 - \gamma_0 U_2)_{\partial\Omega} \\ &\leq \text{cst} \{ \|F_1 - F_2\|_{0,q/2} + \|G_1 - G_2\|_{0,p,\Delta} \} \|U_1 - U_2\|_1, \end{aligned}$$

where we have used the Sobolev and Hölder inequalities. Hence for fixed  $\Theta$ ,  $\mathbf{G}_\lambda$  must be a closed linear operator. Applying the closed graph theorem, we have continuity with respect to  $(F, G)$ . To conclude the proof it suffices to show that  $\mathbf{G}_\lambda$  is continuous with respect to  $\Theta$  for  $F \equiv 0$  and  $G \equiv 0$ . Again this reduces to showing that the graph  $\{(\Theta, \mathbf{G}_\lambda(0, 0, \Theta))\}$  is closed. To see this suppose  $\Theta_i \rightarrow 0$  and  $\mathbf{G}_\lambda(0, 0, \Theta_i) \rightarrow W$ ; then, by extending standard arguments (see e.g., [2]) to the multi-component case (cf. (7))

it can be shown that

$$0 = \int_{\Omega} (L + \lambda - H)U_i W \, dx = A_{\lambda}(U_i, W) - \int_{\Gamma} \frac{\partial U_i}{\partial N} W \, dS,$$

where  $\partial/\partial N : \mathcal{H}^1(\Omega) \rightarrow \mathcal{H}^{-1/2}(\partial\Omega)$  is a continuous linear map corresponding to  $\{v_i(a_k^{ij}D_j + d_k^i)\}_{k=1}^m$  (Since we will never use continuity with respect to  $\Theta_i$ , we omit the details). Therefore, for  $\lambda > \mu_0$  we have

$$\varepsilon_0 \|U_i\|_1^2 \leq A_{\lambda}(U_i, U_i) = \int_{\Gamma} \frac{\partial U_i}{\partial N} \gamma_0 \Theta_i \, dS \leq \text{cst} \times \|U_i\|_1 \|\Theta_i\|_1.$$

One can show, using the methods used early in the proof of the previous theorem, that the graph is closed in the  $\mathcal{L}^2$ -topology:  $\|U_i\| \leq \text{cst} \times \|\Theta\|_{2,1,\Omega}$ . Either of these inequalities can be used in conjunction with the closed graph theorem, or these inequalities may be used together without resorting to the closed graph theorem.  $\square$

It will be convenient to introduce the following notation:

DEFINITION.  $\bar{\mu} = \max(\mu_0, \mu_1)$ .

**3. The nonlinear elliptic problem.** Throughout the rest of the paper we will assume that hypotheses (I)–(IV) are satisfied. Let us consider

$$(4') \quad (L_k + \lambda)u_k - h_k^i u_i = f_k(x, U) \quad \text{in } \Omega,$$

$$(5') \quad B_k u_k - e_k^i u_i = g_k(x, U) \quad \text{on } \Delta_k,$$

$$(6') \quad u_k = \theta_k \quad \text{on } \Gamma_k.$$

Using the more concise notation we define the formal nonlinear operator  $\mathcal{N}$  by defining  $\mathcal{N}(U) = V$  if  $V$  is a generalized solution of

$$(L + \lambda)V - HV = F(x, U) \quad \text{in } \Omega,$$

$$BV - EV = G(x, U) \quad \text{on } \Delta,$$

$$V = \Theta \quad \text{on } \Gamma.$$

Then solving (4')–(6') is tantamount to finding a fixed point for  $\mathcal{N}$ . We shall be interested in the case where  $F$  and  $G$  are dominated by affine functions. This is a reasonable assumption for many practical applications. For one thing, it means that positive solutions to the associated parabolic equations (i.e., reaction-diffusion equations) grow no faster than exponentially, thus ensuring existence of a global solution whenever local solutions exist. In other words, we want to assume that there exist a matrix  $H_F(x)$  whose entries are all positive, and some vector  $D(x)$  such that

$$(9) \quad F(x, U) \leq H_F(x)U + D(x).$$

Obviously, due to the presence of  $H$  on the left side of our equations, we may subtract  $H_F U$  on both sides, therefore assuming that  $F$ , and similarly  $G$ , are bounded from above for all  $U$ . As a specific example, let us consider the case where one models the processes of chemical reactor kinetics or of flame propagation (see [4] for the equations). In both these cases one of the components is temperature, and the boundary condition is obtained from heat flux considerations at the boundary. If a significant amount of heat is lost by radiation, one expects a boundary condition of the form

$$(10) \quad K \frac{\partial u}{\partial \nu} = g(u) \equiv \alpha + \beta u - \gamma_1(u^4 - u_0^4) \quad \text{on } \partial\Omega,$$

where  $u \geq u_0$ ,  $u_0$  is the temperature of the exterior region,  $\alpha \geq 0$ ,  $\gamma_1$  is a positive constant obtained as the product of the emissivity of the container's surface and the Stefan-Boltzmann constant [3], and  $K$  is the heat conductivity. If, on the other hand one assumes natural convection at the boundary, one obtains [3]

$$(11) \quad K \frac{\partial u}{\partial \nu} = g(u) \equiv -\gamma_2(u - u_0)^{5/4},$$

where  $\gamma_2 > 0$  and  $u \geq u_0$ . In the interval  $0 < u < u_0$ , the remaining physically meaningful range of the temperatures, one might have some other boundary condition which matches at  $u_0$ . In any case we notice that in both cases  $g(u)$  is dominated by a linear function for  $u \geq 0$ . For  $u < 0$  we can apparently define  $g$  to be whatever is convenient in order to satisfy mathematical hypotheses. That this causes no problems follows from a result which we shall prove; the existence theorem stated below is still valid even if the linear domination hypothesis fails in some region, provided some other condition holds. In the above example this condition amounts to observing that if we set  $\tilde{u} \equiv 0$  we get

$$K \frac{\partial \tilde{u}}{\partial \nu} \leq g(\tilde{u}).$$

When the corresponding partial differential equation is also nonhomogeneous, we must require a similar inequality there. For example if we are dealing with a one-component case  $Lu = f(u)$ , we also require  $L\tilde{u} \leq f(\tilde{u})$ . (Following standard terminology one may call  $\tilde{u}$  a subsolution, a term which we however have already used.) In addition to domination by an affine map, we also must require some reasonable local behavior.

DEFINITION. Let  $(S, \mu)$  be a measure space and  $T$  a function mapping  $S \times R^m$ , or a subset thereof, into  $R^k$ . Then  $T$  is said to satisfy the *Caratheodory condition* if  $T(x, U)$  is measurable in  $x$  for each fixed  $U \in R^m$ , and is continuous in  $U$  for almost all  $x$  in  $S$ .

DEFINITION. Let  $\mathcal{S} \subset S \times R^m$ . Then  $\mathcal{F}_r(\mathcal{S})$  denotes the class of all functions  $T : \mathcal{S} \rightarrow R^m$  which satisfy the Caratheodory condition and also satisfy:

- (i) There exists a  $D \in \mathcal{L}^r(S)$  such that  $T(x, U) \leq D(x)$  for all  $(x, U) \in \mathcal{S}$ .
- (ii) For each real number  $\nu$  there exists a  $T_\nu \in \mathcal{L}^r(S)$  such that  $F(x, U) \geq T_\nu(x)$  for all  $(x, U) \in \mathcal{S}$  with  $U \leq (\nu, \nu, \dots, \nu)$ .

A simple example of a map  $T \in \mathcal{F}_r(\Omega \times R^m)$  is one which is continuous, nonincreasing, and bounded from above. Another example is a continuous function which is bounded. In particular, if  $\mathcal{S}$  is bounded and closed, and  $T$  continuous on  $\mathcal{S}$ , then  $T \in \mathcal{F}_r(\mathcal{S})$  for any  $0 < r \leq \infty$ .

We introduce another hypothesis which will be needed for almost all subsequent results.

- (V). There exist numbers  $\gamma_1 \geq 0, \gamma_2 \geq 0$ , such that for all  $(x, U), (x, V)$

$$\begin{aligned} (F(x, U) - F(x, V)) \cdot (U - V) &\leq \gamma_1 |U - V|^2, \\ (G(x, U) - G(x, V)) \cdot (U - V) &\leq \gamma_2 |U - V|^2. \end{aligned}$$

Using the notation of Lemma 2 (i) and (ii) we define  $\gamma = \gamma_1 + \gamma_2 C \frac{\epsilon_0}{\gamma_2}$ .

THEOREM 9. Suppose (I)–(IV) are satisfied,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $F \in \mathcal{F}_{q/2}(\Omega \times R^m)$ ,  $G \in \mathcal{F}_p(\partial\Omega \times R^m)$  and  $\lambda > \bar{\mu}$ . Then (4')–(6') has a generalized solution. If (V) is also satisfied, then this solution is unique for  $\lambda > \mu_0 + \gamma$ .

Proof. Let  $D_F \in \mathcal{L}^{q/2}(\Omega)$  and  $D_G \in \mathcal{L}^p(\partial\Omega)$  such that  $F(x, U) \leq D_F(x)$  and

$G(x, U) \leq D_G(x)$  for all  $U \in R^m$ . By Theorem 7 we know there exists a number  $\nu > 0$  such that  $N_\nu \equiv (\nu, \nu, \dots, \nu) \geq \mathcal{G}_\lambda(D_F, D_G, \Theta)$ . Let

$$\mathfrak{K} = \{U \in \mathcal{H}^1(\Omega) \mid \mathcal{G}_\lambda(F_\nu, G_\nu, \Theta) \leq U \leq N_\nu\},$$

where  $F(x, U) \geq F_\nu(x) \in \mathcal{L}^{q/2}(\Omega)$  and  $G(x, U) \geq G_\nu(x) \in \mathcal{L}^p(\partial\Omega)$  for all  $U \leq N_\nu$ . Now  $\mathfrak{K}$  is mapped into itself by  $\mathcal{N}$ , for if  $U \in \mathfrak{K}$  then

$$\mathcal{N}(U) = \mathcal{G}_\lambda(F(x, U), G(x, U), \Theta) \leq \mathcal{G}_\lambda(D_F, D_G, \Theta) \leq N_\nu,$$

and

$$\mathcal{N}(U) \geq \mathcal{G}_\lambda(F_\nu, G_\nu, \Theta).$$

It remains to prove that  $\mathcal{N}$  is compact continuous (in the  $\mathcal{H}^1(\Omega)$ -topology) because then the result follows from Schauder's fixed point theorem. Suppose  $\{U_i\}$  is a sequence in  $\mathfrak{K}$  which is bounded with respect to the norm  $\|\cdot\|_1$  in  $\mathcal{H}^1(\Omega)$ . We can, by Rellich's lemma, find a subsequence  $\{U_{i'}\}$  which converges in  $\mathcal{L}^2(\Omega)$ . Also, since  $\gamma_0: \mathcal{H}^1(\Omega) \rightarrow \mathcal{L}^2(\partial\Omega)$  is compact continuous, we may assume that  $\gamma_0 U_{i'}$  converges in  $\mathcal{L}^2(\partial\Omega)$  (a fortiori to  $\gamma_0 U$ , where  $U$  is the  $\mathcal{L}^2(\Omega)$ -limit of the sequence  $\{U_{i'}\}$ ). We have (Lemma 2),

$$\begin{aligned} \varepsilon_0 \|\mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'})\|_1^2 &\leq A_\lambda(\mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'}), \mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'})) \\ &= \int_\Omega (F(x, U_{i'}) - F(x, U_{j'})) \cdot (\mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'})) \, dx \\ &\quad + \int_{\partial\Omega} (G(x, U_{i'}) - G(x, U_{j'})) \cdot \gamma_0(\mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'})) \, dS. \end{aligned}$$

By Theorem 7  $\mathcal{N}(\mathfrak{K})$  is bounded in the norm  $\|\cdot\| \equiv \|\cdot\|_{1,2,\Omega} + \|\cdot\|_{0,\infty,\Omega}$  of  $\mathcal{H} \equiv \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ . Therefore there exists a constant  $c$  such that

$$(12) \quad \begin{aligned} \varepsilon_0 \|\mathcal{N}(U_{i'}) - \mathcal{N}(U_{j'})\|_1^2 \\ \leq c\{\|F(x, U_{i'}) - F(x, U_{j'})\|_{0,1,\Omega} + \|G(x, U_{i'}) - G(x, U_{j'})\|_{0,1,\partial\Omega}\} \end{aligned}$$

Since  $\mathfrak{K}$  is a bounded set in  $\mathcal{L}^\infty(\Omega)$  the Nemytskii operator  $F$  takes  $\mathfrak{K}$  into  $\mathcal{L}^{q/2}(\Omega)$ . Similarly the image of  $\gamma_0(\mathfrak{K})$  under the Nemytskii operator  $G$  is bounded in  $\mathcal{L}^p(\partial\Omega)$ . But this means [14, p. 22] that these operators, being defined through functions satisfying the Caratheodory condition, are continuous on  $\mathfrak{K}$  and  $\gamma_0(\mathfrak{K})$  in their respective  $\mathcal{L}^1$ -topologies. Hence, by (12),  $\{\mathcal{N}(U_{i'})\}$  is a Cauchy sequence in  $\mathcal{H}^1(\Omega)$ . We have incidentally shown that (12) also implies continuity. To prove uniqueness we suppose that  $\mathcal{N}(U) = U$  and  $\mathcal{N}(V) = V$ ; then using Lemma 2 we get

$$\begin{aligned} \varepsilon_0 \|U - V\|_1^2 + (\lambda - \mu_0) \|U - V\|_0^2 \\ \leq A_\lambda(U - V, U - V) \\ = A_\lambda(\mathcal{N}(U) - \mathcal{N}(V), U - V) \\ \leq \gamma_1 \|U - V\|_0^2 + \frac{1}{2} \varepsilon_0 \|U - V\|_1^2 + \gamma_2 C(\varepsilon_0/2\gamma_2) \|U - V\|_0^2 \\ = \frac{1}{2} \varepsilon_0 \|U - V\|_1^2 + \gamma \|U - V\|_0^2. \end{aligned}$$

Therefore, if  $\lambda > \mu_0 + \gamma$  then  $U = V$ .  $\square$

Of course the above theorem is also valid if the conditions on  $F$  and  $G$  are replaced by  $-F \in \mathcal{F}_{q/2}(\Omega \times R^m)$  and  $-G \in \mathcal{F}_p(\partial\Omega \times R^m)$ . The above result as well as the next theorem generalize similar results obtained in [12] for one-component equations.

**THEOREM 10.** *Suppose (I)–(IV) are satisfied,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $F$  and  $G$  satisfy the Caratheodory condition on  $\Omega \times R^m$  and  $\partial\Omega \times R^m$  respectively, and  $\lambda > \bar{\mu}$ . Suppose there exist nonincreasing functions  $\phi$  and  $\psi$  from  $R^1$  into itself such that for all  $k \cong k_0 > 0$ ,*

$$(13) \quad \limsup_{s \rightarrow \infty} \frac{k\phi(k\psi(s))}{s} < 1,$$

and for each  $1 \cong i < m$  we have the growth conditions

$$f_i(x, u_1, u_2, \dots, u_m), g_i(x, u_1, u_2, \dots, u_m) \cong \phi(s) \quad \text{if } u_j \cong s \forall j,$$

$$f_i(x, u_1, u_2, \dots, u_m), g_i(x, u_1, u_2, \dots, u_m) \cong \psi(s) \quad \text{if } u_j \cong s \forall j.$$

Then (4')–(6') has a generalized solution. If (V) is also satisfied and  $\lambda > \mu_0 + \gamma$ , then the solution is unique.

*Remark.* Simple examples of functions satisfying (13) are  $\psi(s) = -a - b \max(0, s)^{\gamma_1}$  and  $\phi(s) = a + b \max(0, -s)^{\gamma_2}$ , where  $a, b, \gamma_1$ , and  $\gamma_2$  are positive constants satisfying  $\gamma_1\gamma_2 < 1$ .

*Proof.* Let  $N_1 = (1, 1, \dots, 1) \in R^m$  and  $k_1 = \|\mathbf{G}_\lambda(N_1, N_1, |\Theta|)\|_\infty + k_0$ , and let

$$\mathfrak{R}_y = \{U \in \mathcal{H}^1(\Omega) \mid k_1\psi(y) \cong u_i \cong y\},$$

where  $y > 0$  is chosen so large that  $k_1\phi(k_1(\psi(y)))/y < 1$ . Then  $\mathcal{N}$  maps  $\mathfrak{R}_y$  into itself. To see this we may assume without loss of generality that  $\phi(0) = -\psi(0) \cong 1$ .

$$\begin{aligned} \mathcal{N}(U) &= \mathbf{G}_\lambda(F(x, U), G(x, U), \Theta) \cong \mathbf{G}_\lambda(\psi(y)N_1, \psi(y)N_1, |\Theta|) \\ &\cong \psi(y)\mathbf{G}_\lambda(N_1, N_1, |\Theta|) \cong k_1\psi(y)N_1. \end{aligned}$$

Also

$$\mathcal{N}(U) \cong \mathbf{G}_\lambda(\phi(k_1\psi(y))N_1, \phi(k_1\psi(y))N_1, |\Theta|) \cong \phi(k_1\psi(y))k_1N_1 \cong yN_1.$$

As in the proof of Theorem 9, we have all the necessary components to justify the use of Schauder's fixed point theorem. Uniqueness follows from the same argument that was used in the proof of the previous theorem.  $\square$

We conclude this section with a theorem on invariant sets which constitutes the crucial ingredient in the proof of the invariant set theorem for the reaction-diffusion equations discussed in the next section. Instead of viewing the result as an invariant set theorem we might, maybe more appropriately, regard it is a nonlinear generalization of Lemma 5, i.e., as a sort of maximum principle. We would then expect to need the following conditions: i) the  $i$ th component of  $F(x, U) + HU$  is nondecreasing in  $u_j$  for each  $j \neq i$  (corresponding to the hypothesis  $H \cong 0$  in Lemma 5); ii) (V) is satisfied (corresponding to the coerciveness requirement of  $A_\lambda$ ). Also in order to be able to treat nonlinearities of the type occurring in (10) and (11) we certainly want to allow  $f_i(x, U)$  to decrease "rapidly" with respect to  $u_i$ . This last requirement has tended to make our proof rather lengthy. Before we proceed we must introduce some more notation.

**DEFINITIONS.**

$$(i) \quad \hat{F}(x, U) = F(x, U) + H(x)U, \quad \hat{F} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m),$$

$$\hat{G}(x, U) = G(x, U) + E(x)U, \quad \hat{G} = (\hat{g}_1, \hat{g}_2, \dots, \hat{g}_m).$$

(ii) We use  $+\infty$  (resp.  $-\infty$ ) to also denote the extended real valued function  $x \rightarrow +\infty$  (respectively  $x \rightarrow -\infty$ ). For convenience we define  $(L_i + \lambda)(\pm\infty) = \pm\infty$  and  $B_i(\pm\infty) = \pm\infty$ .

(iii)  $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_m)$  where  $\varphi_i \in H^1(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)$  for  $1 \leq i \leq d$  and  $\varphi_i = -\infty$  for  $i > d$ .  $\Psi = (\psi_1, \psi_2, \dots, \psi_m)$ , where  $\psi_i \in H^1(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)$  for  $\delta \leq i \leq l$  and  $\psi_i = +\infty$  for  $i < \delta$  or  $i > l$ . Also we assume the indexing is such that  $\delta \leq d + 1$ . In other words the indices  $1 \leq i < \delta$  are those for which  $\varphi_i$  is finite-valued and  $\psi_i$  is  $+\infty$ , the indices  $\delta \leq i \leq d$  are those for which both  $\varphi_i$  and  $\psi_i$  are finite-valued, the indices  $d + 1 \leq i \leq l$  are those for which  $\varphi_i$  is  $-\infty$  but  $\psi_i$  is finite-valued, and the indices  $l < i \leq m$  are those for which both  $\varphi_i = -\infty$  and  $\psi_i = +\infty$ . We also use  $[\Phi, \Psi]$  to denote  $\{U \in \mathcal{H}^1(\Omega) \mid \Phi \leq U \leq \Psi\}$ .

(iv)  $\mathcal{S}_\Phi^\Psi = \{(x, U) \in \Omega \times R^m \mid \Phi(x) \leq U \leq \Psi(x)\},$

$\partial \mathcal{S}_\Phi^\Psi = \{(x, U) \in \partial\Omega \times R^m \mid \gamma_0 \Phi(x) \leq U \leq \gamma_0 \Psi(x)\}.$

(v) For any  $U \in \mathcal{H}^1(\Omega)$ ,  $U_\Phi = (\varphi_1, \varphi_2, \dots, \varphi_d, u_{d+1}, \dots, u_m)$ , and  $U^\Psi = (u_1, u_2, \dots, u_{\delta-1}, \psi_\delta, \psi_{\delta+1}, \dots, \psi_l, u_{l+1}, \dots, u_m)$ . That is to say  $U_\Phi$  is obtained from  $\Phi$  by replacing all components which are  $-\infty$  by corresponding components from  $U$ , and similarly  $U^\Psi$  is obtained from  $\Psi$  by replacing components which are  $+\infty$  by corresponding components from  $U$ .

**THEOREM 11.** *Suppose that (I)–(V) are satisfied and  $\lambda > \bar{\mu} + \gamma$ ,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $F \in \mathcal{F}_{q/2}(\mathcal{S}_\Phi^\Psi)$ ,  $G \in \mathcal{F}_p(\partial \mathcal{S}_\Phi^\Psi)$ , and that  $\hat{f}_i(x, u_1, u_2, \dots, u_m)$  and  $\hat{g}_i(x, u_1, u_2, \dots, u_m)$  are nondecreasing in  $u_j$  for all  $1 \leq j \leq l$  with  $j \neq i$ . Suppose  $\Phi \leq \Theta \leq \Psi$  and that for all  $U \in [\Phi, \Psi]$ :*

(IQ)  $(L + \lambda)\Phi \leq HU_\Phi + F(x, U_\Phi) \quad \text{and} \quad (L + \lambda)\Psi \geq HU^\Psi + F(x, U^\Psi) \quad \text{in } \Omega,$   
 $B\Phi \leq EU_\Phi + G(x, U_\Phi) \quad \text{and} \quad B\Psi \geq EU^\Psi + G(x, U^\Psi) \quad \text{on } \Delta.$

Then (4')–(6') has a unique generalized solution  $\bar{U} \in [\Phi, \Psi]$ .

We shall postpone the proof until the end of this section. This theorem can be viewed as an invariant set theorem in the following way. For  $\lambda$  sufficiently large, let  $\mathcal{T}_\lambda : \mathcal{L}^{q/2}(\Omega) \times \mathcal{L}^p(\partial\Omega) \times \mathcal{H} \rightarrow \mathcal{H} \cap \mathcal{C}(\Omega)$  be the operator defined by  $\mathcal{T}_\lambda(F_0, G_0, \Theta_0) = V$ , where  $V$  is the unique solution of

$$\begin{aligned} (L + \lambda - H)V - F(x, V) &= F_0 \quad \text{in } \Omega, \\ (B - E)V - G(x, V) &= G_0 \quad \text{on } \Delta, \\ V &= \Theta_0 \quad \text{on } \Gamma. \end{aligned}$$

The fact that  $V \in \mathcal{C}(\Omega)$  follows from known regularity results [15, p. 201]. Suppose  $F \in \mathcal{F}_{q/2}(\Omega \times R^m)$ ,  $G \in \mathcal{F}_p(\partial\Omega \times R^m)$ , and that the inequalities (IQ) are satisfied for all  $U \in [\Phi, \Psi]$ . Then for fixed  $\Theta_0 \in [\Phi, \Psi]$  and  $\mu > 0$ , the map  $W \rightarrow \mathcal{T}_{\lambda+\mu}(\mu W, 0, \Theta_0)$  leaves  $[\Phi, \Psi]$  invariant. The proof of this follows immediately from the inequalities  $(L + \lambda + \mu)\Phi \leq HU_\Phi + F(x, U_\Phi) + \mu W$  and  $(L + \lambda + \mu)\Psi \geq HU^\Psi + F(x, U^\Psi) + \mu W$ . It is also easy to prove the following generalization of Lemma 5.

**COROLLARY 12.** *Suppose (I)–(V) are satisfied,  $\lambda > \bar{\mu} + \gamma$ ,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $F \in \mathcal{F}_{q/2}(\Omega \times R^m)$ ,  $G \in \mathcal{F}_p(\partial\Omega \times R^m)$  and that for all  $1 \leq i \leq l : \hat{f}_i(x, u_1, u_2, \dots, u_m)$  and  $\hat{g}_i(x, u_1, u_2, \dots, u_m)$  are nondecreasing in  $u_j$  for all  $j \neq i$ . Then  $\mathcal{T}_\lambda$  is an order preserving map; i.e.,*

if  $F_1 \geq F_0, G_1 \geq G_0, \Theta_1 \geq \Theta_0,$  then  $\mathcal{T}_\lambda(F_1, G_1, \Theta_1) \geq \mathcal{T}_\lambda(F_0, G_0, \Theta_0).$

To prove this let  $\Phi = \mathcal{T}_\lambda(F_0, G_0, \Theta_0)$  and  $U = \mathcal{T}_\lambda(F_1, G_1, \Theta_1)$ , and apply the theorem.

In Corollary 12 we have a lot of monotonicity available. At the other extreme we may delete the monotonicity requirement entirely from the statement of Theorem 11, provided we replace (IQ) by the requirement that for all  $U \in [\Phi, \Psi]$ ,

$$\begin{aligned} L_k \varphi_k + \lambda \varphi_k &\leq h_k^i u_i + f_k(x, u_1, u_2, \dots, u_{k-1}, \varphi_k, u_{k+1}, \dots, u_m), \quad (1 \leq k \leq d), \\ B_k \varphi_k &\leq e_k^i u_i + g_k(x, u_1, u_2, \dots, u_{k-1}, \varphi_k, u_{k+1}, \dots, u_m), \quad (1 \leq k \leq d), \\ L_k \psi_k + \lambda \psi_k &\geq h_k^i u_i + f_k(x, u_1, \dots, u_{k-1}, \psi_k, u_{k+1}, \dots, u_m), \quad (\delta \leq k \leq l), \\ B_k \psi_k &\geq e_k^i u_i + g_k(x, u_1, \dots, u_{k-1}, \psi_k, u_{k+1}, \dots, u_m), \quad (\delta \leq k \leq l), \end{aligned}$$

yielding a result akin to Theorem 8 in [13]. Since we only assume that the inequalities are satisfied for  $U \in [\Phi, \Psi]$ , instead of for all  $U \in \mathcal{H}^1(\Omega)$ , this result is not just a repeated application of the theorem. We will return to this point with a remark at the end of this section.

The following lemma will be necessary for the proof of Theorem 11.

LEMMA 13. *Suppose  $u \in H^1_S(\Omega) \cap L^\infty(\Omega) \cap C(\Omega)$ ,  $G = \{x \in \Omega \mid u(x) > 0\}$ , and  $R = S \cap \partial G$ . Then the restriction of  $u$  to  $G$  is a member of  $H^1_k(G)$ .*

*Proof.* Let

$$E_k = \left\{ x \in \Omega \mid \frac{1}{k} < |u(x)| \leq \frac{2}{k} \right\}.$$

Then there must exist a subsequence  $\{k(n)\}$  of positive integers such that  $\lim_{n \rightarrow \infty} m(E_{k(n)}) = 0$ , where  $m$  is the usual Lebesgue measure on  $\Omega$ . We define

$$\xi_n(x) = \max [0, \min (1, 2 - k(n)|u(x)|)],$$

a function which is a member of  $H^1(\Omega)$ , which is equal to 0 whenever  $|u(x)| \geq 2/k(n)$ , and equal to 1 when  $|u(x)| \leq 1/k(n)$ . Moreover,

$$D_i \xi_n(x) = \begin{cases} -\operatorname{sgn}(u(x))k(n)D_i u(x) & \text{if } x \in E_{k(n)}, \\ 0 & \text{if } x \in \Omega \setminus E_{k(n)}. \end{cases}$$

One easily verifies that  $D_i \xi_n u = u D_i \xi_n + \xi_n D_i u$ . We first show that  $\xi_n u \rightarrow 0$  in  $H^1(\Omega)$ . Let

$$S_n = \{x \in \Omega \mid |u(x)| \leq 2/k(n)\};$$

then we have

$$\int_{\Omega} (\xi_n u)^2 dx \leq \int_{S_n} (\xi_n u)^2 dx \leq \frac{4m(\Omega)}{k(n)^2},$$

while

$$\begin{aligned} \int_{\Omega} (D_i \xi_n u)^2 dx &\leq 2 \int_{\Omega} [(D_i \xi_n)^2 u^2 + (D_i u)^2 \xi_n^2] dx \\ &\leq 2 \int_{E_{k(n)}} k(n)^2 (D_i u)^2 u^2 dx + 2 \int_{S_n} (D_i u)^2 dx \\ &\leq 8 \int_{E_{k(n)}} (D_i u)^2 dx + 2 \int_S (D_i u)^2 dx. \end{aligned}$$

We note that since  $m(\Omega) < \infty$ , the last term tends to  $2 \int_{S_\infty} (D_i u)^2 dx$  where  $S_\infty = u^{-1}(\{0\})$ . But by the remarks made just before the proof of Lemma 4 it follows that this integral is zero. The next-to-last term tends to zero because  $m(E_{k(n)}) \rightarrow 0$ . There



exists a function  $J : (0, \infty) \rightarrow (0, \infty)$  such that

$$\max_{1 \leq i \leq n} \int_{\sigma} (D_i u)^2 dx < \varepsilon \quad \text{whenever } m(\sigma) < J(\varepsilon).$$

Since  $u \in H^1_S(\Omega)$ , there exists a sequence  $\{u_n\} \subset H^1(\Omega)$  such that for each positive integer  $n$ , there exists an open neighborhood  $N_n$  of  $\partial\Omega \setminus S$  such that  $u_n$  vanishes on  $N_n$ . We may assume without loss of generality that there exists a positive number  $K$  such that  $|u_n(x) - u(x)| \leq K$  a.e. for  $n = 1, 2, \dots$ , and that  $|u(x) - u_n(x)| \leq 1/k(n)$  except on a set  $\sigma_n$  of measure less than  $J(1/k(n)^3)$ . Clearly  $(1 - \xi_n)u_n$ , restricted to  $G$ , is a member of  $H^1(G)$  which vanishes on a neighborhood of  $\partial G \setminus S$ . We observe that

$$u - (1 - \xi_n)u_n = -\xi_n(u - u_n) + (u - u_n) + \xi_n u,$$

where the last two terms tend to zero in  $H^1(\Omega)$ . Obviously  $\xi_n(u - u_n)$  tends to zero in the  $L^2(\Omega)$ -topology, so that we only need to examine convergence of its derivatives.

$$\begin{aligned} \int_{\Omega} \{D_i[\xi_n(u - u_n)]\}^2 dx &\leq \int_{E_{k(n)}} 2(D_i \xi_n)^2 (u - u_n)^2 dx + \int_{\Omega} 2\xi_n^2 [D_i(u - u_n)]^2 dx \\ &\leq \int_{E_{k(n)} \cap (\Omega \setminus \sigma_n)} 2(D_i u)^2 dx + \int_{E_{k(n)} \cap \sigma_n} 2k(n)^2 |D_i u|^2 K^2 dx + 2\|u - u_n\|_1^2 \\ &\leq \int_{E_{k(n)}} 2(D_i u)^2 dx + 2K^2/k(n) + 2\|u - u_n\|_1^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad \square \end{aligned}$$

*Remark.* In the proof of Theorem 11 we will use the following fact. Suppose we alter  $E$  and  $H$  by setting certain entries equal to zero. This changes the quadratic functional  $A_{\lambda}(U, U)$  to a new one,  $A_{\lambda}^*(U, U)$ . Let  $U^* = (|u_1|, |u_2|, \dots, |u_m|)$ ; then  $A_{\lambda}^*(U, U) \geq A_{\lambda}^*(U^*, U^*) \geq A_{\lambda}(U^*, U^*) \geq \varepsilon_0 \|U^*\|_1^2 - \mu_0 \|U^*\|_0^2 = \varepsilon_0 \|U\|_1^2 - \mu_0 \|U\|_1^2$ .

Hence all the results which we have proven are still true, for the same values of  $\lambda$ , for the problem obtained by setting one or more of entries  $h_k^i$  and  $e_k^i$  equal to zero.

*Proof of Theorem 11.* We will use  $u \vee v$  to denote the function  $x \rightarrow \max(u(x), v(x))$ , and if  $U = (u_1, u_2, \dots, u_m)$  and  $V = (v_1, v_2, \dots, v_m)$  then  $U \vee V = (u_1 \vee v_1, u_2 \vee v_2, \dots, u_m \vee v_m)$ . We similarly define the greatest lower bounds  $u \wedge v$  and  $U \wedge V$ . Next we introduce a notation which can be used to denote certain matrices obtained from  $H$  and  $E$  by replacing one or more columns by columns of zeros. If  $M = (m_{ij})$  is an  $m \times m$  matrix, then the matrix  $^{[k,r]}M = (^{[k,r]}m_{ij})$  is the matrix defined by

$$^{[k,r]}m_{ij} = \begin{cases} m_{ij} & \text{if } k \leq j \leq r, \\ 0 & \text{otherwise.} \end{cases}$$

Let  $S \geq 0$  be a member of  $\mathcal{H}^1(\Omega) \cap \mathcal{L}^{\infty}(\Omega)$  which we shall choose later. Next we define  $F_0 : \Omega \times R^m \rightarrow R^m$  and  $G_0 : \partial\Omega \times R^m \rightarrow R^m$  by

$$\begin{aligned} F_0(x, U) &= F(x, (U \vee \Phi) \wedge \Psi) + {}^{[1,\delta-1]}H(x) ((U \vee \Phi) \wedge S) \\ &\quad + {}^{[\delta,d]}H(x) ((U \vee \Phi) \wedge \Psi) + {}^{[d+1,d]}H(x) ((U \wedge \Psi) - U), \end{aligned}$$

and similarly

$$\begin{aligned} G_0(x, U) &= G(x, (U \vee \Phi) \wedge \Psi) + {}^{[1,\delta-1]}E(x) ((U \vee \Phi) \wedge S) \\ &\quad + {}^{[\delta,d]}E(x) ((U \vee \Phi) \wedge \Psi) + {}^{[d+1,d]}E(x) ((U \wedge \Psi) - U). \end{aligned}$$

Let  $\varphi^* = \max_{1 \leq i \leq d} \|\varphi_i\|_{0,\infty,\Omega}$ , and  $\psi^* = \max_{\delta \leq i \leq l} \|\psi_i\|_{0,\infty,\Omega}$ . We note that the first term appearing on the right-hand side of the definition of  $F_0$  is a member of  $\mathcal{F}_{q/2}(\Omega \times R^m)$ . The second term is  $\leq HS$ ; for  $U \leq (\nu, \nu, \dots, \nu)$ , where  $\nu$  is a real number, it is bounded from below by  $-H(x) (\varphi^*, \varphi^*, \dots, \varphi^*) \in \mathcal{L}^{q/2}(\Omega)$ . Hence the second term is a member of  $\mathcal{F}_{q/2}(\Omega \times R^m)$ . The third term also belongs to  $\mathcal{F}_{q/2}(\Omega \times R^m)$ , because it is bounded from above by  $H(x) (\psi^*, \psi^*, \dots, \psi^*) \in \mathcal{L}^{q/2}(\Omega)$ , and from below by  $-H(x) (\varphi^*, \varphi^*, \dots, \varphi^*) \in \mathcal{L}^{q/2}(\Omega)$ . Finally, the last term is  $\leq 0$  and, for  $U \leq (\nu, \nu, \dots, \nu)$ , is bounded from below by  $H((0, 0, \dots, 0) \wedge (-\nu - \psi^*, -\nu - \psi^*, \dots, -\nu - \psi^*))$ . Hence  $F_0 \in \mathcal{F}_{q/2}(\Omega \times R^m)$ . Similarly, we have  $G_0 \in \mathcal{F}_p(\partial\Omega \times R^m)$ . Therefore we may apply Theorem 9, which says that we have a (generalized) solution  $U_0$  to the problem

$$\begin{aligned} (L + \lambda)U_0 - {}^{[d+1,m]}HU_0 &= F_0(x, U_0) \quad \text{in } \Omega, \\ BU_0 - {}^{[d+1,m]}EU_0 &= G_0(x, U_0) \quad \text{on } \Delta, \\ U_0 &= \Theta \quad \text{on } \Gamma. \end{aligned}$$

Let  $0 \leq \bar{F} \in \mathcal{L}^{q/2}(\Omega)$  be an upper bound for the sum of the first, third and fourth terms in the definition of  $F_0$ . Similarly,  $0 \leq \bar{G} \in \mathcal{L}^p(\partial\Omega)$  is an upper bound for the sum of the first, third, and fourth terms in the definition of  $G_0$ . Let  $S$  be the solution of

$$\begin{aligned} (L + \lambda)S - HS &= \bar{F} - {}^{[d,d]}HS \quad \text{in } \Omega, \\ BS - ES &= \bar{G} - {}^{[d,d]}ES \quad \text{on } \Delta, \\ S &= \Theta \vee 0 \quad \text{on } \Gamma. \end{aligned}$$

Applying Lemma 5 to

$$\begin{aligned} (L + \lambda)(S - U_0) - {}^{[d+1,m]}H(S - U_0) &\geq 0 \quad \text{in } \Omega, \\ B(S - U_0) - {}^{[d+1,m]}E(S - U_0) &\geq 0 \quad \text{on } \Delta, \\ S - U_0 &\geq 0 \quad \text{on } \Gamma, \end{aligned}$$

we get  $S \geq U_0$ . But this means that

$$\begin{aligned} (L + \lambda)U_0 &\leq \hat{F}(x, (U_0 \vee \Phi) \wedge \Psi) \quad \text{in } \Omega, \\ BU_0 &\leq \hat{G}(x, (U_0 \vee \Phi) \wedge \Psi) \quad \text{on } \Delta, \\ U_0 &= \Theta \quad \text{on } \Gamma. \end{aligned}$$

Let  $\delta_i = \psi_i - u_{0i}$ . For each  $1 \leq i \leq m$

$$\begin{aligned} (L_i + \lambda)\psi_i &\geq \hat{f}_i(x, (U_0 \vee \Phi)^\Psi) \quad \text{in } \Omega, \\ B\psi_i &\geq \hat{g}_i(x, (U_0 \vee \Phi)^\Psi) \quad \text{on } \Delta_i, \\ \psi_i &\geq \theta_i \quad \text{on } \Gamma_i. \end{aligned}$$

Let  $G_k = \{x \mid u_{0k}(x) > \psi_k(x)\}$ . We claim that  $G_k = \emptyset$ . Suppose this were not the situation. We can show  $\delta_k \wedge 0$  is a member of  $H^1_{\Delta_k}(\Omega)$ . To see this, we first observe that since  $u_{0k} - \theta_k \in H^1_{\Delta_k}(\Omega)$ , there exists a sequence  $\{v_j\} \in H^1(\Omega)$ , such that  $v_j \rightarrow u_{0k} - \theta_k$  in  $H^1(\Omega)$ , and such that  $v_j = 0$  on a neighborhood of  $\Gamma_k$ . But then  $[\psi_k - (v_j + \theta_k)] \wedge 0 \rightarrow \delta_k \wedge 0$  in  $H^1(\Omega)$  as  $j \rightarrow \infty$ , which implies that  $\delta_k \wedge 0 \in H^1_{\Delta_k}(\Omega)$ . Applying

the lemma,  $\delta_k \in H^1_{R_k}(G)$  where  $R_k = \Delta_k \cap \partial G$ . Therefore,

$$\begin{aligned} (L_k + \lambda)\delta_k &\cong \hat{f}_k(x, u_{01} \vee \varphi_1, \dots, u_{0\delta-1} \\ &\quad \vee \varphi_{\delta-1}, \psi_\delta, \dots, \psi_l, u_{0l+1}, \dots, u_{0m}) \\ &\quad - \hat{f}_k(x, u_{01} \vee \varphi_1 \wedge \psi_1, \dots, u_{0l} \\ &\quad \quad \vee \varphi_l \wedge \psi_l, u_{0l+1}, \dots, u_{0m}) \cong 0 \text{ in } G_k, \\ B_k\delta_k &\cong \hat{g}_k(x, u_{01} \vee \varphi_1, \dots, u_{0\delta-1} \\ &\quad \vee \varphi_{\delta-1}, \psi_\delta, \dots, \psi_l, u_{0l+1}, \dots, u_{0m}) \\ &\quad - \hat{g}_k(x, u_{01} \vee \varphi_1 \wedge \psi_1, \dots, u_{0l} \\ &\quad \quad \vee \varphi_l \wedge \psi_l, u_{0l+1}, \dots, u_{0m}) \cong 0 \text{ on } R_k, \\ \delta_k &= 0 \text{ on } \partial G \setminus R_k, \end{aligned}$$

where we used the monotonicity and the fact that  $(u_{0k} \vee \varphi_k) \wedge \psi_k = \psi_k$  on  $G$ . Actually, some care must be taken to verify that the boundary condition on  $R_k$  is truly satisfied for the problem on  $G$ . To prove this we first show that if  $u \in H^1_{R_k}(G)$  then  $\bar{u} \in H^1_{R_k}(\Omega)$ , where we define  $\bar{u}$  simply as

$$\bar{u}(x) = \begin{cases} u(x) & \text{if } x \in G, \\ 0 & \text{if } x \notin G. \end{cases}$$

In order to do this we may, without loss of generality, assume that  $u(x) = 0$  on a neighborhood  $N$  of  $\partial G \setminus R_k$ . Let  $v \in \mathcal{D}(\Omega)$ , and define  $\bar{v}$  to be a function in  $\mathcal{D}(\Omega)$ , which agrees with  $v$  on  $\text{supp } u \cap \text{supp } v$  and such that  $\text{supp } \bar{v} \subset G$ . This is possible since  $\text{supp } u \cap \text{supp } v$  and  $\partial G$  are disjoint compact sets. Now

$$\int_\Omega \bar{u} D_i v \, dx = \int_G u D_i \bar{v} \, dx = - \int_G (D_i u) \bar{v} \, dx.$$

Hence, for each  $i$ ,  $D_i \bar{u}(x)$  equals  $D_i u(x)$  on  $G$ , and is 0 outside  $G$ , i.e.,  $D_i \bar{u} \in L^2(\Omega)$ . Moreover  $\bar{u} = 0$  on a neighborhood of  $\partial\Omega \setminus R_k$ , namely  $\bar{\Omega} \setminus \text{supp } u$ . Now, since  $\delta_k$  is defined on all of  $\Omega$  and satisfies  $B_k \delta_k = B_k \psi_k - B_k u_{0k}$  on  $\Delta_k$ , (in the generalized sense via  $A_\lambda$ ), it follows that the boundary condition is satisfied on  $R_k$ , i.e., via the bilinear functional  $A_\lambda$  defined on  $\mathcal{H}^1(G) \times \mathcal{H}^1(G)$ . Therefore, by Lemma 5,  $\delta_k \cong 0$  on  $G_k$ , which implies  $G_k = \emptyset$ . Hence  $U_0 \leq \Psi$ . Applying the inequalities which we know hold for  $U_{0\Phi}$ , we obtain

$$\begin{aligned} (L + \lambda)(U_0 - U_{0\Phi}) &\cong \hat{F}(x, U_0 \vee \Phi) - \hat{F}(x, U_{0\Phi}) \quad \text{in } \Omega, \\ B(U_0 - U_{0\Phi}) &\cong \hat{G}(x, U_0 \vee \Phi) - \hat{G}(x, U_{0\Phi}) \quad \text{on } \Delta, \\ (U_0 - U_{0\Phi}) &\cong 0 \quad \text{on } \Gamma. \end{aligned}$$

Using an argument entirely analogous to the one used to show that  $U_0 \leq \Psi$ , we obtain from the above inequalities the fact that  $U_0 \cong \Phi$ , thus concluding the proof of the theorem since  $U_0$  also solves (4')–(6').  $\square$

*Remark.* Suppose one has several pairs  $(\Phi^{(j)}, \Psi^{(j)})$ ,  $1 \leq j \leq r$ , as in the statement of Theorem 11, and suppose that

$$\begin{aligned} (L + \lambda)\Phi^{(j)} &\cong \hat{F}(x, U_{\Phi^{(j)}}) \quad \text{and} \quad (L + \lambda)\Psi^{(j)} \cong \hat{F}(x, U^{\Psi^{(j)}}) \quad \text{in } \Omega, \\ B\Phi^{(j)} &\cong \hat{G}(x, U_{\Phi^{(j)}}) \quad \text{and} \quad B\Psi^{(j)} \cong \hat{G}(x, U^{\Psi^{(j)}}) \quad \text{on } \Delta, \end{aligned}$$

and that  $\Phi^{(j)} \leq \Theta \leq \Psi^{(j)}$  for all  $1 \leq j \leq m$  and all  $U \in [\Phi, \Psi]$ , where  $\Phi = \Phi^{(1)} \vee$

$\Phi^{(2)} \vee \dots \vee \Phi^{(r)}$  and  $\Psi = \Psi^{(1)} \wedge \Psi^{(2)} \wedge \dots \wedge \Psi^{(r)}$ . Then there exists a solution  $U_0 \in [\Phi, \Psi]$  to (4')–(6'). To see this one merely notes that the first part of the proof of Theorem 11 still shows that there exists a solution  $U_0$  to

$$\begin{aligned} (L + \lambda)U_0 &= \hat{F}(x, (U_0 \vee \Phi) \wedge \Psi) && \text{in } \Omega, \\ BU_0 &= \hat{G}(x, (U_0 \vee \Phi) \wedge \Psi) && \text{on } \Delta, \\ U_0 &= \Theta && \text{on } \Gamma. \end{aligned}$$

Next we note that

$$(L_i + \lambda)\psi_k^{(j)} \geq \hat{F}_k(x, ((U_0 \vee \Phi) \wedge \Psi)^{j'}) \text{ in } \Omega,$$

and a corresponding inequality holds on  $\Delta$ . Letting  $\delta_k = \psi_k^{(j)} - u_{0k}$  we obtain the appropriate inequalities for  $\delta_k$  which show that  $u_{0k} \leq \psi_k^{(j)}$ . Hence  $U_0 \leq \Psi$  and similar arguments lead to the conclusion that  $U_0 \geq \Phi$ .

**4. The nonlinear parabolic problem.** We turn our attention to the system

$$(14) \quad \frac{\partial u_k}{\partial t} + L_k u_k = \hat{f}_k(x, t, U) \quad \text{in } \Omega \times (0, T),$$

$$(15) \quad B_k u_k = \hat{g}_k(x, U) \quad \text{on } \Delta_k \times (0, T),$$

$$(16) \quad u_k(x, t) = \theta_k(x) \quad \text{on } \Gamma_k \times (0, T),$$

$$(17) \quad u_k(x, 0) = u_k^0(x) \quad \text{in } \Omega.$$

We assume that the only explicit time dependence appears in the  $\hat{f}_k$ 's, although this can be generalized. For example, if the coefficients of  $L_k$  are regular enough, then we can allow time dependence in the principal coefficients without complicating matters too much. Time-dependent boundary conditions, however, seem to lead to more serious difficulties.

In order to obtain our results we shall employ the nonlinear semigroup theory of Crandall, Liggett, and Pazy [7], [8]. This seems to be appropriate for the investigation of invariant sets, since this type of semigroup ‘‘lives’’ on a closed set which does not necessarily have to be an entire Banach space. We first briefly describe the nonlinear semigroup results which will be used.

Let  $X$  be a Banach space, and for each  $t \geq 0$  let  $\mathcal{A}(t)$  be an operator from  $\mathcal{D}(t) \subset X$ , its domain, into  $X$ , which satisfies

$$\|x + \lambda \mathcal{A}(t)x - (y + \lambda \mathcal{A}(t)y)\| \geq (1 - \lambda\omega) \|x - y\|$$

for all  $x, y \in \mathcal{D}(t)$  and all  $0 < \lambda < 1/\omega$ , where  $\omega$  is some given positive number. Suppose that the closure,  $\overline{\mathcal{D}(t)}$ , of the domain is independent of time and

$$\overline{\mathcal{D}(t)} = \overline{\mathcal{D}(0)} \subset \text{Range } (I + \lambda \mathcal{A}(t)) \text{ for all } 0 \leq t < T,$$

and all  $0 < \lambda < 1/\omega$ . Finally, we suppose that  $J_\lambda(t) \equiv (I + \lambda \mathcal{A}(t))^{-1}$  satisfies

$$\|J_\lambda(t)x - J_\lambda(\tau)x\| \leq \lambda |\mu(t) - \mu(\tau)| M(\|x\|)$$

for all  $0 \leq t, \tau < T$ , and  $x \in \overline{\mathcal{D}(t)}$ , where  $\mu : [0, T] \rightarrow X$  is a continuous function of bounded variation and  $M : [0, \infty) \rightarrow [0, \infty)$  is a nondecreasing function. Under these assumptions,

$$\mathcal{U}(t, s)x \equiv \lim_{n \rightarrow \infty} \prod_{i=1}^n J_{(t-s)/n}(s + i(t-s)/n)x$$

exists for all  $x \in \overline{\mathcal{D}(0)}$ ,  $0 \leq s < t < T$ , and

$$\lim_{t \downarrow s} \mathcal{U}(t, s)x = x \quad \forall x \in \overline{\mathcal{D}(0)}.$$

$\mathcal{U}(t, s)$  is called the *propagation operator* because if  $y: [0, T] \rightarrow X$  is a continuous, strongly differentiable map satisfying

$$\frac{dy}{dt} + \mathcal{A}(t)y = 0, \quad y(s) = y_0 \in \overline{\mathcal{D}(0)},$$

then  $y(t) = \mathcal{U}(t, s)y_0$  [8, Theorem 3.1].

Our aim will be to find an invariant set which is equal to  $\overline{\mathcal{D}(0)}$  for an appropriate nonlinear semigroup. This means that we must find  $\mathcal{D}(t)$  such that (15) is satisfied. This makes it necessary to determine exactly the domain of the operator  $L$ . The difficulty in this lies in the interpretation of the boundary condition (2). Since  $a_k^{ij} \in L^\infty(\Omega)$  and  $D_i u_k \in L^2(\Omega)$ , their traces on  $\partial\Omega$  are not well defined. However we can circumvent this problem as follows. Define  $\tilde{B}$  to be the unique linear operator

$$\tilde{B} : \mathcal{V}(L) \equiv \{u \in \mathcal{H}^1(\Omega) \mid LU - HU \in \mathcal{L}^2(\Omega)\} \rightarrow \mathcal{H}^{-1/2}(\Omega),$$

which satisfies

$$A_0(U, V) - (LU - HU, V) = (\tilde{B}U, \gamma_0 V)_{\partial\Omega},$$

for all  $V \in \mathcal{H}^1(\Omega)$ . The existence of  $\tilde{B}$  is easily established via the Riesz representation theorem [2]. One can also check to see that if  $U$  and the coefficients of  $L, H, B$ , and  $E$  are sufficiently well behaved, then

$$\tilde{B}_k u_k = \nu_i [a_k^{ij} D_j u_k + d_k^i u_k] + \sigma_k u_k - e_k^i u_i,$$

where the right hand side can be evaluated pointwise. Since  $\mathcal{H}^1(\Omega)$  is a Hilbert space, there exists an orthogonal projection operator  $\pi_\Delta : \mathcal{H}^1(\Omega) \rightarrow \mathcal{H}^1(\Omega)$  with  $\pi_\Delta \mathcal{H}^1(\Omega) = \mathcal{H}_\Delta^1(\Omega)$ . Suppose  $\alpha \in \mathcal{H}^{1/2}(\partial\Omega)$ . Let  $U \in \gamma_0^{-1}(\alpha)$ , and define  $\tilde{\pi}_\Delta \alpha = \gamma_0 \pi_\Delta U \in \mathcal{H}_\Delta^{1/2}(\partial\Omega) \equiv \gamma_0 \mathcal{H}_\Delta^1(\Omega) \subset \mathcal{H}^{1/2}(\partial\Omega)$ . This is a well-defined map, since if  $\gamma_0 U = \gamma_0 U'$  then  $U - U' \in \mathcal{H}_0^1(\Omega) \subset \mathcal{H}_\Delta^1(\Omega)$ , so that  $\gamma_0 \pi_\Delta (U - U') = \gamma_0 (U - U') = 0$ . Hence we have a projection operator  $\tilde{\pi}_\Delta$  satisfying  $\tilde{\pi}_\Delta \gamma_0 = \gamma_0 \pi_\Delta$ . We also have the corresponding adjoints  $\pi_\Delta^* : \mathcal{H}^1(\Omega)' \rightarrow \mathcal{H}^1(\Omega)'$  and  $\tilde{\pi}_\Delta^* : \mathcal{H}^{-1/2}(\partial\Omega) \rightarrow \mathcal{H}^{-1/2}(\partial\Omega)$ .

LEMMA 14. Suppose  $F \in \mathcal{L}^2(\Omega)$ ,  $G \in \mathcal{H}^{-1/2}(\partial\Omega)$  and  $\Theta \in \mathcal{H}^1(\Omega)$ ; then  $U$  is a generalized solution of (1)–(3) iff

- (a)  $(L + \lambda)U - HU = F$ , (as distributions),
- (b)  $\tilde{\pi}_\Delta^* [\tilde{B}U - G] = 0$ ,
- (c)  $\pi_\Delta [U - \Theta] = U - \Theta$ .

Proof. If  $U$  is a generalized solution, then

$$A_\lambda(U, V) = (F, V) + (G, V)_{\partial\Omega} \quad \forall V \in \mathcal{H}_\Delta^1(\Omega).$$

In particular,

$$A_\lambda(U, V) = (LU + \lambda U - HU, V) = (F, V),$$

for all  $V \in \mathcal{C}^\infty(\Omega)$  with compact support in  $\Omega$ , and therefore (a) is satisfied. By the definition of  $\tilde{B}$ ,

$$A_\lambda(U, V) = (LU + \lambda U - HU, V) + (\tilde{B}U, \gamma_0 V)_{\partial\Omega} = (F, V) + (G, \gamma_0 V)_{\partial\Omega},$$

for all  $V \in \mathcal{H}_\Delta^1(\Omega)$ , and hence

$$(\tilde{B}U, \gamma_0 \pi_\Delta V)_{\partial\Omega} = (G, \gamma_0 \pi_\Delta V) \quad \forall V \in \mathcal{H}^1(\Omega),$$

so that

$$\gamma_0^* \tilde{\pi}_\Delta^* [\tilde{B}U - G] = \pi_\Delta^* \gamma_0^* [\tilde{B}U - G] = 0.$$

Since  $\gamma_0$  is surjective, hence  $\gamma_0^*$  injective, (b) follows. Because  $U - \Theta \in \mathcal{H}_\Delta^1(\Omega)$ ,  $\pi_\Delta(U - \Theta) = U - \Theta$ . Conversely, suppose (a), (b), (c) are satisfied. Obviously  $U - \Theta \in \mathcal{H}_\Delta^1(\Omega)$ . Using the definition of  $\tilde{B}$  together with (a), (b) and the fact that  $\mathcal{D}(\Omega)$  is dense in  $L^2(\Omega)$  yields

$$A_\lambda(U, V) = (F, V) + (G, \gamma_0 V)_{\partial\Omega} \quad \forall V \in \mathcal{H}_\Delta^1(\Omega). \quad \square$$

Returning to the problem (14)–(17), we see that the above lemma implies that

$$\mathcal{A}(t) : U \rightarrow LU - \hat{F}(x, t, U)$$

is a well-defined operator from

$$\{U \in \mathcal{H}^1(\Omega) \mid LU \in \mathcal{L}^2(\Omega), \pi_\Delta(U - \Theta) = U - \Theta, \hat{F}(x, t, U) \in \mathcal{L}^2(\Omega), \\ \tilde{\pi}_\Delta^* [\tilde{B}U - j\hat{G}(x, U)] = 0\}$$

into  $\mathcal{L}^2(\Omega)$ , where  $j : \mathcal{L}^p(\partial\Omega) \subset \mathcal{H}^{-1/2}(\partial\Omega)$ . The closure of this set will, if the coefficients of  $L$  and  $B$  are “nice”, be all of  $\mathcal{L}^2(\Omega)$ . In order to obtain invariant sets of the type described in the Introduction, we will instead define the domain by

$$\mathcal{D} = \{U \in \mathcal{H}^1(\Omega) \mid LU \in \mathcal{L}^2(\Omega), \pi_\Delta(U - \Theta) = U - \Theta, \tilde{\pi}_\Delta^* [\tilde{B}U - j\hat{G}(x, U)] \\ = 0, \hat{F}(x, t, U) \in \mathcal{L}^2(\Omega), \Phi \leq U \leq \Psi\}.$$

The following hypothesis will also be needed.

(VI).

(i) There exists a constant  $K > 0$  such that

$$|\hat{F}(x, t, U) - \hat{F}(x, \tau, U)| \leq K|U| |t - \tau|.$$

We also assume that  $F$  satisfies (V) with  $\gamma_1$  independent of  $t$ , and  $\hat{F} \in \mathcal{F}_2(\mathcal{S}\Psi)$ .

(ii) There exists a collection  $\mathcal{D}_{00} \subset (\mathcal{D}(\Omega))^m$  such that  $\mathcal{D}_{00}$  is a dense subset (with respect to the  $\mathcal{L}^2(\Omega)$ -topology) of

$$\mathcal{D}_0 \equiv \{U \in \mathcal{H}^1(\Omega) \mid LU - HU \in \mathcal{L}^2(\Omega), \pi_\Delta U = U, \tilde{\pi}_\Delta^* \tilde{B}U = 0\}.$$

Condition (i) is more restrictive than needed. Condition (ii) is a technical necessity which can be replaced by additional regularity requirements on the coefficients. For example, if  $a_k^j$  and  $d_k^i$  are of class  $C^1(\Omega)$  then we can set  $\mathcal{D}_{00} = (\mathcal{D}(\Omega))^m$ . If these coefficients are sectionally  $C^1$  with discontinuities across surfaces in  $\Omega$  whose union  $\Gamma$  has a closure whose  $n$ -dimensional measure is zero, then  $\mathcal{D}_{00}$  can be taken to be the collection of all  $\mathcal{C}^\infty(\Omega)$  functions with compact support in  $\Omega \setminus \bar{\Gamma}$ .

Please recall that  $\hat{F}(x, t, U) = H(x, t)U + F(x, t, U)$  and  $\hat{G}(x, U) = E(x)U + G(x, U)$ .

LEMMA 15. Suppose (I)–(VI) are satisfied,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ , and for each fixed  $t \in [0, T)$ , with  $T < \infty$ , we have  $F \in \mathcal{F}_{q_2}(\mathcal{S}\Psi)$  and  $G \in \mathcal{F}_p(\partial\mathcal{S}\Psi)$ . For  $1 \leq i \leq l$ , we suppose  $\hat{f}_i(x, t, u_1, u_2, \dots, u_m)$  and  $\hat{g}_i(x, u_1, u_2, \dots, u_m)$  are nondecreasing in  $u_j$  for all  $j \neq i$ ,  $1 \leq j \leq l$ . Finally suppose that  $\Phi \leq \Theta \leq \Psi$ , and that for all  $U \in$

$[\Phi, \Psi]$ ,

$$\begin{aligned} L\Phi &\cong \hat{F}(x, t, U_\Phi) \quad \text{and} \quad L\Psi \cong \hat{F}(x, t, U^\Psi) \quad \text{in } \Omega \times [0, T), \\ B\Phi &\cong \hat{G}(x, U_\Phi) \quad \text{and} \quad B\Psi \cong \hat{G}(x, U^\Psi) \quad \text{on } \Delta \times [0, T). \end{aligned}$$

Then  $\mathcal{A}(t)$  satisfies:

$$(i) \quad \|(U + \lambda\mathcal{A}(t)U) - (V + \lambda\mathcal{A}(t)V)\|_0 \cong (1 - \lambda\omega)\|U - V\|_0,$$

for all  $U, V \in \mathcal{D}$  and all  $0 < \lambda < \omega^{-1}$ , where  $\omega$  is some fixed positive number.

(ii) The  $\mathcal{L}^2(\Omega)$ -closure of  $\mathcal{D}$  is  $\overline{\mathcal{D}} = \{U \in \mathcal{L}^2(\Omega) \mid \Phi \cong U \cong \Psi\}$ , a set we will denote by  $[\overline{\Phi}, \overline{\Psi}]$ .

(iii) Range  $(I + \lambda\mathcal{A}(t)) \supset [\overline{\Phi}, \overline{\Psi}]$

(iv) If  $W(t) + \lambda\mathcal{A}(t)W(t) = F_0 \in [\overline{\Phi}, \overline{\Psi}]$  (i.e.  $W(t) = J_\lambda(t)F_0$ ), then

$$\|W(t_1) - W(t_2)\|_0 \cong C_\lambda |t_1 - t_2| (\|F\|_0 + 1),$$

for all  $t_1, t_2 \in [0, T)$ , where  $C$  is a constant.

*Proof.*

$$\begin{aligned} (i) \quad &\|U - V\|_0 \|U + \lambda\mathcal{A}(t)U - V - \lambda\mathcal{A}(t)V\|_0 \\ &\cong (U - V, U - V) + \lambda A_0(U - V, U - V) \\ &\quad - \lambda(F(x, t, U) - F(x, t, V), U - V) \\ &\quad - \lambda(G(x, U) - G(x, V), U - V)_{\partial\Omega} \\ &\cong [1 - \lambda(\mu_0 + \gamma m(\Omega))] \|U - V\|_0^2, \end{aligned}$$

where we use (VI), Lemma 2, and the definition of  $\gamma$ .

(ii) Let  $U$  be the generalized solution of

$$\begin{aligned} (L + \lambda)U &= \hat{F}(x, t, U) \quad \text{in } \Omega, \quad (\lambda > \bar{\mu}), \\ BU &= \hat{G}(x, U) \quad \text{on } \Delta, \\ U &= \Theta \quad \text{on } \Gamma. \end{aligned}$$

Suppose  $W$  is any element in  $\mathcal{D}_{00}$  such that  $\Phi \cong U + W \cong \Psi$ . Since  $W \equiv 0$ , on some neighborhood of  $\partial\Omega$  we have

$$\begin{aligned} (a) \quad &(L + \lambda - H)(U + W) = F(x, t, U) + (L + \lambda - H)W \in \mathcal{L}^2(\Omega), \\ &\hat{F}(x, t, U + W) \in \mathcal{L}^2(\Omega), \\ (b) \quad &\tilde{\pi}_\Delta^* [\tilde{B}(U + W) - j\hat{G}(x, \gamma_0(U + W))] = \tilde{\pi}_\Delta^* [\tilde{B}U - j\hat{G}(x, \gamma_0 U)] = 0, \\ (c) \quad &\pi_\Delta(U + W - \Theta) = W + \pi_\Delta(U - \Theta) = W + U - \Theta. \end{aligned}$$

Hence  $\mathcal{D} \supset \{U + W \mid W \in \mathcal{D}_{00}, \Phi \cong U + W \cong \Psi\}$ , which upon taking closure with respect to the  $\mathcal{L}^2(\Omega)$ -topology yields

$$\overline{\mathcal{D}} \supset \{U + W \mid W \in \mathcal{L}^2(\Omega), \Phi \cong U + W \cong \Psi\} = [\overline{\Phi}, \overline{\Psi}].$$

(iii) Let  $F_0 \in [\overline{\Phi}, \overline{\Psi}]$ , and consider

$$\begin{aligned} \frac{1}{\lambda} U + LU &= \hat{F}(x, t, U) + \frac{1}{\lambda} F_0 \quad \text{in } \Omega, \\ BU &= \hat{G}(x, U) \quad \text{on } \Delta, \\ U &= \Theta \quad \text{on } \Gamma. \end{aligned}$$

By Theorem 11 we have a generalized solution  $U \in [\Phi, \Psi]$  if  $1 - (\bar{\mu} + \gamma)\lambda > 0$ .

(iv) Let  $W_i + \lambda \mathcal{A}(t_i)W_i = F_i$ ,  $(i = 1, 2)$ . Then

$$\begin{aligned} & \left(\frac{1}{\lambda} - \mu_0\right) \|W_1 + W_2\|_0^2 + \varepsilon_0 \|W_1 - W_2\|_1^2 \leq A_{1/\lambda}(W_1 - W_2, W_1 - W_2) \\ & \leq (F(x, t_1, W_1) - F(x, t_2, W_1), W_1 - W_2) \\ & \quad + (F(x, t_2, W_1) - F(x, t_2, W_2), W_1 - W_2) \\ & \quad + (G(x, W_1) - G(x, W_2), W_1 - W_2)_\Delta + \frac{1}{\lambda} (F_1 - F_2, W_1 - W_2) \\ & \leq C' \{ |t_1 - t_2| \|W_1\|_0 \|W_1 - W_2\|_0 + \|W_1 - W_2\|_0^2 + \|W_1 - W_2\|_{0,2,\partial\Omega}^2 \} \\ & \quad + \frac{1}{\lambda} \|F_1 - F_2\|_0 \|W_1 - W_2\|_0. \end{aligned}$$

Using Lemma 2, we see that there exists a constant  $C'(\varepsilon_0)$  such that the above inequality implies

$$\begin{aligned} (18) \quad & \left(\frac{1}{\lambda} - \mu_0 - C'(\varepsilon_0)\right) \|W_1 - W_2\|_0 + \varepsilon_0/2 \|W_1 - W_2\|_1 \\ & \leq C' |t_1 - t_2| \|W_1\|_0 + \frac{1}{\lambda} \|F_1 - F_2\|_0. \end{aligned}$$

First we let  $t_2 = t_0$ , some fixed value in  $(0, T)$ , and  $W_2 = W_0$ , the solution corresponding to the case where  $F_2 = 0$ . We then obtain

$$\left(\frac{1}{\lambda} - \mu_0 - C'(\varepsilon_0)\right) \|W_1\| \leq \left(\frac{1}{\lambda} - \mu_0 - C'(\varepsilon_0)\right) \|W_0\| + C' T \|W_1\| + \frac{1}{\lambda} \|F_1\|,$$

or

$$\|W_1\| \leq \frac{(1 - \lambda[\mu_0 + C'(\varepsilon_0)])\|W_0\| + \|F_1\|}{1 - \lambda_0[\mu_0 + C'(\varepsilon_0) + C' T]}, \quad (\lambda < \lambda_0),$$

where  $\lambda_0$  is chosen so small that the denominator is larger than  $\frac{1}{2}$ . Hence we have

$$\|W_1\| \leq 2\|W_0\| + 2\|F_1\|.$$

Returning to inequality (18) and setting  $F_1 = F_2 = F$ , we obtain, for  $0 < \lambda < \lambda_0$ ,

$$\|J_\lambda(t_1)F - J_\lambda(t_2)F\| = \|W_1 - W_2\| \leq 4\lambda C' |t_1 - t_2| \{\|W_0\| + \|F\|\}.$$

This concludes the proof of the lemma which guarantees the existence of a propagation operator for a nonlinear semigroup on  $[\overline{\Phi}, \overline{\Psi}]$  generated by  $\mathcal{A}$ .  $\square$

DEFINITIONS.  $U : [0, T) \rightarrow \mathcal{L}^2(\Omega)$  is called a *strong solution* of (14)–(17) if

(a)  $U$  is continuous on  $[0, T)$  and  $U(0) = U^0$ .

(b)  $U$  is absolutely continuous on compact subsets of  $(0, T)$ .

(c)  $U$  is differentiable almost everywhere on  $(0, T)$ , and is a generalized solution of (14)–(16) (regarded as an elliptic problem) for almost all  $t \in (0, T)$ . A subset  $\mathcal{S} \subset \mathcal{L}^2(\Omega)$  is called an invariant set for (14)–(16) if  $U(t) \in \mathcal{S}$  for all  $t \in (s, T)$  whenever  $U$  is a strong solution of (14)–(17) with  $U(s) \in \mathcal{S}$ .

THEOREM 16. *Suppose the hypotheses of Lemma 15 are satisfied (except that  $T = \infty$  is allowed). Then there exists a propagation operator  $\mathcal{U}(t, s)$  defined on  $[\overline{\Phi}, \overline{\Psi}]$  corresponding to problem (14)–(16). In particular,  $[\overline{\Phi}, \overline{\Psi}]$  is an invariant set*



for this problem. Moreover, if the graph of  $\mathcal{A}(t)$  is closed, then  $\mathcal{U}(t, 0)U^0$  is a strong solution for each  $U^0 \in \mathcal{D}$ .

*Proof.* The existence of  $\mathcal{U}(t, s)$  follows from the lemma. By Theorem 3.1 in [8], any strong solution  $U$  of (14)–(17) with  $U(s) \in [\overline{\Phi}, \overline{\Psi}]$  must satisfy  $U(t) = \mathcal{U}(t, s)U^0$ ,  $t \geq s$ , and hence  $[\overline{\Phi}, \overline{\Psi}]$  is an invariant set. The last assertion of the theorem follows from Theorem 3.4 in [8].  $\square$

As in the elliptic case, there are various possible corollaries we could state. One such result was stated in the introduction. We shall state two more.

**COROLLARY 17.** *Suppose (I)–(VI) are satisfied,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $\Phi$  and  $\Psi$  are members of  $\mathcal{H}^1(\Omega) \cap \mathcal{C}(\Omega) \cap \mathcal{L}^\infty(\Omega)$ ,  $F$  and  $G$  are continuous on  $\overline{\Omega} \times R^{m+1}$  and  $\overline{\Omega} \times R^m$  respectively, and for each  $1 \leq k \leq m$ ,  $\hat{f}_k(x, t, u_1, u_2, \dots, u_m)$  and  $\hat{g}_k(x, u_1, u_2, \dots, u_m)$  are nondecreasing in  $u_j$  for  $j \neq k$ . Suppose  $\Phi \leq \Theta \leq \Psi$  and*

$$\begin{aligned} L\Phi &\leq \hat{F}(x, t, \Phi) \quad \text{and} \quad L\Psi \geq \hat{F}(x, t, \Psi) \quad \text{in } \Omega \times [0, T), \\ B\Phi &\leq \hat{G}(x, \Phi) \quad \text{and} \quad B\Psi \geq \hat{G}(x, \Psi) \quad \text{on } \Delta \times [0, T). \end{aligned}$$

Then  $[\overline{\Phi}, \overline{\Psi}]$  is an invariant set, and  $\mathcal{U}(t, 0)U^0$  is a strong solution whenever  $U^0 \in \mathcal{D} \subset [\Phi, \Psi]$ .

*Proof.* We only have to establish that  $\mathcal{A}(t)$  is closed. Suppose  $U_n \in \mathcal{D}$  and  $U_n \rightarrow U$  in  $\mathcal{L}^2(\Omega)$ , and  $\mathcal{A}(t)U_n \equiv F_n \rightarrow F$  in  $\mathcal{L}^2(\Omega)$ . Using (18) one easily sees that this means  $U_n \rightarrow U$  in  $\mathcal{H}^1(\Omega)$  and hence  $\gamma_0 U_n \rightarrow \gamma_0 U$  in  $\mathcal{L}^2(\partial\Omega)$ . But since  $F$  and  $G$  satisfy the Caratheodory condition, for each  $t$  this means  $F(x, t, U_n) \rightarrow F(x, t, U)$  in  $\mathcal{L}^{q_1}(\Omega)$ , while  $G(x, U_n) \rightarrow G(x, U)$  in  $\mathcal{L}^{p_1}(\partial\Omega)$ . To see this we use the fact that the  $U_n$ 's are uniformly bounded and a standard continuity results for Nemytskii operators [14, p. 22]. Using (18) once again, we see that  $J_\lambda(t)$ , and hence  $\mathcal{A}(t)$ , is closed.  $\square$

We also have the following result for the case where we have no monotonicity requirement on the coupling.

**COROLLARY 18.** *Suppose (I)–(VI) are satisfied,  $\Theta \in \mathcal{H}^1(\Omega) \cap \mathcal{L}^\infty(\Omega)$ , and for each fixed  $t \in [0, T)$  we have  $F \in \mathcal{F}_{q_2}(\mathcal{S}_\Phi^{\mathcal{Y}})$  and  $G \in \mathcal{F}_p(\partial\mathcal{S}_\Phi^{\mathcal{Y}})$ , where  $\Phi \leq \Theta \leq \Psi$ . Then  $[\overline{\Phi}, \overline{\Psi}]$  is an invariant set provided that for all  $U \in [\Phi, \Psi]$  we have*

$$\begin{aligned} L_k \varphi_k &\leq \hat{f}_k(x, u_1, u_2, \dots, u_{k-1}, \varphi_k, u_{k+1}, \dots, u_m), & (1 \leq k \leq d), \\ B_k \varphi_k &\leq \hat{g}_k(x, u_1, u_2, \dots, u_{k-1}, \varphi_k, u_{k+1}, \dots, u_m), & (1 \leq k \leq d), \\ L_k \psi_k &\geq \hat{f}_k(x, u_1, u_2, \dots, u_{k-1}, \psi_k, u_{k+1}, \dots, u_m), & (\delta \leq k \leq l), \\ B_k \psi_k &\geq \hat{g}_k(x, u_1, u_2, \dots, u_{k-1}, \psi_k, u_{k+1}, \dots, u_m), & (\delta \leq k \leq l). \end{aligned}$$

These inequalities are essentially requirements that the ‘‘velocity’’ on the ‘‘faces’’  $\{U | u_k = \varphi_k\}$  and  $\{U | u_k = \psi_k\}$  is in the right direction. If one has monotonicity, this ‘‘velocity’’ only needs to be checked at the ‘‘edges’’  $\{U | u_i = \varphi_i, 1 \leq i \leq d\}$  and  $\{U | u_i = \psi_i, \delta \leq i \leq l\}$  (the statement of the theorem) while in the extreme case of totally monotonic coupling (Corollary 17) we only need to check the ‘‘velocities’’ at the ‘‘vertices’’  $\Phi$  and  $\Psi$ .

*Proof.* By the remark at the end of the section on elliptic systems we see that it suffices for the inequalities to hold for all  $U \in [\Phi, \Psi]$ . Hence part (iii) of Lemma 15 is still true. The other parts of Lemma 15 are obviously also still true since the relevant hypotheses are those which this lemma and Theorem 11 have in common. Therefore the proof of Theorem 16 again applies here.  $\square$

The results on invariant sets are of course still true for the case where  $F$  depends on the first order partial derivatives of  $U$  with respect to  $x$  ( $F = F(x, t, U, \nabla U)$ ), provided we know that the solution  $u$  is sufficiently regular and the map  $W \rightarrow F(x, t,$

$W, \nabla U(x, t)$ ) satisfies the hypotheses of Theorem 16. Furthermore, Theorem 16 can be used to derive results on the existence of *positively invariant sets* (using the terminology of [4]) for classical solutions to problem (14)–(17).

In conclusion, we mention that these results can be generalized to problems involving even more general, but still time-independent, boundary conditions on Lipschitz continuous boundaries. We can also allow time dependence in the elliptic operators  $L_k$ , provided the coefficients are sufficiently regular. This is done by applying the full power of the semigroup results in [8].

**Appendix.**

**THEOREM.** *Suppose that  $u \in H^1(\Omega)$  and that for all  $v$  in  $H^1(\Omega)$ ,*

$$a_\lambda(u, v) \equiv \int_\Omega [a^{ij}(D_j u) (D_i v) + d^i u D_i v + b^i v D_i u + (c + \lambda) uv] dx = \int_{\partial\Omega} g v dS,$$

where we assume that **(I)**–**(III)** are satisfied, (since we are dealing with the one-component case,  $m = 1$ , the subscript  $k$  is deleted), that  $\lambda > \bar{\mu}$ , and  $g \in L^p(\partial\Omega)$ . Then  $u \in L^\infty(\Omega)$ .

**LEMMA.** ([19, p. 24]). *Let  $\zeta = \zeta(t)$  be a nonnegative, nonincreasing function on the half-line  $t \geq 0$  such that there are positive constants  $C, \alpha$  and  $\beta$  such that*

$$\zeta(h) \leq C(h - k)^{-\alpha} [\zeta(k)]^\beta \quad \text{for } h > k \geq 0.$$

Then, if  $\beta > 1$ , there exists a constant  $d \geq 0$  such that  $\zeta(d) = 0$ , (e.g.,  $d = C^{1/\alpha} [\zeta(0)]^{(\beta-1)/\alpha} 2^{\beta(\beta-1)}$ ).

*Proof of the theorem.* Let  $v = (\text{sgn } u) \max(|u| - k, 0) = (u - k) \vee 0 + (u + k) \wedge 0 \in H^1(\Omega)$ . We have, letting  $E(k) = \{x \in \bar{\Omega} \mid |u(x)| \geq k\}$ ,

$$\begin{aligned} a_\lambda(u, v) &= \left( \int_{E(k)} + \int_{\Omega \setminus E(k)} \right) ((a^{ij} D_j u + d^i u) D_i v + (b^i D_i u + (c + \lambda) u) v) dx \\ &= \int_{E(k)} ((a^{ij} D_j v + d^i v) D_i v + v b^i D_i v + (c + \lambda) v^2) dx \\ &\quad + k \int_{\{u(x) \geq k\}} (d^i D_i v + c v + \lambda v) dx \\ &\quad - k \int_{\{u(x) \leq -k\}} (d^i D_i v + c v + \lambda v) dx \\ &= a_\lambda(v, v) + k \int_{E(k)} (d^i D_i |v| + c |v| + \lambda |v|) dx \\ &= a_\lambda(v, v) + k \int_\Omega (d^i D_i |v| + c |v| + \lambda |v|) dx \\ &= a_\lambda(v, v) + k \int_\Omega (c + \lambda - D_i d^i) |v| + k \int_{\partial\Omega} \nu_i d^i |v| dS. \end{aligned}$$

Hence  $a_\lambda(u, v) \geq a_\lambda(v, v)$ . Also there exists a constant  $K \geq 0$  such that

$$\|v\|_1^2 \leq K a_\lambda(v, v) \leq K a_\lambda(u, v) = K \int_{\partial\Omega} g v dS.$$

If we set  $F(k) = E(k) \cap \partial\Omega$ , then

$$\int_{\partial\Omega} g v dS = \int_{F(k)} g v dS \leq \|g\|_{0,r,F(k)} \|\gamma_0 v\|_{0,\rho,F(k)} \leq C_0 \|g\|_{0,r,F(k)} \|v\|_1,$$

where  $\rho = 2(n - 1)/(n - 2)$  and  $r = 2(n - 1)/n$ . We note that  $p > r$ , hence another application of Hölder's inequality yields

$$\|g\|_{0,r,F(k)} \leq \|g\|_{0,p,F(k)} m(F(k))^{1/r-1/p}.$$

Using the Sobolev inequality  $\|\gamma_0 v\|_{0,\rho,\partial\Omega} \leq C_0 \|v\|_1$ , and the fact that  $v = 0$  on  $\partial\Omega \setminus F(k)$ , we have

$$\|\gamma_0 v\|_{0,\rho,F(k)} \leq KC_0^2 \|g\|_{0,p,\partial\Omega} m(F(k))^{1/r-1/p}.$$

Hence if  $h > k$  then

$$m(F(h))^{1/\rho}(h - k) \leq \|\gamma_0 v\|_{0,\rho,F(k)} \leq KC_0^2 \|g\|_{0,p,\partial\Omega} m(F(k))^{1/r-1/p}.$$

Letting  $\zeta(h) = m(F(h))$ , we have, for  $h > k > 0$ ,

$$\zeta(h) \leq (KC_0^2 \|g\|_{0,p,\partial\Omega})^\rho (h - k)^{-\rho} \zeta(k)^{(1/r-1/p)\rho}.$$

An application of the lemma concludes the proof.  $\square$

#### REFERENCES

- [1] H. AMANN, *Invariant sets and existence theorems for semilinear parabolic and elliptic systems*, J. Math. Anal. Appl., 65 (1978), pp. 432–467.
- [2] J.-P. AUBIN, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.
- [3] H. S. CARSLAW AND J. C. JAEGER, *Conduction of Heat in Solids*, Oxford University Press, London, 1959.
- [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. 373–392.
- [5] E. D. CONWAY AND J. A. SMOLLER, *Diffusion and predator-prey interaction*, SIAM J. Appl. Math., 33 (1977), pp. 673–686.
- [6] ———, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.
- [7] M. G. CRANDALL AND T. LIGGETT, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 113 (1971), pp. 265–298.
- [8] M. G. CRANDALL AND A. PAZY, *Nonlinear evolution equations in Banach spaces*, Israel J. Math., 11 (1972), pp. 57–100.
- [9] R. FIORENZA, *Sulla Hölderianità delle soluzioni dei problemi di derivata obliqua regolare del secondo ordine*, Ricerche Mat., 14 (1965), pp. 102–123.
- [10] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, 1969.
- [11] G. R. GAVALAS, *Nonlinear Diffusion Equations of Chemically reacting Systems*, Springer, New York, 1968.
- [12] H. J. KUIPER, *Some nonlinear boundary value problems*, SIAM J. Math. Anal., 7 (1976), pp. 551–564.
- [13] ———, *Existence and comparison theorems for nonlinear diffusion systems*, J. Math. Anal. Appl., 60 (1977), pp. 166–181.
- [14] M. A. KRASNOSELSKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Macmillan, New York, 1964.
- [15] O. A. LADYZHENSKAYA AND N. N. URALTSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [16] J. L. LIONS AND E. MAGENES, *Problemi ai limiti non omogenei (III)*, Ann. Scuola Norm. Sup. Pisa, (3) 15 (1961), pp. 41–103.
- [17] ———, *Problemi ai limiti non omogenei (V)*, Annali Scuola Norm. Sup. Pisa, 16 (1962), pp. 1–44.
- [18] ———, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer, New York, 1972.
- [19] M. K. V. MURTHY AND G. STAMPACCHIA, *Boundary value problems for some degenerate- elliptic operators*, Ann. Mat. Pura. Appl., 80 (1968), pp. 1–122.
- [20] J. NEČAS, *Les méthodes directes en théorie des equations elliptiques*, Masson, Paris, 1967.

- [21] G. STAMPACCHIA, *Equations elliptiques à données discontinues*, Seminaire Schwartz, 5<sup>e</sup> année, Faculté des Sciences, Paris, 1960–61.
- [22] ———, *Equations elliptiques du second ordre à coefficients discontinus*, Les Presses de l'Université de Montréal, Montreal, 1966.
- [23] W. WALTER, *Differential and Integral Inequalities*, Springer, New York, 1970.
- [24] N. WEINBERGER, *Invariant sets for weakly coupled parabolic and elliptic systems*, *Rend. Mat.*, 8 (1975), pp. 295–310.
- [25] S. A. WILLIAMS AND P.-L. CHOW, *Nonlinear reaction-diffusion models for interacting populations*, *J. Math. Anal. Appl.*, 62 (1978), pp. 157–169.

## A UNIQUENESS THEOREM FOR HELMHOLTZ' EQUATION: PENETRABLE MEDIA WITH AN INFINITE INTERFACE\*

GERHARD KRISTENSSON†

**Abstract.** In this paper we will prove the uniqueness of a solution to Helmholtz' equation for two halfspaces of different media in  $n$  dimensions. The theorem allows a finite number of bounded inhomogeneities in each halfspace. The surface separating the halfspaces is assumed to be a cone of arbitrary cross section far away from the origin and is furthermore assumed to be smooth. We assume all space to be lossless, and in each halfspace we assume a radiation condition to be fulfilled. The boundary conditions at the interface are a general coupling in the field and its normal derivative with constant coefficients.

**1. Introduction.** The first uniqueness theorem for Helmholtz' equation for the exterior problem was shown by A. Sommerfeld [23]. In the exterior problem the field outside a bounded surface  $S$  satisfies

$$(1.1) \quad (\nabla^2 + k^2)\psi = 0.$$

Here  $k$  is a real or complex constant, and at the surface  $S$  certain boundary conditions are assumed to be satisfied. To obtain a well-defined problem he introduced a radiation condition for large distances from the obstacle—a boundary value at infinity. The solutions of (1.1) separate in two classes, satisfying either

$$\frac{\partial\psi}{\partial r} - ik\psi = o(r^{-1}), \quad r \rightarrow \infty,$$

or

$$\frac{\partial\psi}{\partial r} + ik\psi = o(r^{-1}), \quad r \rightarrow \infty.$$

The first class holds for the outgoing spherical waves (if we take the conventional time dependence to be  $e^{-i\omega t}$ ) while the second is satisfied by the ingoing spherical waves. From potential theory this property was unfamiliar, and in his paper Sommerfeld clarifies the differences between the static and the wave solution. He adopts the outgoing spherical wave, and the radiation condition thus reads

$$(1.2) \quad \frac{\partial\psi}{\partial r} - ik\psi = o(r^{-1})$$

uniformly in all angles as  $r \rightarrow \infty$ . An additional condition for large distances was also introduced:

$$(1.3) \quad \psi = O(r^{-1}), \quad r \rightarrow \infty.$$

This is the "condition of finiteness" which was later proved to be superfluous by W. Magnus [13]. In a number of papers [2], [12], [14], [21], [29] the results have been sharpened and also generalized to an arbitrary number of dimensions. Some of the papers use a slightly weakened form of Sommerfeld's radiation condition, first found

---

\*Received by the editors July 19, 1979, and in final revised form January 23, 1980.

†Institute of Theoretical Physics, S-41296 Göteborg, Sweden. This work was supported by the National Swedish Board for Technical Development (STU).

in Rellich [21]:

$$(1.4) \quad \iint_{\Sigma(r)} \left| \frac{\partial \psi}{\partial r} - ik\psi \right|^2 dS = o(1), \quad r \rightarrow \infty.$$

Here  $\Sigma(r)$  is a large sphere of radius  $r$ .

If the bounding surface  $S$  is infinite, we then have a more limited number of results. The pioneer paper is [21], which establishes uniqueness for the solution for Helmholtz' equation in  $R^n$ , when the infinite surface  $S$  intersects the plane  $x_n = \text{constant}$  for each  $x_n$  and furthermore

$$(1.5) \quad \hat{\nu} \cdot \hat{x}_n < 0 \quad \text{on } S,$$

where  $\hat{\nu}$  is the normal into the exterior of the volume considered. The radiation condition to be satisfied at infinity is a modified version of (1.4):  $\Sigma$  is now a plane  $x_n = \text{constant}$ , and the radial derivative is replaced by  $(\partial\psi/\partial x_n)$ . The importance of (1.5) is also discussed, and an example proving the non-uniqueness of a solution for Helmholtz' equation for a geometry violating (1.5) is given. Further results are given in [22].

Additional results for boundary value problems where the surface  $S$  is infinite are given in [7], [16], [19], [20]. D. S. Jones [7] gives a uniqueness theorem for surfaces which for large distances are cones of arbitrary cross section; these results are extended by F. M. Odeh [19] who shows uniqueness if  $(\partial r/\partial \nu) \leq 0$  on the surface for large  $r$ . By analytic arguments W. L. Miranker [16] shows uniqueness results for domains in which a cone with an angle greater than  $\pi/2$  can be inscribed, but certain restrictions which must be introduced on the normal derivative make the result less general. A two-dimensional formulation is found in [20], where the infinite boundary is a straight line for large  $r$ . A number of Russian authors [3], [4], [5], [24], have also studied various aspects of the problem, mostly extensions to differential equations of more general elliptic type, and in the limiting case where the losses vanish. Some results for boundary value problems with infinite boundary for a general type of elliptic differential equation have recently been published by V. Vogelsang [25], [26]. These theorems are essentially extensions of the results of Rellich.

The geometry in all these theorems proving uniqueness for Helmholtz' equation with an infinite boundary is such that the surface gets wider for increasing  $r$ . This is achieved by assuming conical shapes or by assuming that (1.5) is satisfied. This guarantees that the energy radiates to infinity as required by the radiation condition, e.g., (1.4).

The results for Helmholtz' equation in infinite domains are, as may be seen from the brief review above, both diverse and comprehensive. Uniqueness is established for many situations of interest in applications for both finite or infinite bounding surfaces as well as for real or complex wave numbers. Now focusing on geometrics with penetrable media, we find that the results here are very scarce. Werner [28] has analyzed the uniqueness of the solution for Helmholtz' equation in the case where we have penetrable obstacles of finite extension. A very specialized situation where the surface is infinite is found in [19]. Odeh here analyzes two halfspaces separated by a flat interface. The boundary conditions on the interface are very restricted, e.g., continuity in the field  $\psi$  and  $k^2\partial\psi/\partial n$ , but the result holds for real wave numbers.

The aim of this paper is to derive a uniqueness theorem for penetrable media for a more general geometry in the lossless case, and for a more general type of boundary conditions compared to Odeh [19]. In §2 we will give the principal definitions and

symbols used in this paper. The lemmas and the uniqueness theorem are proved in §3, while conclusions and a discussion of applications are found in §4.

**2. Principal definitions and notations.** In this section we will define the notations found in this paper and state the problem more precisely.

Consider two infinite halfspaces  $V_1$  and  $V_2$  in  $R^n$  (the radial distance is defined in the usual way as  $r^2 = \sum_{j=1}^n x_j^2$ ) separated by an infinite surface  $S$  as depicted in Fig. 1. We will assume the interface  $S$  to be sufficiently smooth, so that an application of Green's theorem at every finite part of  $V_1$  and  $V_2$  is valid. Sufficient conditions for this to hold are found, e.g., in [1], [9], [17]. The volumes  $V_1$  and  $V_2$  are assumed to be homogeneous and isotropic with wave numbers  $k_1$  and  $k_2$  respectively, except for a finite number of inhomogeneities  $V'_1$  and  $V'_2$  (if several inhomogeneities are present, let  $V'_1$  and  $V'_2$  be a notation for the sum of obstacles in each volume respectively, even though the boundary conditions and wave numbers may differ). For simplicity we take  $V'_1$  and  $V'_2$  homogeneous and isotropic (wave numbers  $k'_1$  and  $k'_2$ ) and bounded by  $S_1$  and  $S_2$  respectively. Let  $O$  be an arbitrarily chosen origin (this will be specified later), and let  $V_1(R)$  and  $V_2(R)$  denote the interior of a hypersphere centered at the origin, of radius  $R$ , in  $V_1$  and  $V_2$ , respectively. The portion of the hypersphere in  $V_1$  is denoted  $\Sigma_1(R)$  and  $\Sigma_2(R)$  is defined similarly. The intersection of the hypersphere and  $S$  is called  $C(R)$ , and the part of  $S$  enclosed by the hypersphere is denoted  $S(R)$ . The normal  $\hat{\nu}$  of  $S$  is directed into  $V_1$  while the normals  $\hat{\nu}$  for  $S_1$  and  $S_2$  are conventionally taken as directed outwards.

We will assume all space to be sourcefree, since in proving the uniqueness theorem we study the difference between two solutions having the same sources. Thus, when we consider the difference, the source term disappears, and we will in this paper study fields satisfying the following conditions.

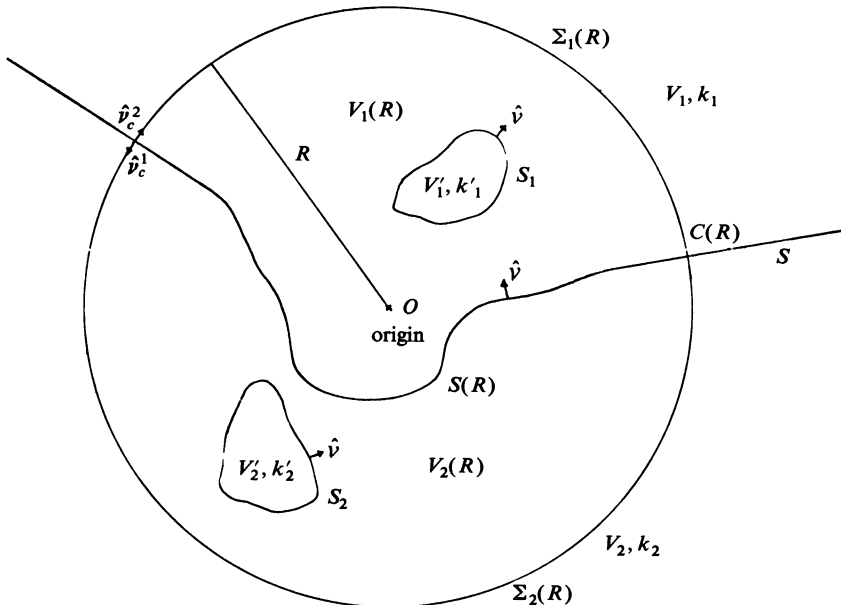


FIG. 1.

Condition 1.

$$\begin{aligned} (\nabla^2 + k_j^2)\psi_j(\vec{r}) &= 0, \quad \vec{r} \in V_j, \quad \psi_j \in C^2(V_j), \\ (\nabla^2 + k_j'^2)\psi_j'(\vec{r}) &= 0, \quad \vec{r} \in V_j', \quad \psi_j' \in C^2(V_j'), \end{aligned} \quad j=1,2.$$

Condition 2.

$$k_j^2 > 0, \quad k_j'^2 > 0, \quad j=1,2.$$

Condition 3 (boundary conditions).

$$\begin{aligned} A_1 \frac{\partial \psi_1}{\partial \nu} &= A_2 \frac{\partial \psi_2}{\partial \nu}, \\ B_1 \psi_1 &= B_2 \psi_2 \end{aligned} \quad \text{on } S,$$

$$\begin{aligned} A_j \frac{\partial \psi_j}{\partial \nu} &= A_j' \frac{\partial \psi_j'}{\partial \nu}, \\ B_j \psi_j &= B_j' \psi_j' \end{aligned} \quad \text{on } S_j, \quad j=1,2.$$

Here  $A_j, A_j', B_j, B_j', j=1,2$  are arbitrary complex constants such that  $C_j \equiv \bar{A}_j B_j = A_j \bar{B}_j$ , and  $C_j' \equiv \bar{A}_j' B_j' = A_j' \bar{B}_j'$ , where  $j=1,2$ , are nonnegative real numbers.

Condition 4. The surface  $S$  is conical outside a given radius, i.e.,  $\exists R_0$  such that for  $r \geq R_0$  we have

- a)  $\hat{\nu} \cdot \hat{r} = 0$  on  $S$  (This specifies our origin),
- b) a finite number of inhomogeneities and  $V_j(R_0) \supset V_j', j=1,2$ .

Condition 5 (radiation condition).

$$\iint_{\Sigma_j(R)} \left| \frac{\partial \psi_j}{\partial R} - ik_j \psi_j \right|^2 dS = o(1), \quad R \rightarrow \infty, \quad j=1,2.$$

Here and below  $\bar{A}$  will denote the complex conjugate of the complex number  $A$ .

The second condition states that all media are lossless. The case with loss will not be analyzed here, but it can be expected to be easier due to damping. The third condition gives the conditions at the boundaries; in this paper we will assume both the field and its normal derivative to be discontinuous. The coupling constants  $A$  and  $B$  can be arbitrary complex numbers such that the  $C \equiv \bar{A}B$  are nonnegative real numbers. The theory allows any  $A$  or  $B$  to be zero, and in this special case the theory is the problem treated by Jones [7]. The fourth condition requires the shape of  $S$  to be conical for large  $r$ , and also limits inhomogeneities in  $V_1$  and  $V_2$  to a finite number. The choice of origin is now also fixed due to the condition  $\hat{r} \cdot \hat{\nu} = 0$  for large  $r$ . The radiation condition which will be adopted here is the radiation condition discussed, e.g., by Rellich [21] and Jones [7]. See also Wilcox [30] for additional comments on the choice of radiation condition. Notice that this fifth item is nonlinear. However, Minkowski's inequality proves that the sum or difference of two fields satisfying Condition 5 still satisfies the radiation condition.

It is convenient to work with field quantities where a specific radial dependence has been extracted, so as to make the remaining radial dependence of the fields easier



to study. We will therefore adopt the following notation.

$$\begin{aligned}
 \phi_j(\vec{r}) &\equiv \sqrt{C_j} r^{(n-1)/2} \psi_j(\vec{r}), \\
 \phi'_j(\vec{r}) &\equiv \frac{\partial}{\partial r} \phi_j(\vec{r}), \quad j=1,2. \\
 \nabla_1 &\equiv r \left( \nabla - \hat{r} \frac{\partial}{\partial r} \right),
 \end{aligned}
 \tag{2.1}$$

The  $\nabla_1$  operator is defined on the unit sphere  $\{\vec{r} \in R^n: |\vec{r}| = 1\}$ , and is independent of the radial coordinate  $r$ . A direct computation of the gradient in spherical coordinates,

$$\begin{cases}
 x_n = r \cos \vartheta_{n-1}, & 0 \leq \vartheta_{n-1} \leq \pi, \\
 x_{n-1} = r \sin \vartheta_{n-1} \cos \vartheta_{n-2}, & 0 \leq \vartheta_{n-2} \leq 2\pi, \\
 \vdots & \\
 x_1 = r \sin \vartheta_{n-1} \sin \vartheta_{n-2} \cdots \sin \vartheta_1, & 0 \leq \vartheta_1 \leq 2\pi,
 \end{cases}$$

shows this. It is also a general result from differential geometry (for more details see e.g. [6]). If the Laplace-Beltrami operator on the unit sphere in  $R^n$  is denoted  $\nabla_1^2$  and is denoted  $\nabla^2$  in  $R^n$  itself, we have the following relation (see [27, p.6]):

$$\nabla_1^2 = r^2 \nabla^2 - r^{3-n} \frac{\partial}{\partial r} \left( r^{n-1} \frac{\partial}{\partial r} \right).
 \tag{2.2}$$

Helmholtz' equation can be rewritten in terms of the field  $\phi_j$  with  $\nabla_1^2$  and radial derivatives as

$$\phi_j'' + \frac{1}{r^2} \nabla_1^2 \phi_j + (k_j^2 - p_n(r)) \phi_j = 0,
 \tag{2.3}$$

where

$$p_n(r) \equiv \frac{(n-1)(n-3)}{4r^2}.
 \tag{2.4}$$

We notice that the quantity  $p_n(r)$  which depends on the dimension  $n$  is nonnegative, except for  $n=2$ . We also define the solid angle  $\Omega$  for a part of a hypersphere  $\Sigma$  in  $R^n$  as the projection on the unit sphere

$$\Omega \equiv \Sigma / r^{n-1}
 \tag{2.5}$$

( $\Sigma$  here is used both as a notation for the surface and for its measure). This solid angle  $\Omega$  for  $\Sigma_j(R)$  is a constant for  $r \geq R_0$ , with  $R_0$  chosen as in Condition 4.

Green's theorem for two fields  $u$  and  $v$  defined on the hypersphere will be used extensively (see, e.g., [6]):

$$\frac{1}{r^2} \iint_{\Sigma} [v \nabla_1^2 u + \nabla_1 u \cdot \nabla_1 v] dS = \frac{1}{r} \int_C v \hat{v}_c \cdot \nabla_1 u dl.
 \tag{2.6}$$

Here  $\Sigma$  is a part of the hypersphere of radius  $r$ ,  $C$  is its periphery defined in  $n-2$  dimensions, and  $\hat{v}_c$  is the outward normal to the periphery  $C$  (see Fig. 1).

In the following section we will prove the uniqueness of the fields  $\psi_j$  and  $\psi'_j$ ,  $j=1,2$ , satisfying Conditions 1-5 defined above; i.e., we will show that the only possible solution to Conditions 1-5 is the trivial solution  $\psi_j \equiv \psi'_j \equiv 0$ ,  $j=1,2$ . The main building blocks in this theorem will be four lemmas which will be proven first. The first three will make no use of the radiation condition, i.e., they give some general

features of fields satisfying Conditions 1–4 for large radial distances. These lemmas are extensions to and modifications for the present situation of lemmas given by Kato [8]. However, in this paper we will not rely on the symmetry properties of  $\nabla^2$  as in [8], but will make explicit use of Green's theorem on the hypersphere. The last lemma, which includes the radiation condition, will serve as a contradiction to the former, leaving just the trivial zero-solution as the only remaining solution.

Some quantities will appear often, so for convenience we define, for  $r \geq R_0$ , the following functions depending on the radial distance  $r$ .

$$(2.7) \quad E(r) \equiv \sum_{j=1}^2 \iint_{\Omega_j} [|\phi_j'|^2 + k_j^2 |\phi_j|^2] d\Omega,$$

$$(2.8) \quad G(r) \equiv E(r) - \frac{1}{r^2} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega.$$

Here  $\Omega_j$  is the solid angle for  $\Sigma_j(r)$ , which, as pointed out earlier, is a constant for  $r \geq R_0$ .

**3. Uniqueness theorem for permeable media.** We will in this section prove the uniqueness theorem for a configuration as depicted in Fig. 1 and with the assumptions and definitions stated in §2.

LEMMA 1. Consider two fields  $\phi_j$  satisfying Conditions 1–4 in the preceding section. We then have for  $r \geq R_0$ :

$$(3.1) \quad G'(r) = \frac{2}{r^3} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega + p_n(r) \frac{d}{dr} \sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega.$$

*Proof.* We take the derivative of (2.7) for  $r \geq R_0$ . Then we get, since  $\Omega_j$  is constant for  $r \geq R_0$ ,

$$E'(r) = 2 \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} [\phi_j'' \bar{\phi}_j' + k_j^2 \phi_j \bar{\phi}_j'] d\Omega.$$

We insert (2.3) and get

$$E'(r) = 2 \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \left[ p_n \phi_j - \frac{1}{r^2} \nabla_1^2 \phi_j \right] \bar{\phi}_j' d\Omega.$$

We apply Green's theorem (2.6) and get

$$E'(r) = 2 \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \left[ p_n \phi_j \bar{\phi}_j' + \frac{1}{r^2} \nabla_1 \phi_j \cdot \nabla_1 \bar{\phi}_j' \right] d\Omega.$$

The contribution from  $C$  vanishes. To see this, notice that

$$(3.2) \quad \frac{1}{r} \sum_{j=1}^2 \int_C \bar{\phi}_j' \hat{\nu}_c^j \cdot \nabla_1 \phi_j dl = - \int_C \left[ \bar{\phi}_1' \frac{\partial \phi_1}{\partial \nu} - \bar{\phi}_2' \frac{\partial \phi_2}{\partial \nu} \right] dl,$$

since for  $r \geq R_0$  we have  $\hat{\nu} \cdot \hat{r} = 0$  and

$$\frac{1}{r} \hat{\nu}_c^j \cdot \nabla_1 \phi_j = \mp \hat{\nu} \cdot \nabla \phi_j = \mp \frac{\partial \phi_j}{\partial \nu}.$$

By use of the boundary conditions on  $S$  (see Condition 3 in §2) and the definition in

(2.1) we can show the continuity of  $\bar{\phi}'_j \frac{\partial \phi_j}{\partial \nu}$  across  $S$ ,

$$\begin{aligned}
 \bar{\phi}'_2 \frac{\partial \phi_2}{\partial \nu} &= C_2 \frac{\partial}{\partial r} (r^{(n-1)/2} \bar{\psi}_2) \frac{\partial}{\partial \nu} (r^{(n-1)/2} \psi_2) \\
 (3.3) \quad &= A_2 \bar{B}_2 r^{n-1} \left( \frac{n-1}{2r} \bar{\psi}_2 + \frac{\partial \bar{\psi}_2}{\partial r} \right) \frac{\partial \psi_2}{\partial \nu} \\
 &= A_1 \bar{B}_1 r^{n-1} \left( \frac{n-1}{2r} \bar{\psi}_1 + \frac{\partial \bar{\psi}_1}{\partial r} \right) \frac{\partial \psi_1}{\partial \nu} = \bar{\phi}'_1 \frac{\partial \phi_1}{\partial \nu},
 \end{aligned}$$

since  $\hat{\nu} \cdot \hat{r} = 0$  for  $r \geq R_0$  and  $\psi_j$  can be differentiated along  $\hat{r}$  on  $S$ . The contribution from (3.2) is thus zero and we have

$$E'(r) = \sum_{j=1}^2 \left\{ p_n \frac{d}{dr} \iint_{\Omega_j} |\phi_j|^2 d\Omega + \frac{1}{r^2} \frac{d}{dr} \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega \right\},$$

since  $\nabla_1$  is independent of  $r$ .

From (2.8) we get, with this expression of  $E'(r)$ ,

$$G'(r) = \frac{2}{r^3} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega + p_n(r) \frac{d}{dr} \sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega,$$

and the lemma is proved.

The quantity  $F(m, r)$ , which will be used in the following lemma, is defined as

$$\begin{aligned}
 (3.4) \quad F(m, r) &\equiv \sum_{j=1}^2 \iint_{\Omega_j} \left[ |\phi_j^m|^2 + \left( k_j^2 - \frac{a}{r} + \frac{m(m+1)}{r^2} \right) |\phi_j^m|^2 - \frac{1}{r^2} |\nabla_1 \phi_j^m|^2 \right] d\Omega, \\
 \phi_j^m &\equiv r^m \phi_j, & j=1, 2, \quad m \text{ is an arbitrary positive integer} \\
 \phi_j^m &\equiv \frac{\partial}{\partial r} \phi_j^m, & j=1, 2, \\
 a &\equiv R_0 \min_{j=1,2} k_j^2 \equiv R_0 \kappa^2.
 \end{aligned}$$

LEMMA 2. *Let Conditions 1-4 in §2 be fulfilled. Then there are positive integers  $m_0, m_1$  and a number  $r_1 \geq R_0$  such that*

- a)  $\frac{d}{dr} (r^2 F(m, r)) \geq 0$  for all  $m \geq m_1$  and all  $r \geq R_0$ ,
- b)  $F(m_0, r) > 0$  for all  $r \geq r_1$  unless  $\phi_j \equiv 0, j=1, 2$ .

*Proof.* For  $r \geq R_0$  we have

$$\begin{aligned}
 \frac{d}{dr} (r^2 F(m, r)) &= 2r^2 \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \left[ \phi_j^m \overline{\phi_j^m} + \left( k_j^2 - \frac{a}{r} + \frac{m(m+1)}{r^2} \right) \phi_j^m \overline{\phi_j^m} \right. \\
 &\quad \left. - \frac{1}{r^2} \nabla_1 \overline{\phi_j^m} \cdot \nabla_1 \phi_j^m + \frac{1}{r} \left( k_j^2 - \frac{a}{2r} \right) |\phi_j^m|^2 + \frac{1}{r} |\phi_j^m|^2 \right] d\Omega.
 \end{aligned}$$

It is straightforward to prove, using (2.3), that  $\phi_j^m$  satisfies the following differential equation:

$$\phi_j^m + \frac{1}{r^2} \nabla_1^2 \phi_j^m - \frac{2m}{r} \phi_j^m + \frac{m(m+1)}{r^2} \phi_j^m + (k_j^2 - p_n(r)) \phi_j^m = 0.$$

We thus get, after some algebra,

$$(3.5) \quad \frac{d}{dr}(r^2 F(m, r)) = 2r \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \left[ |\phi_j^{\prime m}|^2 (2m+1) + \left(k_j^2 - \frac{a}{2r}\right) |\phi_j^m|^2 + \overline{\phi_j^{\prime m}} \phi_j^m (rp_n - a) - \frac{1}{r} (\nabla_1 \overline{\phi_j^{\prime m}} \cdot \nabla_1 \phi_j^m + \overline{\phi_j^{\prime m}} \nabla_1^2 \phi_j^m) \right] d\Omega.$$

The last term in the integrand disappears by use of Green's theorem (2.6) and the continuity of  $\overline{\phi_j^{\prime m}}(\partial \phi_j^m / \partial \nu)$ , since on  $C$  we have

$$\begin{aligned} \overline{\phi_1^{\prime m}} \frac{\partial \phi_1^m}{\partial \nu} &= (mr^{m-1} \overline{\phi_1} + r^m \overline{\phi_1'}) r^m \frac{\partial \phi_1}{\partial \nu} = (mr^{m-1} \overline{\phi_2} + r^m \overline{\phi_2'}) r^m \frac{\partial \phi_2}{\partial \nu} \\ &= \overline{\phi_2^{\prime m}} \frac{\partial \phi_2^m}{\partial \nu}. \end{aligned}$$

Furthermore, from Hölder's inequality for integrals on the hypersphere we have the following estimate:

$$\begin{aligned} \operatorname{Re} \sum_{j=1}^2 (rp_n - a) \iint_{\Omega_j} \overline{\phi_j^{\prime m}} \phi_j^m d\Omega &\geq -(r|p_n| + a) \sum_{j=1}^2 \iint_{\Omega_j} |\phi_j^{\prime m}| |\phi_j^m| d\Omega \\ &\geq -(r|p_n| + a) \sum_{j=1}^2 \sqrt{\iint_{\Omega_j} |\phi_j^{\prime m}|^2 d\Omega} \sqrt{\iint_{\Omega_j} |\phi_j^m|^2 d\Omega}. \end{aligned}$$

Thus we can write (3.5) as

$$\begin{aligned} \frac{d}{dr}(r^2 F(m, r)) &= 2r \sum_{j=1}^2 \iint_{\Omega_j} \left[ |\phi_j^{\prime m}|^2 (2m+1) + \left(k_j^2 - \frac{a}{2r}\right) |\phi_j^m|^2 + \operatorname{Re} \{ (rp_n - a) \overline{\phi_j^{\prime m}} \phi_j^m \} \right] d\Omega \\ &\geq 2r \sum_{j=1}^2 \left\{ \iint_{\Omega_j} \left[ |\phi_j^{\prime m}|^2 (2m+1) + \left(k_j^2 - \frac{a}{2r}\right) |\phi_j^m|^2 \right] d\Omega - (r|p_n| + a) \sqrt{\iint_{\Omega_j} |\phi_j^{\prime m}|^2 d\Omega} \sqrt{\iint_{\Omega_j} |\phi_j^m|^2 d\Omega} \right\}. \end{aligned}$$

The right hand side is a quadratic form which is greater than zero if

$$(3.6) \quad (2m+1) \left(k_j^2 - \frac{a}{2r}\right) \geq \frac{1}{4} (r|p_n| + a)^2 = \frac{1}{4} \left( \frac{|n-1||n-3|}{4r} + R_0 \kappa^2 \right)^2.$$

Since  $k_j^2 - a/(2r) = k_j^2 - (R_0/2r)\kappa^2 > 0$  for  $r \geq R_0$ , and the right-hand side of (3.6) is independent of  $m$  and bounded for large  $r$ , we can find  $m_1$  such that (3.6) is fulfilled for all  $m \geq m_1$  and  $r \geq R_0$ , i.e.,

$$\frac{d}{dr}(r^2 F(m, r)) \geq 0 \quad \text{for all } m \geq m_1 \text{ and all } r \geq R_0,$$

and the first part of the lemma is proved. If  $\phi_j \neq 0$ , then there exists an  $r_1 \geq R_0$  such that

$$\sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega > 0, \quad r = r_1.$$

(If this is false for every  $r_1 \geq R_0$  then  $\phi_j = 0$  for all  $r \geq R_0$  and by the properties of solutions to Helmholtz' equation  $\phi_j$  is zero everywhere.) We can write  $F(m, r)$  in (3.4) as

(3.7)

$$F(m, r) = r^{2m} \sum_{j=1}^2 \iint_{\Omega_j} \left[ \left| \frac{m}{r} \phi_j + \phi_j' \right|^2 + \left( k_j^2 - \frac{a}{r} + \frac{m(m+1)}{r^2} \right) |\phi_j|^2 - \frac{1}{r^2} |\nabla_1 \phi_j|^2 \right] d\Omega.$$

For an  $r_1$  chosen as above let  $m_0 \geq m_1$ , where  $m_1$  is given by the first part of this lemma, such that  $F(m_0, r_1) > 0$ . But according to the first part of this lemma we then have  $F(m_0, r) > 0$  for all  $r \geq r_1 \geq R_0$ , since  $r^2 F(m_0, r)$  is a nondecreasing function in  $r$ ; i.e.  $F(m_0, r)$  can not change sign. The proof of Lemma 2 is thus completed.

The next lemma, in which we will use the previous one, reads

LEMMA 3. *Let Conditions 1-4 be fulfilled and furthermore let*

$$\sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega = o(1), \quad r \rightarrow \infty.$$

If  $\phi_j \not\equiv 0$  then there exists an infinite sequence of real numbers  $\{r_\mu\}_{\mu=1}^\infty$  such that  $r_\mu \rightarrow \infty$ ,  $\mu \rightarrow \infty$  and  $G(r_\mu) > 0$ .

*Proof.* Define a set  $T$  such that

$$T = \left\{ r \geq R_0 : \frac{d}{dr} \sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega \leq 0 \right\}.$$

$T$  is an infinite set and furthermore contains arbitrarily big elements. This is a consequence of the assumption

$$\sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega = o(1), \quad r \rightarrow \infty.$$

For an  $r_\mu \in T$  we have

$$\operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \bar{\phi}_j \phi_j' d\Omega = \frac{1}{2} \frac{d}{dr} \sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega \leq 0.$$

Thus we have, for  $r_\mu \in T$ ,

$$\sum_{j=1}^2 \iint_{\Omega_j} \left| \phi_j' + \frac{m}{r} \phi_j \right|^2 d\Omega \leq \sum_{j=1}^2 \iint_{\Omega_j} \left[ |\phi_j'|^2 + \frac{m^2}{r^2} |\phi_j|^2 \right] d\Omega.$$

We can now estimate  $F(m, r_\mu)$  (for definitions see (3.4) or (3.7)),

$$F(m, r_\mu) \leq r_\mu^{2m} \sum_{j=1}^2 \iint_{\Omega_j} \left[ |\phi_j'|^2 + \left( k_j^2 - \frac{a}{r_\mu} + \frac{m(2m+1)}{r_\mu^2} \right) |\phi_j|^2 - \frac{1}{r_\mu^2} |\nabla_1 \phi_j|^2 \right] d\Omega.$$

Let  $m = m_0$ , where  $m_0$  is chosen according to Lemma 2 b). For all  $r_\mu \in T$  such that  $r_\mu \geq r_1 \geq R_0$  ( $r_1$  given by Lemma 2 b)) we have

$$0 < F(m_0, r_\mu) \leq r_\mu^{2m_0} \left\{ G(r_\mu) - \sum_{j=1}^2 \iint_{\Omega_j} \left( \frac{a}{r_\mu} - \frac{m_0(2m_0+1)}{r_\mu^2} \right) |\phi_j|^2 d\Omega \right\}.$$

Let  $r_2 \geq r_1$  such that  $\frac{a}{r_2} = \frac{R_0}{r_2} \kappa^2 \geq m_0(2m_0 + 1)/r_2^2$ . Then for all  $r_\mu \in T$  such that  $r_\mu \geq r_2$  we have

$$0 < F(m_0, r_\mu) \leq r_\mu^{2m_0} G(r_\mu) \Rightarrow G(r_\mu) > 0,$$

and the lemma is proved.

The last lemma makes explicit use of the radiation condition (Condition 5 in §2), and furthermore contradicts Lemma 3 as will be seen in the theorem below.

LEMMA 4. *Let Conditions 1–5 be fulfilled. Then*

$$\begin{aligned} \iint_{\Sigma_j(R)} |\psi_j|^2 dS &= o(1), \\ \iint_{\Sigma_j(R)} \left| \frac{\partial \psi_j}{\partial R} \right|^2 dS &= o(1), \end{aligned} \quad R \rightarrow \infty, \quad i=1,2.$$

*Proof.* The radiation condition (Condition 5) can explicitly be written as

$$\iint_{\Sigma_j(R)} \left| \frac{\partial \psi_j}{\partial R} - ik_j \psi_j \right|^2 dS = \iint_{\Sigma_j(R)} \left[ \left| \frac{\partial \psi_j}{\partial R} \right|^2 + k_j^2 |\psi_j|^2 + 2k_j \operatorname{Im} \left( \psi_j \frac{\partial \bar{\psi}_j}{\partial R} \right) \right] dS$$

Multiplying both sides by  $C_j/k_j$  and summing over  $j=1,2$  ( $C_j = A_j \bar{B}_j$ ), we get

$$(3.8) \quad o(1) = \sum_{j=1}^2 \frac{C_j}{k_j} \iint_{\Sigma_j(R)} \left[ \left| \frac{\partial \psi_j}{\partial R} \right|^2 + k_j^2 |\psi_j|^2 \right] dS + 2 \sum_{j=1}^2 \operatorname{Im} \left[ \iint_{\Sigma_j(R)} C_j \psi_j \frac{\partial \bar{\psi}_j}{\partial R} dS \right].$$

We now apply Green's first formula in  $R^n$  (this is analogous to (2.6), which holds on the hypersphere) on the field  $\psi_j$  and its complex conjugate in  $V_j$  and  $V'_j, j=1,2$ . We get, for  $j=1,2$ ,

$$\iint_{\Sigma_j(R)} \psi_j \frac{\partial \bar{\psi}_j}{\partial R} dS = \pm \iint_{S(R)} \psi_j \frac{\partial \bar{\psi}_j}{\partial \nu} dS + \iint_{S_j} \psi_j \frac{\partial \bar{\psi}_j}{\partial \nu} dS + \iiint_{V_j(R)} [|\nabla \psi_j|^2 - k_j^2 |\psi_j|^2] dV,$$

$$\iint_{S_j} \psi_j \frac{\partial \bar{\psi}_j}{\partial \nu} dS = \iiint_{V'_j} [|\nabla \psi'_j|^2 - k_j'^2 |\psi'_j|^2] dV.$$

The plus sign holds for  $j=1$  and the minus for  $j=2$ . This is a consequence of the direction of the surface normal  $\hat{\nu}$  on  $S$ .

We have so far not specified the smoothness properties assumed for the surface  $S, S_1$  and  $S_2$ . These properties are here assumed to guarantee the finiteness of the surface integrals over  $S, S_1$  and  $S_2$  above. This property was not used in the preceding lemmas. The last term in (3.8) can be rewritten as

$$\begin{aligned} \operatorname{Im} \left[ \sum_{j=1}^2 \iint_{\Sigma_j(R)} C_j \psi_j \frac{\partial \bar{\psi}_j}{\partial R} dS \right] &= \operatorname{Im} \left[ \sum_{j=1}^2 \iint_{S_j} C_j \psi_j \frac{\partial \bar{\psi}_j}{\partial \nu} dS \right] \\ &+ \iint_{S(R)} \left[ C_1 \psi_1 \frac{\partial \bar{\psi}_1}{\partial \nu} - C_2 \psi_2 \frac{\partial \bar{\psi}_2}{\partial \nu} \right] dS = 0, \end{aligned}$$

since  $C_j \psi_j \partial \bar{\psi}_j / \partial \nu$  is continuous over  $S, S_1$  and  $S_2$  by the boundary conditions. Thus we have, from (3.8),

$$\sum_{j=1}^2 \frac{C_j}{k_j} \iint_{\Sigma_j(R)} \left[ \left| \frac{\partial \psi_j}{\partial R} \right|^2 + k_j^2 |\psi_j|^2 \right] dS = o(1), \quad R \rightarrow \infty.$$

Since each term is positive the statement of the lemma is proved.

We have now collected results enough to prove the main theorem of this paper.

**THEOREM.** *Let  $\psi_j$  and  $\psi'_j$   $j=1,2$  be fields satisfying Conditions 1–5, defined and discussed in §2. The only possible solution to this problem is the solution which is zero everywhere.*

*Proof.* Lemma 1 gives, for  $r \geq R_0$ ,

$$G'(r) = \frac{2}{r^3} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega + 2p_n(r) \operatorname{Re} \sum_{j=1}^2 \iint_{\Omega_j} \phi_j \bar{\phi}'_j d\Omega,$$

Furthermore we have ( $\kappa \equiv \min_{j=1,2} k_j$ )

$$0 \leq \sum_{j=1}^2 \iint_{\Omega_j} |\phi'_j \pm \kappa \phi_j|^2 d\Omega = \sum_{j=1}^2 \iint_{\Omega_j} [|\phi'_j|^2 + \kappa^2 |\phi_j|^2 \pm 2\kappa \operatorname{Re} \phi_j \bar{\phi}'_j] d\Omega,$$

and we can write (if  $n \neq 2$  use the plus sign, if,  $n=2$  the minus sign)

$$\begin{aligned} G'(r) &\geq -\frac{|p_n(r)|}{\kappa} \sum_{j=1}^2 \iint_{\Omega_j} [|\phi'_j|^2 + \kappa^2 |\phi_j|^2] d\Omega + \frac{2}{r^3} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega \\ &\geq -\frac{|p_n(r)|}{\kappa} \sum_{j=1}^2 \iint_{\Omega_j} [|\phi'_j|^2 + k_j^2 |\phi_j|^2] d\Omega + \frac{2}{r^3} \sum_{j=1}^2 \iint_{\Omega_j} |\nabla_1 \phi_j|^2 d\Omega \\ &= -\frac{|p_n(r)|}{\kappa} G(r) + \sum_{j=1}^2 \iint_{\Omega_j} \left( \frac{2}{r^3} - \frac{|p_n(r)|}{r^2 \kappa} \right) |\nabla_1 \phi_j|^2 d\Omega \\ &\geq -\frac{|p_n(r)|}{\kappa} G(r) \quad \text{for all } r \geq r_3 \geq R_0, \end{aligned}$$

where  $r_3$  satisfies

$$\frac{2}{r_3^3} \geq \frac{|p_n(r)|}{r_3^2 \kappa} = \frac{|n-1||n-3|}{4r_3^4 \kappa}.$$

Thus we have the following differential inequality:

$$(3.9) \quad G'(r) + f(r)G(r) \geq 0,$$

where  $f(r) = |p_n(r)|/\kappa = b_n/r^2$ . The solution to (3.9) is

$$G(r) \geq G(r_0) \exp \left[ - \int_{r_0}^r f(t) dt \right] = G(r_0) \exp \left[ b_n \left( \frac{1}{r} - \frac{1}{r_0} \right) \right] \quad \text{for } r \geq r_0.$$

Lemma 4 gives

$$\sum_{j=1}^2 \iint_{\Omega_j} |\phi_j|^2 d\Omega = o(1), \quad R \rightarrow \infty.$$

Lemma 3 now states that there exists an arbitrarily large  $r_0$  such that  $G(r_0) > 0$  provided  $\phi_j \not\equiv 0$ ; we thus have  $\lim_{r \rightarrow \infty} G(r) > 0$  unless  $\phi_j \equiv 0$ . On the other hand we have

$$\lim_{r \rightarrow \infty} G(r) \leq \lim_{r \rightarrow \infty} E(r) = \lim_{r \rightarrow \infty} \sum_{j=1}^2 \iint_{\Omega_j} [|\phi_j'|^2 + k_j^2 |\phi_j|^2] d\Omega.$$

Furthermore Lemma 4 gives

$$\sum_{j=1}^2 \iint_{\Omega_j} |\phi_j'|^2 d\Omega = \sum_{j=1}^2 C_j \iint_{\Sigma_j(R)} \left| \frac{n-1}{2R} \psi_j + \frac{\partial \psi_j}{\partial R} \right|^2 dS = o(1), \quad R \rightarrow \infty.$$

This last step can be shown by Hölder's inequality. We thus have  $\lim_{r \rightarrow \infty} G(r) \leq 0$ . This contradicts  $\lim_{r \rightarrow \infty} G(r) > 0$ , and we have  $\phi_j \equiv 0$  everywhere.

**4. Conclusions and applications.** We have in this paper shown the uniqueness of the solution for Helmholtz' equation in the lossless case for a special class of infinite surfaces, namely those which eventually become conical. As a special case, if one of  $A_j$  or  $B_j$  equals zero, we have the result of Jones [7].

The proof of the theorem relies in several places on the fact that  $\hat{r} \cdot \hat{v} = 0$  for  $r \geq R_0$ . This property makes it possible to differentiate one of the boundary conditions in the radial direction. A uniqueness theorem valid for a more general geometry thus must in some parts use different techniques and arguments. In [19] it is stated that the case with losses (complex  $k_j$ ) can be proved by simple boundedness conditions, but this is not carried out in detail. The theorem proved in this paper can not be extended as it stands to complex values, but we expect that only slight modifications will be required in order to make this extension possible. The boundary conditions assumed in this paper are fairly general, but an interesting extension would be to investigate how more general conditions would affect the uniqueness. At present this is an unsolved problem. The volumes  $V_1'$  and  $V_2'$  were assumed homogeneous and lossless, but these assumptions can easily be relaxed.

The uniqueness theorem for Helmholtz' equation together with the derived growth properties at large distances is of great interest in many situations. One application which has recently appeared is the question of completeness of various systems of functions on a given surface, finite or not; see, e.g., [10], [11], [15], [18]. This question is of special interest for eigenfunctions to Helmholtz' equation for surfaces which are not coordinate surfaces to the eigenfunctions. The technique used



by Millar [15] relies on the uniqueness results (in the use of either Dirichlet or Neumann boundary value problems) for the corresponding geometry, in the interior and the exterior case.

**Acknowledgment.** The author wishes to thank Dr. Staffan Ström for a careful reading of the manuscript. The work reported in the present paper is part of a project sponsored by the National Swedish Board for Technical Development (STU) and their support is gratefully acknowledged.

## REFERENCES

- [1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, 1965.
- [2] F. V. ATKINSON, *On Sommerfeld's "Radiation Condition,"* Phil. Mag., 40 (1949), pp. 645–651.
- [3] D. M. EIDUS, *The principle of limiting absorption*, Am. Math. Soc. Transl., Series 2, 47 (1965), pp. 157–191.
- [4] ———, *Some boundary-value problems in infinite regions*, Am. Math. Soc. Transl., Series 2, 53 (1966), pp. 139–166.
- [5] ———, *The Principle of limit amplitude*, Russian Math. Surveys, 24 (1969), pp. 97–167.
- [6] H. FLANDERS, *Differential Forms with Applications to the Physical Sciences*, Academic Press, New York, 1963.
- [7] D. S. JONES, *The eigenvalues of  $\nabla^2 u + \lambda u = 0$  when the boundary conditions are given on semi-infinite domains*, Proc. Camb. Phil. Soc., 49 (1953), pp. 668–684.
- [8] T. KATO, *Growth properties of solutions of the reduced wave equation with a variable coefficient*, Comm. Pure Appl. Math., 12 (1959), pp. 403–425.
- [9] O. D. KELLOG, *Foundation of Potential Theory*, Dover, New York, 1953.
- [10] G. KRISTENSSON AND S. STRÖM, *Scattering from buried inhomogeneities—a general three-dimensional formalism*, J. Acoust. Soc. Am., 64 (1978), pp. 917–936.
- [11] G. KRISTENSSON, *Electromagnetic scattering from buried inhomogeneities—a general three-dimensional formalism*, J. Appl. Phys., 51 (1980), pp. 3486–3500.
- [12] L. M. LEVINE, *A uniqueness theorem for the reduced wave equation*, Comm. Pure Appl. Math., 17 (1964), pp. 147–176.
- [13] W. MAGNUS, *Über Eindeutigkeitsfragen bei einer Randwertaufgabe von  $\Delta u + k^2 u = 0$* , Jber. Deutschen Math. Verein., 52 (1942), pp. 177–188.
- [14] ———, *Fragen der Eindeutigkeit und des Verhaltens im Unendlichen für Lösungen von  $\Delta u + k^2 u = 0$* , Abhandl. Math. Sem. Hamburg. Bd. 16 (1949), pp. 77–94.
- [15] R. F. MILLAR, *The Rayleigh hypothesis and a related least-squares solution to scattering problems for periodic surfaces and other scatterers*, Radio Sci. 8 (1973), pp. 785–796.
- [16] W. L. MIRANKER, *Uniqueness and representation theorems for solutions of  $\Delta u + k^2 u = 0$  in infinite domains*, J. Math. Mech., 6 (1957), pp. 847–858.
- [17] C. MÜLLER, *Foundation of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, New York, 1969.
- [18] C. MÜLLER, *Boundary values and diffraction problems*, in Symposia Mathematical XVIII, Academic Press, New York, 1976, pp. 353–367.
- [19] F. M. ODEH, *Uniqueness theorems for the Helmholtz equation in domains with infinite boundaries*, J. Math. Mech., 12 (1963), pp. 857–868.
- [20] A. S. PETERS AND J. J. STOKER, *A uniqueness theorem and a new solution for Sommerfeld's and other diffraction problems*, Comm. Pure Appl. Math., 7 (1954), pp. 565–585.
- [21] F. RELICH, *Über das asymptotische Verhalten der Lösungen von  $\Delta u + \lambda u = 0$  in unendlichen Gebieten*, Jber. Deutschen Math. Verein., 53 (1943), pp. 57–64.
- [22] ———, *Das Eigenwertproblem von  $\Delta u + \lambda u = 0$  in Halbröhren*, in Studies and Essays presented to R. Courant, Interscience, New York, 1948, pp. 329–344.
- [23] A. SOMMERFELD, *Die Greensche Funktion der Schwingungsgleichung*, Jber. Deutschen Math. Verein., 21 (1912), pp. 309–353.
- [24] B. R. VAINBERG, *Principle of radiation, limit absorption and limit amplitude in general theory of partial differential equations*, Russian Math. Surveys, 21 (1966), pp. 115–193.

- [25] V. VOGELSANG, *Das Ausstrahlungsproblem für elliptische Differentialgleichungen in Gebieten mit unbeschränktem Rand*, Math. Z., 144 (1975), pp. 101–124.
- [26] ———, *Elliptische Differentialgleichungen mit variablen Koeffizienten in Gebieten mit unbeschränktem Rand*, Manuscripta Math., 14 (1975), pp. 379–401.
- [27] N. R. WALLACH, in *Symmetrical Spaces*, W. M. Boothby and G. L. Weiss, eds., Marcel Dekker, New York, 1972.
- [28] P. WERNER, *Zur mathematischen Theorie akustischer Wellenfelder*, Arch. Rat. Mech. Anal., 6 (1960), pp. 231–260.
- [29] C. H. WILCOX, *A Generalization of theorems of Rellich and Atkinson*, Proc. Am. Math. Soc., 7 (1956), pp. 271–276.
- [30] C. H. WILCOX, *Spherical means and radiation conditions*, Arch. Rat. Mech. Anal., 3 (1959), pp. 133–148.